



Project Deliverable

D2.1

In-depth analysis of SIEMs extensibility

Project Number	700692
Project Title	DiSIEM – Diversity-enhancements for SIEMs
Programme	H2020-DS-04-2015

Deliverable type	Report
Dissemination level	PU
Submission date	28.02.2017

Responsible partner	ATOS
Editor	Susana González Zarzosa
Revision	1.0



The DiSIEM project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700692.

Editor

Susana González Zarzosa, Atos

Contributors

Antonio Galán Corroto, Atos

Ignacio Robla, Atos

Elsa Prieto Pérez, Atos

Susana González Zarzosa, Atos

Alysson Bessani, FFCUL

Ana Respício, FFCUL

João Alves, FFCUL

Luís Ferreira, FFCUL

Adriano Serckumecka, FFCUL

Pedro Dias Rodrigues, EDP

Pedro Daniel Ferreira, EDP

Gonçalo Santos Martins, EDP

Zayani Dabbabi, Amadeus

Miruna-Mihaela Mironescu, Amadeus

Frances Buontempo, City

Cagatay Turkay, City

Executive Summary

This report documents the current state of the art in SIEM systems with a forward vision on potential enhancements and extensions to be developed in DiSIEM with diversity-related technology.

The main goal of this document is to describe the present scenario of Security Information and Event Management (SIEM) systems as the starting point for future enhancements of these systems proposed in DiSIEM. Main features to be considered for comparing these systems are exposed and, based on these features, several available SIEM offers and related systems which are not SIEMs but have been selected by their relevance in the current security landscape (such as Elastic Stack or Splunk) are analysed. Potential extensions to be developed are also analysed together with main factors that could affect the future of SIEMs, including potential exploitation aspects to be considered.

The main features of this document are the following:

- General overview of relevant features to be considered in the analysis and extension of SIEM systems.
- Details on features currently provided by some of the most relevant SIEM solutions available in the market with focus on the systems used by the partners involved in the project, where DiSIEM outcomes will be validated. More specifically, the following SIEMs have been analysed: HP ArcSight, IBM Q Radar, Intel McAfee Enterprise Security Manager, Alienvault OSSIM/USM and Atos XL-SIEM.
- Analysis of features provided by the following emerging tools which are also relevant in a SIEM context: Elastic Stack and Splunk.
- First analysis of potential SIEM enhancements to be investigated later with more detail and implemented throughout the different work packages WP3, WP4, WP5 and WP6.
- PEST analysis with a list of political, economic, socio-cultural and technological factors that can act as enablers or barriers to the development of SIEMs in the mid and long-term.

Table of Contents

1.	Introduction.....	12
1.1	Organization of the Document.....	12
2	Analysis of SIEM features and capabilities.....	13
2.1	Introduction.....	13
2.2	Data sources supported.....	13
2.3	Data storage capabilities.....	13
2.4	Processing capabilities.....	13
2.5	Flexibility in security directives.....	14
2.6	Behavioural analysis at application-level.....	14
2.7	Risk analysis capacity.....	14
2.8	Exposed APIs.....	15
2.9	Resilience.....	15
2.10	Security event management and visualization capabilities.....	15
2.11	Reaction capabilities.....	15
2.12	Simplicity of deployment and support provided.....	16
2.13	Licensing.....	16
2.14	Comparison between these criteria and others.....	16
2.14.1	Gartner Magic Quadrant.....	16
2.14.2	Other Criteria.....	17
3	Analysis of SIEM solutions.....	19
3.1	Introduction.....	19
3.2	HP ArcSight.....	21
3.2.1	Data sources supported.....	23
3.2.2	Data storage capabilities.....	23
3.2.3	Processing capabilities.....	23
3.2.4	Flexibility in security directives.....	24
3.2.5	Behavioural analysis at application-level.....	24
3.2.6	Risk analysis capacity.....	25
3.2.7	Exposed APIs.....	25
3.2.8	Resilience.....	25
3.2.9	Security event management and visualization capabilities.....	26
3.2.10	Reaction capabilities.....	26
3.2.11	Deployment and support.....	26
3.2.12	Licensing.....	26
3.2.13	Position in Gartner Magic Quadrant.....	26
3.3	IBM QRadar.....	28
3.3.1	Data sources supported.....	28
3.3.2	Data storage capabilities / Processing capabilities.....	29
3.3.3	Flexibility in security directives.....	30
3.3.4	Behavioural analysis at application-level.....	31
3.3.5	Risk analysis capacity.....	32
3.3.6	Exposed APIs.....	32
3.3.7	Resilience.....	32
3.3.8	Security event management and visualization capabilities.....	33
3.3.9	Reaction capabilities.....	34
3.3.10	Deployment and support.....	35
3.3.11	Licensing.....	37

3.3.12	Position in Gartner Magic Quadrant	38
3.4	Intel McAfee Enterprise Security Manager	39
3.4.1	Data sources supported	40
3.4.2	Data storage capabilities	40
3.4.3	Processing capabilities	41
3.4.4	Flexibility in security directives	42
3.4.5	Behavioural analysis at application-level	42
3.4.6	Risk analysis capacity	43
3.4.7	Exposed APIs	43
3.4.8	Resilience	44
3.4.9	Security event management and visualization capabilities	44
3.4.10	Reaction capabilities	45
3.4.11	Deployment and support	46
3.4.12	Licensing	46
3.4.13	Position in Gartner Magic Quadrant	46
3.5	Alienvault OSSIM and USM	47
3.5.1	Data sources supported	48
3.5.2	Data storage capabilities	50
3.5.3	Processing capabilities	51
3.5.4	Flexibility in security directives	52
3.5.5	Behavioural analysis at application-level	54
3.5.6	Risk analysis capacity	54
3.5.7	Exposed APIs	55
3.5.8	Resilience	56
3.5.9	Security event management and visualization capabilities	56
3.5.10	Reaction capabilities	57
3.5.11	Deployment and support	58
3.5.12	Licensing	58
3.5.13	Position in Gartner Magic Quadrant	58
3.6	XL-SIEM	59
3.6.1	Data sources supported	60
3.6.2	Data storage capabilities	60
3.6.3	Processing capabilities	61
3.6.4	Flexibility in security directives	62
3.6.5	Behavioural analysis at application-level	64
3.6.6	Risk analysis capacity	64
3.6.7	Exposed APIs	64
3.6.8	Resilience	65
3.6.9	Security event management and visualization capabilities	65
3.6.10	Reaction capabilities	66
3.6.11	Deployment and support	66
3.6.12	Licensing	67
3.6.13	Position in Gartner Magic Quadrant	67
3.7	Splunk	68
3.7.1	Introduction	68
3.7.2	Data sources supported	68
3.7.3	Data storage capabilities	70
3.7.4	Processing capabilities	71
3.7.5	Flexibility in security directives	76

3.7.6	Behavioural analysis at application-level	76
3.7.7	Risk analysis capacity	77
3.7.8	Exposed APIs.....	79
3.7.9	Resilience.....	80
3.7.10	Security event management and visualization capabilities.....	81
3.7.11	Reaction capabilities	81
3.7.12	Deployment and support.....	82
3.7.13	Licensing.....	83
3.7.14	Position in Gartner Magic Quadrant	85
3.8	Elastic Stack	87
3.8.1	Introduction	87
3.8.2	Data sources supported	88
3.8.3	Data storage capabilities	89
3.8.4	Processing capabilities.....	91
3.8.5	Flexibility in security directives	95
3.8.6	UEBA integration.....	95
3.8.7	Risk analysis capacity	95
3.8.8	Exposed APIs.....	96
3.8.9	Resilience.....	97
3.8.10	Event management and Visualization capabilities.....	99
3.8.11	Reaction capabilities	101
3.8.12	Deployment and Support	102
3.8.13	Licensing.....	103
3.8.14	Position in Gartner Magic Quadrant	103
4	Analysis of potential extensions to be developed in DiSIEM.....	105
4.1	Introduction	105
4.2	Security metrics and probabilistic modelling.....	106
4.2.1	Risk assessment and security metrics.....	106
4.2.2	Probabilistic modelling for diversity and defence in depth.....	109
4.2.3	Statistical analysis for anomaly detection	112
4.3	OSINT data fusion and analysis.....	113
4.3.1	Data Collection	113
4.3.2	Data storage.....	114
4.3.3	Monitored IT infrastructure specification	114
4.3.4	Threat detection	115
4.3.5	Incidence reporting and analysis	116
4.4	Visualization capabilities	116
4.5	Infrastructure enhancements	119
4.5.1	UBA	119
4.5.2	Event consolidation.....	119
4.5.3	Storage capabilities enhancement.....	119
5	Role of SIEMs in the future: barriers and enablers	122
5.1	Introduction	122
5.2	Technological factors	122
5.2.1	SIEM technology improvements	122
5.2.2	Technology trends	123
5.3	Societal factors.....	126
5.4	Economic factors.....	127
5.5	Political factors	128

5.5.1	EU regulation in Data protection.....	128
5.5.2	Investment R&I on EU Cybersecurity.....	129
6	Summary and Conclusions.....	131
7	List of Acronyms.....	135
8	References.....	137
9	Appendix I: Elasticsearch Products for Log Ingestion	143
10	Appendix II: Elasticsearch deployment	147

List of Figures

Figure 3-1: Gartner Magic Quadrant for SIEM 2016	20
Figure 3-2: ArcSight architecture at EDP.....	22
Figure 3-3: ArcSight Hadoop ecosystem.....	25
Figure 3-4: IBM QRadar event and flow components.....	36
Figure 3-5: IBM Q Radar Geographically Distributed Deployment.....	37
Figure 3-6: McAfee SIEM architecture.....	39
Figure 3-7: McAfee Enterprise Security Manager dashboard	45
Figure 3-8 : Alienvault OSSIM Architecture [12]	47
Figure 3-9 : Alienvault USM Architecture [13]	48
Figure 3-10: One sensor data to several USM	52
Figure 3-11: SIEMs data to several loggers.....	52
Figure 3-12: OSSIM Graphical Interface.....	57
Figure 3-13: XL-SIEM Architecture	59
Figure 3-14: XL-SIEM Web Graphical Interface	66
Figure 3-15: Splunk Data Indexing	72
Figure 3-16 : Indexing process in Splunk.....	73
Figure 3-17 Splunk Risk Analysis Framework.....	78
Figure 3-18 Splunk Deployment units – Dependencies.....	83
Figure 3-19: Splunk Enterprise Price.....	85
Figure 3-20 Advanced Threat Detection – Gartner	86
Figure 3-21: The standard components of Elastic Stack	88
Figure 3-22: A JSON file example	88
Figure 3-23: Kibana reporting.....	93
Figure 3-24: Kibana Visualize tab.....	100
Figure 3-25: Kibana Dashboard example	100
Figure 3-26: Timelion, a time series data visualizer	101
Figure 3-27: Kibana Alerting	102
Figure 3-28: Elastic Stack types of support	103
Figure 3-29: Gartner – Incident Response	104
Figure 4-1: Data collection and workflow with ArcSight in EDP.....	109
Figure 4-2: OSINT data processing using listening247	114
Figure 9-1 : Elasticsearch Filebeat diagram	144
Figure 9-2 : Elasticsearch Metricbeat diagram	145
Figure 9-3 : Hadoop Elasticsearch ecosystem	146
Figure 10-1: ELK minimal setup	147
Figure 10-2 : ELK mature setup deployment.....	147

List of Tables

Table 1: QRadar REST API endpoints summary	32
Table 2: Chart types in QRadar	34
Table 3: Data Sources supported by Alienvault SIEMs	49
Table 4: Example of Alienvault SIEMs security directive	53
Table 5: Attributes in correlation rules supported by Alienvault SIEMs	54
Table 6: Types of forwarders provided by Splunk.....	70
Table 7: Bucket stages provided by Splunk.....	70
Table 8: REST API Reference Manual Categories provided by Splunk	79
Table 9 : Lucene files	91
Table 10: Summary of SIEM enablers and barriers.....	130
Table 11: Strengths and weaknesses of analysed SIEMs.....	132
Table 12: Evaluation of analysed SIEMs	132
Table 13 : Input, Filter and Output plugins in Elastic Stack.....	143

Revision History

Version	Date	Author	Notes
0.1	11.11.2016	Susana González (ATOS)	ToC
0.2	15.12.2016	Antonio Galán (ATOS), Ignacio Robla (ATOS), Susana González (ATOS)	First contribution to Sections 3.4, 3.5, 3.6, 5
0.3	16.12.2016	Susana González (ATOS)	Integrated chapter 4 available in google doc
0.4	19.12.2016	Pedro Daniel Ferreira (EDP)	Changes in Section 4
0.5	19.12.2016	Ana Respicio (FFCUL)	Updated Section 4.2.1
0.6	19.12.2016	Alysson Bessani (FFCUL)	Integrated first version of chapter 2
0.7	20.12.2016	Pedro Dias Rodrigues (EDP)	Included ArcSight Section.
0.8	20.12.2016	Zayani Dabbabi (Amadeus)	Added Splunk section.
0.9	23.12.2016	Frances Buontempo (City)	Fixed typos.
0.10	28.12.2016	Pedro Dias Rodrigues (EDP)	Contribution to Section 3.2
0.11	09.01.2017	Zayani Dabbabi (Amadeus)	ELK stack added
0.12	09.01.2017	Frances Buontempo (City)	Added some reference texts
0.13	09.01.2017	Antonio Galán Corroto (Atos)	Executive Summary and Introduction. Added bibliography and several edition tasks.
0.14	12.01.2017	Pedro Dias Rodrigues (EDP)	Added text about machine learning in EDP's platform (Section 3.2.5).
0.15	17.01.2017	Alysson Bessani (FFCUL)	Updates and comments on Elastic Stack section.
0.16	18.01.2017	Zayani Dabbabi (Amadeus)	Added QRadar Amadeus
0.17	18.01.2017	Susana González Zarzosa (Atos)	Updates to Section 3.6. Relocated Alienvault section.
0.18	25.01.2017	Pedro Dias Rodrigues (EDP)	Updated information about ArcSight.
0.19	27.01.2017	Susana González Zarzosa (Atos) Ignacio Robla Sánchez (Atos)	Updated the following sections: Table of acronyms, table of references, Sections 3.4, 3.5, 3.6, 5.1, 5.2,

			5.3, 5.4, 5.5, 6. conclusions
0.20	31.01.2017	Susana González Zarzosa (Atos)	Added captions to figures and tables
0.21	06.02.2017	Adriano Serckumecka	Added Section 3.8.7
0.22	13.02.2017	Susana González Zarzosa (Atos)	Completed conclusions. Moved text to appendix. Draft ready for internal review.
0.23	15.02.2017	Alysson Bessani (FFCUL)	Completed Section 2.14
0.24	17.02.2017	Zayani Dabbabi (Amadeus)	Added missing Risk Analysis and Magic Quadrant sections for IBM Q Radar
0.25	22.02.2017	Zayani Dabbabi (Amadeus)	Fixed internal reviewers comments.
0.26	24.02.2017	Susana González Zarzosa (Atos)	Fixed internal reviewers comments.
0.27	27.02.2017	Frances Buontempo (City)	Fixed internal reviewers comments.
0.28	27.02.2017	Zayani Dabbabi (Amadeus)	Fixed internal reviewers comments.
0.29	27.02.2017	Ana Respicio (FFCUL), Ignacio Robla (Atos), Elsa Prieto (Atos), Susana González Zarzosa (Atos), Zayani Dabbabi (Amadeus), Cagatay Turkay (City)	Fixed internal reviewers comments.
0.30	28.02.2017	Miruna Mironescu (Amadeus)	Fixed internal reviewers comments.
0.31	28.02.17	Alysson Bessani (FFCUL)	Minor formatting and typos removal

1. Introduction

Nowadays, in this all technological computer connected world, systems that guarantee security in computer transactions and technological environments play a fundamental role. This is basically the goal of the Security Information and Event Management (SIEM) systems. These systems trace events and correlate them to discover possible threats. Information of these events is obtained by means of different kind of sensors.

In order to improve present SIEMs, the DiSIEM project will develop some enhancements mainly focused on the following four subjects:

- 1)- to extend the types of events detected,
- 2) to collect information about security from open-source intelligence data available on the internet,
- 3) to generate new visualization plugins based on event information to ease the detections of threats,
- 4) to use storage clouds to store the long-term events information data.

This deliverable analyses several SIEM products with two purposes in mind, first how to extend with custom connectors and second how to create new event visualization tools. Future of SIEMs and factors that could affect their evolution are also treated.

This document is the result of Task 2.1 and provides the basis for the definition of the reference architecture (Task 2.2), that will be the generic SIEM model used throughout the project. This architecture will lead to an integration plan (Task 2.3) that will guide the integration activities that will take place within WP3, WP4, WP5 and WP6.

1.1 Organization of the Document

This document is divided into the following sections:

1. Introduction where goals of general project and this deliverable are described.
2. Features and capabilities of SIEMs to be considered in enhancements.
3. Description of several representative SIEMs considering features mentioned in Section 2.
4. Analysis of potential extensions to be developed.
5. Main factors affecting SIEMs in future.
6. Summary and conclusions of this study.

2 Analysis of SIEM features and capabilities

2.1 Introduction

Fundamentally, all SIEMs have the capacity to collect, store and correlate events generated by a managed infrastructure. Besides these key capacities, there are many differences between existing systems that normally reflect the different positions of SIEMs in the market.

In this chapter, we discuss the main criteria we use for comparing the different SIEMs.

2.2 Data sources supported

One of the key features of a SIEM system is the capacity of collecting events from multiple and diverse data sources in the managed infrastructure. For any SIEM analysed, we expect to list the main types of data sources natively supported by the system.

2.3 Data storage capabilities

Another fundamental feature of a SIEM system is the capacity to store and archive the collected events. Usually, two kinds of storage systems are supported: events storage and archival storage. For any analysed SIEM, it is important to know:

- Does the system differentiate standard storage from archival of old events?
- Where the data is stored (a relational database, a NoSQL database, direct to files, another specific technology)?
- How data is transferred between the two kinds of storage? Or, alternatively, what is the log rotation policy supported?
- Does the system support integration with cloud storage services such as Amazon S3?
- Is the storage horizontally scalable (i.e., if is it possible to increase the storage capacity and read/write performance by adding more machines to the system)?
- What are the data formats supported in the storage? This is especially relevant if one needs to fetch events directly from the storage.

2.4 Processing capabilities

The third fundamental capacity of a SIEM system is the correlation of collected and stored events through rules to a correlation engine. Traditionally SIEMs keep collected events on a database on a single server, therefore, their capacity of correlating events is limited by the number and complexity of rules to be run in such single server.

For each of the analysed SIEMs, it is important to understand how the data can be processed and correlated, which means:

- Does the system automatically index the events for reporting?
- Is it possible to create custom rules for generating alarms?
- Is it possible to scale the processing of events to several machines? In case it is possible, is it possible to place rules in different servers to create processing pipelines?

2.5 Flexibility in security directives

Another important feature of the system related with its processing capacities is how the rules, or security directives, can be expressed. There are many important questions that need to be addressed here:

- Does the system support a pre-configured set of rules and actions? How extensive it is?
- How user-defined rules are defined? Is there a custom language (SQL-like or something similar) for doing that? How expressive it is?
- Does the system support the integration with existing languages and platforms for creating advanced rules? For example: is it possible to define event-handlers in Java or Python or is it possible to integrate MapReduce jobs on the processing engine?

2.6 Behavioural analysis at application-level

More recent versions of leading SIEMs have been supporting extensive integration with application- and user-based anomaly detectors. These capabilities are mostly related with User and Entity Behaviour Analysis (UEBA), which include the analysis of the behaviour of employees, third-party contractors and other collaborators of the organization.

UEBA comprises the management of user/application profiles and the use of machine learning techniques for detecting misbehaviour. Concretely this usually implies the use of outlier detectors or classifiers.

It is important to understand exactly which kind of UEBA capabilities SIEMs have, and what are the gaps that DiSIEM can explore.

2.7 Risk analysis capacity

Similarly to UEBA, recent versions of leading SIEM systems already include some features for doing risk analysis on the assets of the managed infrastructure. Since risk analysis is a key contribution of DiSIEM it is important to understand what the capabilities of these tools are.

More precisely, it is important to identify if the analysed SIEMs natively support risk analysis or they can be integrated with external appliances for that purpose.

2.8 Exposed APIs

DiSIEM exploits a fundamental feature of existing SIEMs for improving their capabilities: extensibility. This means we need to precisely understand which kind of APIs are available in the analysed SIEMs. Some of the APIs we are interested are:

- Custom Connectors: how hard is to build new connectors for the system? Which languages are supported?
- The event storage components of the system provide some external APIs for reading, writing, and modifying events stored in the system.
- Does the system support the creation of new correlation engines? More specifically, is it possible to run a custom program for implementing non-standard security analysis of events (e.g., running MapReduce or Spark jobs)?

2.9 Resilience

Resilience or fault tolerance is an important feature of any critical monitoring system. It is important to understand what the fault tolerance capabilities of existing SIEMs are:

- Does the correlation engine support fault tolerance?
- How disaster recovery and replication are supported on the event storage?
- Do the connectors support some high availability features?

Notice that the resilience of these systems can be improved by using techniques on the operating-system level, but we want to know if the analysed SIEM natively support some high-availability features.

2.10 Security event management and visualization capabilities

It is well understood that one of the key factors that hinder the analysis of security events is the lack of support for proper data visualisation methods and little support for interactive exploration of the collected data is provided. Therefore, it is important to understand what the capabilities of the analysed systems in terms of the creation of new data visualisation methods and custom dashboards are. Some of the relevant questions are:

- Does the system support some standard techniques for creating dashboards (e.g., HTML5)?
- Is it possible to implement interactive data exploration methods?

2.11 Reaction capabilities

Traditionally, SIEMs support the creation of security directives for detecting suspect behaviour in the system and report alarms. However, these directives/rules could in principle, be used to trigger actions for modifying the

managed events (e.g., changing the configuration of firewalls or NIDS). What are the actions natively supported? How easy is it to express these actions to the correlation engine?

2.12 Simplicity of deployment and support provided

SIEMs are known for being difficult to deploy and manage. However, it is important to understand if the analysed system can be installed for testing with moderate effort.

2.13 Licensing

Is the analysed system open source or proprietary? If it is open source, which kind of license is the software subject to (e.g., GPL, BSD, Apache)? If proprietary, is it possible to have a free license for testing it?

2.14 Comparison between these criteria and others

In this section, we compare the criteria we defined for evaluating SIEMs with few others currently available.

2.14.1 Gartner Magic Quadrant

Gartner is a leading research and advisory company that provides information technology related insight about the most relevant products available in the market to the IT world, such as government agencies or business leaders. Gartner published two reports: one for 2016 Magic Quadrant (MQ) [1] and one for the 2016 Critical Capabilities for Security Information and Event Management (SIEM) [2].

The two axes composing Gartner's MQ are:

1. **Completeness of Vision:** including the following criteria:

- **Market Understanding:** The vendor's ability to understand buyers' needs and convert them into products and services.
- **Marketing Strategy:** Having a clear, differentiated set of messages consistently conveyed throughout the company and publicized on online channel.
- **Sales Strategy:** A strategy to sell products that uses an appropriate network and communication to extend the vendor's knowledge and capabilities (e.g. market reach, skills, and expertise, technologies, services and customer base).
- **Offering (Product) Strategy:** A vendor's approach to product development and delivery that points to the products current and future capabilities.
 - **Business Model:** The validity and logic of a vendor's underlying business proposition.
 - **Vertical/Industry Strategy:** The strategy of a vendor to meet the needs of market/industry segments and its way to manage its resources in this scope.

- Innovation: The use of the company's resource in order to prevent its acquisition.
- Geographic Strategy: Strategy to meet the geographical market's needs using all its means (e.g. thorough partners, subsidiaries, channels, etc).

2. **Ability to execute:** Including the following criteria:

- Product/Service: Core goods and services offered by the vendor that serve the market.
- Overall Viability: Includes an assessment of the vendor's overall financial health and its probability to continue to invest in the product.
- Sales Execution/Pricing: The vendor's capabilities in presales activities and the structure that supports them.
- Market Responsiveness and Track Record: Vendor's ability to be flexible with respect to market needs.
- Marketing Execution: The vendor's strategy to influence the market and increase brand/product awareness.
- Customer Experience: The clients' satisfaction and support.
- Operations: The vendor's ability to meet its goals and commitments as well as its quality.

The Gartner analysis, although quite complete, is more focused on market aspects, and especially relevant for decision makers. For DiSIEM, we need a more technical set of criteria to ensure a deep understanding of the features and extensibility of existing systems.

2.14.2 Other Criteria

Every comparison of SIEMs, include some type of criteria used for defining the best solutions for different users. In this section, we discuss two of the most interesting criteria found on these comparisons.

In an article to TechTarget, Karen Scarfone defined seven criteria for evaluating SIEMs [3], namely:

- Native support provided for the possible log sources.
- Supplementation of existing source logging capabilities.
- Support for threat intelligence.
- Availability of forensics capabilities.
- Features to assist in performing data examination and analysis.
- Quality of automated response capabilities, if available.
- Built-in reporting support.

These criteria are quite complete from the SIEM end-users perspective, however it does not address extensibility features as required in the DiSIEM project.

A second interesting set of criteria was defined used by InfoSec Nirvana to make a detailed comparison of the leaders of the Gartner magic quadrant [4]. The

criteria employed were based on a score, from one to five in terms of how well a system support the following features:

- Real-time security monitoring.
- Threat intelligence.
- Behaviour profiling.
- Data and end user monitoring.
- Application monitoring.
- Analytics.
- Log management and reporting.
- Deployment and support simplicity.

As it can be seen, their criteria are mainly feature-driven, and therefore many aspects important to the project were not considered.

We believe the set of criteria defined in this chapter, and employed in the next one are a superset of the criteria currently used in most SIEM evaluations.

3 Analysis of SIEM solutions

3.1 Introduction

Several companies have developed SIEM software products in order to detect network attacks and anomalies in an IT system. Among them, we can find classical IT companies (IBM, HP), others based on open source software (Alien Vault) or others coming from the antivirus area (McAfee).

As it has been introduced in previous chapter, Gartner reports are the most relevant reports about this subject which analyse the SIEM tools available in the market provided by the top 14 leading SIEM vendors. According to these reports, the products can be classified in four groups based on two main features, the ability to execute and the completeness of vision:

- **Leaders:** they provide products with good requirements behaviour and have the foresight for future requirements.
- **Challengers:** they provide products that do not comply general market requirements, execute well at present or may control a large segment, but do not demonstrate knowledge of future requirements.
- **Visionaries:** forecast trending of market, but do not yet execute well.
- **Niche Players:** they execute well in a particular segment of market but are unfocused and do not out-innovate or outperform others.

Figure 3-1: Gartner Magic Quadrant for SIEM 2016 shows position of SIEM products in 2016.

The features described in Section 2 of some of the most representative and most used SIEMs that appear in this Gartner quadrant will be analysed in following subsections. Some of these SIEMs are used for partners of this project, so analysis is based in a practical knowledge and experience. In particular, HP ArcSight is used by EDP, XL-SIEM is used by Atos and Splunk is used by Amadeus.



Figure 3-1: Gartner Magic Quadrant for SIEM 2016

3.2 HP ArcSight

ArcSight is the SIEM platform used at EDP. This platform encapsulates all features from a normal SIEM, consisting of three major components that create its architecture: the Connectors, the Loggers, and the Enterprise Security Management (ESM).

The ESM component, also known as the console or correlation engine, provides a graphical interface for the Security Operations Center (SOC) team. This module is responsible for event correlation, processing real-time rules and triggering alerts. It allows operators continuously monitoring data flows and ongoing security threats, providing visualization of alerts and the risk level across the organization. From this interface the SOC team monitors, analyses and deals with collected data, creating custom filters and rules. It also simplifies the creation of investigation queries, with the capability of having multiple threads to concurrently browse archived events. All the real-time events are constantly being updated and presented in elements known as dashboards, with several visualization options, such as charts, tables or network graphs.

The connectors are software components which collect events from sources, either through pull or push methods, normalizing them into a standard format – Common Event Format (CEF). The connectors are also responsible for filtering, aggregating and categorizing the events, which are then sent, in parallel, to the Logger and ESM components. This preprocessing is crucial to save bandwidth and improve the efficiency of the event collection process.

There are two types of connectors used at EDP: the more standard SmartConnectors, made available by the provider and already configured for a specific source technology, and the FlexConnector, which is adaptable for third-party devices, for a particular event source or another specific situation. For environments that include many legacy systems, such as the ones found at EDP, the FlexConnector is a valuable asset, being the only option to gather and parse information from older or customly developed applications.

The ArcSight Logger is an universal log management solution with the purpose of optimization. The extreme high event throughput, efficiency in the long-term storage and the rapid data analysis are the main benefits of using the Logger. This tool collects and stores huge amounts of logs, supports cyber security, IT operations and log analytics with quick searches and reports about the data or the investigated incidents. It also provides a web interface where its features can be used, and the security team member can analyse and investigate the events. Moreover, the Logger serves as a secure event storage for forensic analysis and, if need be, to present to authorities in case of legal procedures.

The ruleset and correlation processes can result in new events being generated as an aggregation of multiple original events. In those cases, the classification of the generated event is, at least, equal to the highest classification of the base events.

Figure 3-2 summarizes the ArcSight's architecture described above.

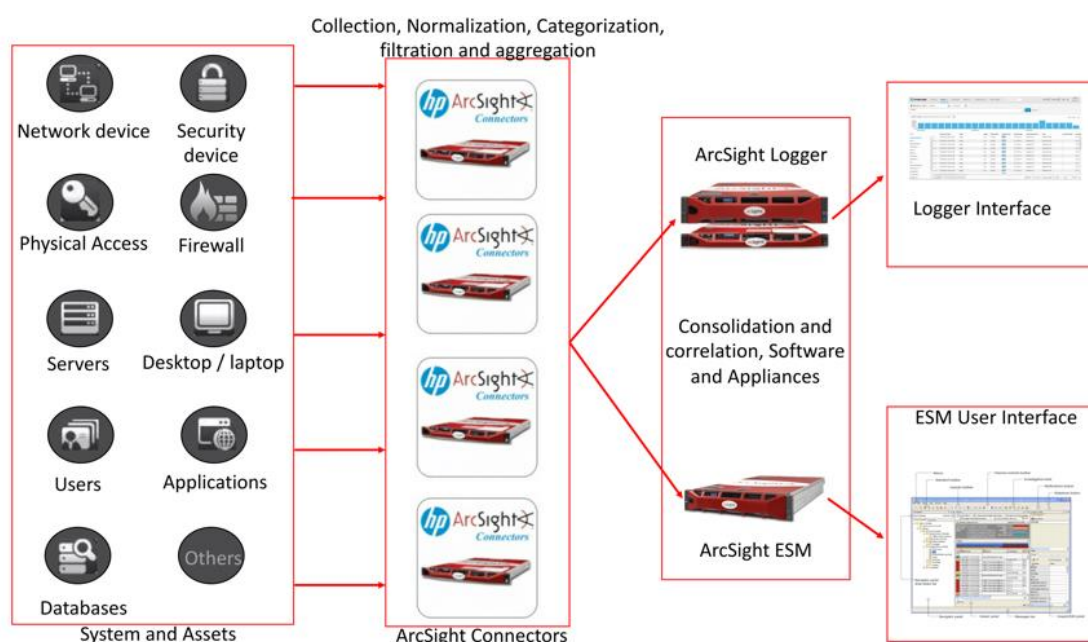


Figure 3-2: ArcSight architecture at EDP

Export Formats

It is possible to export events from either the Logger or ESM modules. The export process allows the user to select not only the events to be included but also the relevant event fields for the extraction. The events are exported in CSV format.

Shortcomings

ArcSight is viewed as more complex to deploy, configure, and operate than other leading solutions. The most commonly identified faults lie in the limited visualization options and intricate correlation rules. For instance, the available dashboards are rigid, not even allowing adequate window arrangement, which is critical for a Security Operations Center. The information associated with events is also immutable, with evident deficits when it comes to adapting the product to company processes and needs.

Thus, EDP, as many other ArcSight clients, uses third-party tools to process and display information. That information might be gathered directly from the Logger database or through simple exports.

ADD-ONS

The Add-on component is a set of applications or external commands that may help the correlation or investigation processes. These add-ons can be applied in different areas, such as user behaviour analysis or system management assistance. One example of practical applications of add-ons is the ArcSight ThreatDetector that, autonomously, identifies patterns of events uncovering zero-day vulnerabilities and advanced persistent threats. With the knowledge

acquired, it automatically creates new rules to the ArcSight ruleset, allowing identifying these threats patterns in the future. Although useful, add-ons result in added complexity and possible issues, especially when performing system updates.

3.2.1 Data sources supported

Although syslog is the commonly used protocol, ArcSight also supports event pulling/pushing from proprietary event sources, using specifically developed connectors.

The SmartConnector can access several types of specific data sources, such as XML, Microsoft SQL, CEF, among others. The FlexConnector does not depend on the type of the data sources. This connector can be customized to any kind of data source, only depending on the development capabilities to parse the intended data format.

3.2.2 Data storage capabilities

The connectors have local cache to allow event storage in case there are communication problems to the destinations. This temporary storage is very limited and depends on the used platform, which in EDP's case is normally a Windows Virtual with a 160GB hard drive. Both the Logger and ESM components use an internal database capable of dealing with large amounts of data.

EDP's event retention policy establishes 180 days for the Logger and 90 days for the ESM as retention periods.

3.2.3 Processing capabilities

ArcSight provides a set of configurable processing capabilities for each main appliance. We divide ArcSight processing capabilities into three categories, as discussed bellow.

Connector. The connector appliance contains the event processing and the destination processing. In the event processing, Arcsight performs four processing methods to the events:

1. Normalization, which parses all the security events to a common format. In this stage, events from multiple sources can be compared and correlated with valuable information;
2. Time correction, which allows the correction of time reported by the device, automatically;
3. Filtering and aggregation significantly decrease the amount of data received and increase data relevancy. A SmartConnector can process up to 5000 EPS;
4. Destination processing offers the possibility to configure batching options and a bandwidth control to send the processed events to the other appliances.

Logger. The Logger appliance contains all the raw data events, collected and processed by the connectors. It has a search engine, which allows the user to perform queries to all the events in a reasonable period, and extract the desired raw data. In EDP's platform, the Logger is processing around 3000 events per second.

Correlation Engine. The ArcSight correlation engine provides a user interface console. The console is processing approximately 2500 EPS in real-time and offers reports, alerts, dashboards and incident management capabilities. It is possible to create rules and/or filters to obtain only the events with meaningful information for future investigation. When displaying the events, a set of fields is displayed, for example the attacker and target address timestamp of the event, its priority, description, etc. The filters can use these fields to reduce the volume of the selected events.

3.2.4 Flexibility in security directives

Rules are a helpful tool for analysing and monitoring specific types of events. A rule is a programmed procedure that evaluates incoming events for specific conditions and patterns, triggering an action when the conditions are met. It can be created using other conditions expressed in filters, other rules and correlation data. Before using a rule, you can put it in a test phase to evaluate it with specific conditions. When a rule is activated, it runs on a live data stream and verifies if events match its conditions. The events which match the conditions of a rule are called correlation events, which can be associated with use cases and ultimately originate security incidents.

Filters are a set of conditions that focus on an event attribute. With filters, ArcSight reduces the number of events processed by the system. Filters also help analysing and monitoring some specific types of events, in the correlation with rules and data monitors. When the conditions of an Arcsight rule are triggered, a notification is created. Notifications help the monitoring efforts of the SOC team, containing the destination resource. The destination resource is the mechanism by which a security team member can send an individual or a user group in the organization a specific notification. The notification messages can be automatically delivered by e-mail, text message, or in the ArcSight Console.

The Dashboards display indicators that convey the security status of the organization. Dashboards aggregate individual data monitors in a variety of graphical and tabular formats. Usually a dashboard is fed by queries, lists and/or monitors. A query contains parameters, which act like filters and select the necessary information, and might have dependencies with other queries, adding complexity but also enabling more focused information.

3.2.5 Behavioural analysis at application-level

To help the investigation process, detect correlation between events and to apply some machine learning algorithm to the security information, it is necessary to send the data to a 3rd party platform. One way to do that is sending all the data with the smart connectors in the CEP syslog format using Raw TCP protocol. The

CEP syslog format is one of the available and, for EDP, is the format with more relevant information. To store all this information, EDP is using the Hadoop ecosystem. This architecture avoids loss of performance in HP Arcsight and enhances the flexibility to analyse the data.

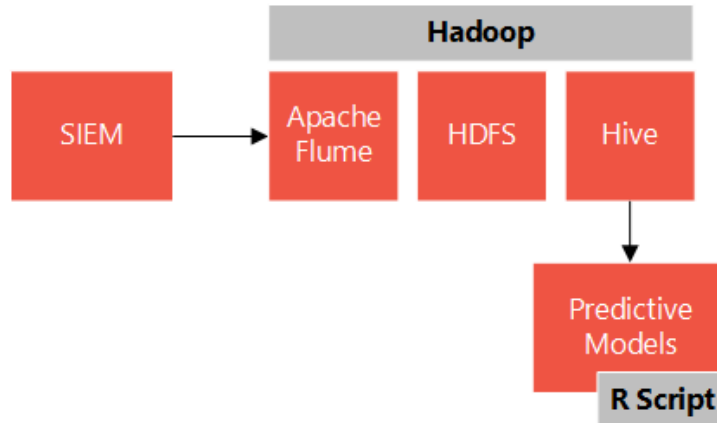


Figure 3-3: ArcSight Hadoop ecosystem

3.2.6 Risk analysis capacity

The Risk analysis approach in ArcSight is based on a per event threat detection generated through a priority formula, which indicates if an event has a higher or a lower priority. It is possible to customize this classification, adapting the automated process to better fit the client's reality. Accurate classification is crucial to select which events should be more carefully analysed.

We do not consider this to be a risk analysis capacity but a prioritization system to consider when correlating security events.

3.2.7 Exposed APIs

ESM Service Layer is a SOA platform and exposes ESM functionalities to web services through one of two options: Java service client API (ESM SDK, with SOAP client API, RESTful scripts, GWTRPC client API) or WSDL protocol. The process that consumes the service must retrieve an authentication token, through a login action, and use it to consume the services. This is a non-comprehensive list of the services provided by the ESM Service Layer through the API:

- ArchiveReportService
- DashboardService
- DataMonitorService
- FileResourceService
- UserResourceService

3.2.8 Resilience

ArcSight provides an Enterprise product with resilience at its core, therefore this is one of their strong suits. All components can be deployment in High Availability configuration, making sure that a single fault will not result in system failure.

3.2.9 Security event management and visualization capabilities

Visualization options include event search outputs, based on rows of events with visual cues related to the flow along a timeline as well as a colour scheme per the criticality of each event.

When considering processed events, the most common visualization options are graphs and charts. The interface is severely limited in terms of customization options, with the client having to adapt to the built-in option or using third-party add-ons.

3.2.10 Reaction capabilities

It is possible to use automatic triggers to perform some external actions, usually through a command line. Although possible, EDP has yet to configure automated actions outside the scope of the SIEM itself.

3.2.11 Deployment and support

ArcSight is a very comprehensive and complex platform with a long learning period. It is very easy to have quick wins, integrating events and providing meaningful outputs in less than a day. However, to fully configure the SIEM platform and explore incident response capabilities, a longer period is required. The quality of support teams is high, with very short response times and immediate expert availability.

3.2.12 Licensing

Commercial, based on number of instances and data flow (GB/day).

3.2.13 Position in Gartner Magic Quadrant

ArcSight is placed among the leaders in the Gartner Magic Quadrant. It has been awarded with this recognition since 2003, three years after its foundation.

In the latest, report Gartner listed both strengths and weaknesses found in ArcSight, with the positive points being:

- ArcSight ESM has the capacity to support a large scale SOC with well-defined workflows and a dedicated deployment management console;
- A User Behavior Analytics module that enables correlation between UBA and SIEM allowing the creation of new use cases;
- It is relatively easy to integrate new sources of logs thanks to the number of connectors and integrations available.

The weak points identified were:

- In its proposals, ArcSight usually include more professional services than other options. This means that there is a need to have Hewlett Packard Enterprise support services resulting in higher costs of implementation;

- The costs to deploy, configure and operate are higher than the other solutions;
- Although ArcSight is in the top four vendors, the trend is decreasing for new installs;
- HPE is rethinking the ArcSight core, and in that process it is possible that some features/functions may be deprecated, creating problems to customers currently using them.

3.3 IBM QRadar

IBM QRadar is an enterprise security information and event management (SIEM) product that can be deployed as a hardware, software or virtual appliance. IBM QRadar can also be deployed as a SaaS on the IBM cloud.

IBM QRadar is used by Amadeus Security Operations Center team. The current Amadeus High Availability [5] QRadar deployment is an on-premise deployment based on both physical and virtual appliances.

IBM QRadar SIEM ingests logs from different sources : Network devices, host assets and operating systems, application logs, etc. IBM QRadar SIEM also includes components for capturing and analysing network traffic.

A central component of the QRadar SIEM is the console which provides a user interface for real-time event and view, reports, offenses, asset information and for managing the product.

In addition to basic SIEM capabilities, IBM QRadar SIEM offers support for threat intelligence feeds and can optionally be extended with IBM Security X-Force Threat Intelligence [6] , a module for identifying malicious IP address, URLs , etc.

IBM QRadar SIEM is a part of the IBM QRadar Security Intelligence Platform [7] which has additional modules for risk management, vulnerability management, forensics analysis and incident response.

3.3.1 Data sources supported

IBM QRadar SIEM data sources supported can be split into three categories: Events, Flows and Vulnerability Assessment information. Below is a description for each data source category:

3.3.1.1 Event data collection

Events are generated by log sources such as firewalls, routers, servers, and intrusion detection systems (IDS) or intrusion prevention systems (IPS). Most log sources send information to QRadar SIEM by using the syslog protocol.

QRadar SIEM also supports the following protocols:

- JDBC
- JDBC - SiteProtector
- Sophos Enterprise Console - JDBC
- Juniper Networks NSM
- OPSEC/LEA
- SDEE
- SNMPv1
- SNMPv2
- SNMPv3
- Sourcefire Defense Center Estreamer
- Log File

- Microsoft Security Event Log
- Microsoft Security Event Log Custom
- Microsoft Exchange
- Microsoft DHCP
- Microsoft IIS
- EMC VMWare
- SMB Tail
- Oracle Database Listener
- Cisco Network Security Event Logging
- PCAP Syslog Combination Protocol
- Forwarded Protocol
- TLS Syslog Protocol
- Juniper Security Binary Log Collector Protocol
- UDP Multiline Syslog Protocol
- IBM Tivoli Endpoint Manager SOAP Protocol

Depending on the event log source, a manual configuration may be required for a proper event ingestion by QRadar SIEM.

3.3.1.2 Flow data collection

Flow Events provide information about network traffic and can be sent to QRadar SIEM in various formats, including Flowlog files, NetFlow, J-Flow, sFlow, and Packeteer. By accepting multiple flow formats simultaneously, QRadar SIEM can correlate flow events with log events to efficiently detect threats and suspicious activities that can be missed otherwise.

QRadar QFlow Collectors provide full application detection of network traffic regardless of the port on which the application is operating.

3.3.1.3 Vulnerability assessment (VA) information

QRadar SIEM can import VA information from various third-party scanners. VA information helps QRadar Risk Manager [8] identify active hosts, open ports, and potential vulnerabilities. QRadar Risk Manager uses VA information to rank the magnitude of offenses on a network.

Depending on the VA scanner type, QRadar Risk Manager can import scan results from the scanner server or can remotely start a scan.

3.3.2 Data storage capabilities / Processing capabilities

IBM QRadar SIEM is a modular. Depending on the scalability, requirements and the processing needed, different appliances can be added to meet the performance needed.

IBM offers several appliances to choose from with different categories and different processing and storage capabilities. Depending on the appliances acquired, processing and storage capabilities vary. IBM QRadar appliances can be divided into different categories:

- QFlow collectors

- Flow processors
- Event collectors
- Event processors
- Combined Event and Flow processors
- All-In-One appliances
- Console
- QRadar Log manager
- QRadar Vulnerability manager
- QRadar Risk manager
- Data node
- QRadar Incident Forensics
- QRadar Packet Capture

The description of the storage and processing capabilities of the QRadar appliances can be found in the official documentation [9].

3.3.3 Flexibility in security directives

In addition to the correlation capabilities, flexibility in defining security rules is essential for any SIEM solution. Rules in QRadar SIEM are applied to events, flows, or offenses to search for or detect anomalies. If all the conditions of a test are met, the rule generates response.

A set of default rules are shipped with QRadar console, those rules can be combined to create new rules using a simple syntax. Additional rules can also be downloaded from the **IBM Security App Exchange**. [10]

Four concepts are important to understand how rules work in QRadar:

- **CRE:** The Custom Rules Engine is used to define, manage and display rules and rule building blocks. The CRE provides information about how the rules are grouped, the types of tests that the rule performs, and the responses that each rule generates. The CRE keeps track of the systems that are involved in incidents, contributes events to offenses, and generates notifications.
- **Building blocks:** Used to build complex logic to define. Unlike rules, building blocks cannot trigger actions.
- **Rules:** A rule is a collection of tests or building blocks that triggers an action when specific conditions are met. Each rule can be configured to capture and respond to a specific event, sequence of events, flow sequence, or offense. The actions that can be triggered include sending an email or generating a syslog message.
- **Offenses:** As event and flow data passes through the CRE, it is correlated against the rules that are configured and an offense can be generated based on this correlation.

Rule types. There are types of rules in QRadar:

- **Event rules:** Test against incoming log source data that is processed in real time by the QRadar Event Processor. Event rules can be used to detect a single event or an event sequences. For example, to monitor network for unsuccessful login attempts, access multiple hosts, or a reconnaissance event followed by an exploit.
- **Flow rules:** Test against incoming log source data that is processed in real time by the QRadar Flow Processor.
- **Common rules:** Test against event and flow data.
- **Offense rules:** Test the parameters of an offense to trigger more responses. An offense rule processes offenses only when changes are made to the offense. For example, when new events are added, or the system scheduled the offense for reassessment. It is common for offense rules to email a notification as a response.

Domain-specific rules. If a rule has a domain test. Rules can be restricted so that it is applied only to events that are happening within a specified domain. An event that has a domain tag that is different from the domain that is set on, the rule does not trigger a response.

3.3.4 Behavioural analysis at application-level

User behaviour analytics in IBM QRadar is somewhat limited. In fact, two QRadar features are used that are based on user behaviour. The first is by using anomaly detection rules and the second is by leveraging the power of IBM QRadar User Behaviour Analytics app.

Anomaly Detection Rules. Anomaly detection rules can be further divided into Anomaly rules, Threshold rules and Behavioural rules.

Anomaly rules are used to test short-term event and flow traffic changes. By performing a comparison between the latest event and flow searches with existing longer time frame searches. An anomaly rule tests event and flow traffic for abnormal activity such as the existence of new or unknown traffic, which is traffic that suddenly ceases or a percentage change in the amount of time an object is active.

Threshold rules are used to test events and flows activity against a specified range. Expert knowledge is leveraged to set and tune thresholds to use. For example, a threshold rule can be used to detect bandwidth usage changes in applications or failed services.

Behavioural rules in QRadar are used to test events or flows for volume changes that occur in regular patterns to detect outliers.

A behaviour rule learns the rate or volume of a property over a pre-defined season. The season defines the baseline comparison timeline for the metric being evaluated. The longer that a behavioural rule runs, the more accurate it is over time. However, using more advanced anomaly detection models is not feasible with anomaly detection and behavioural rules provided by QRadar.

UBA App. IBM QRadar User Behaviour Analytics (UBA) is a module designed to provide early visibility to insider threats. This extension analyses the usage patterns of users inside an enterprise to determine their credentials have been compromised.

Machine Learning algorithms are used to detect anomalous user behaviours by creating a baseline of normal user behaviour and detecting significant deviations. The UBA app is shipped with a user-centric dashboard for monitoring user behaviours with associated QRadar incidents, events and flows. QRadar UBA App is focuses only on Insider Threat.

3.3.5 Risk analysis capacity

IBM QRadar SIEM has a risk analysis extension used to prioritize application vulnerabilities to reduce risk: QRadar Risk Manager. The extension has also a policy engine for automating compliance checks and comes with a risk dashboard.

QRadar Risk Manager monitors network topology and configuration, and correlates vulnerability data with events and flows to sense security risks.

3.3.6 Exposed APIs

QRadar Console offers a RESTful API to interact with QRadar. Sending HTTPS requests to URL endpoints can be done with any programming language that has an HTTP implementation.

The table below is a summary of QRadar REST API interfaces:

REST API	Description
/api/ariel	Query databases, searches, search IDs, and search results.
/api/auth	Log out and invalidate the current session.
/api/help	Returns a list of API capabilities.
/api/siem	Returns a list of all offenses.
/api/reference_data	View and manage reference data collections.
/api/qvm	Retrieves assets, vulnerabilities, networks, open services, networks, and filters.
/api/scanner	View, create, or start a remote scan that is related to a scan profile.
/api/asset_model	Returns a list of all assets in the model.

Table 1: QRadar REST API endpoints summary

QRadar API is not described in details in the IBM knowledge centre, links to API forum and code samples are provided though.

3.3.7 Resilience

Data resiliency for QRadar SIEM can be maintained only for a “High Availability” (HA) deployment. In the event of a hardware or network failure, HA ensures that QRadar continues to collect, store, and process data.

A HA QRadar deployment consists of duplicating all QRadar appliances used to have a primary and secondary component for each node. Disk Synchronization or shared external storage can be used to ensure that primary and secondary nodes have the same data. In case of failures, secondary nodes assume the responsibility of failed primary nodes.

Scenarios that may cause failures include:

- Network failure that is detected by network connectivity testing.
- Management interface failure on the primary HA host.
- Complete Redundant Array of Independent Disks (RAID) failure on the primary HA host.
- Power supply failure.
- Operating system malfunction that delays or stops the heartbeat ping.

QRadar HA does not protect against software errors. Failover occurs when the primary HA host experiences a failure, loses network connectivity, or if a manual failover is performed. During failover, the secondary HA host assumes the responsibilities of the primary HA host.

Failovers are triggered in case of once of the following scenarios:

- Primary Network Failures
- Primary Disk Failure
- Secondary HA Host Network or Disk Failure
- Manual Failovers

3.3.8 Security event management and visualization capabilities

Creating and managing dashboards in QRadar is simple. The default view in QRadar Console when logging in is a Dashboard tab. It provides a workspace environment that supports multiple dashboards used to display views of network security, activity, or data that is collected.

Dashboards allow organizing visualisation items into functional views, which enable to focus on specific areas of a network. Dashboards in QRadar are customizable and QRadar users can choose from default dashboards or create custom ones to investigate log or network activity.

Both reports and dashboards are using charts as building block. User can specify the chart types when building a dashboard or a report. Below is a list of chart types with a description.

Chart Type	Description
None	Used as a white space in dashboards and reports
Asset Vulnerabilities	Used to view vulnerability data for each defined
Connections	Used to view network connection information and trends
Device Rules	Used to view firewall rules and the event count of firewall rules
Device Unused Objects	Used to display object references of unused resources in a network. Ex. IP address, hostnames, etc.
Events/Logs	Used to view event information
Log Sources	Used to export or report on log sources.
Flows	Used to view flow information.
Top Destination IPs	Used to display the top destination IPs in a network location selected.
Top Offenses	Used to display the top offenses that occur at present time for a network location selected.
Offenses Over Time	Used to display offenses in a timeline.
Top Source IPs	Used to display and sort the top offense sources (IP addresses)
Vulnerabilities	Used to display vulnerabilities

Table 2: Chart types in QRadar

3.3.9 Reaction capabilities

Reports. QRadar SIEM provides default report templates that can be customized, rebranded, and distributed to QRadar SIEM users. Report templates are grouped into report types, such as compliance, device, executive, and network reports. The reporting interface in the QRadar console allows the following functionalities:

- Create, distribute, and manage reports for QRadar SIEM data.
- Create customized reports for operational and executive use.
- Combine security and network information into a single report.
- Use or edit preinstalled report templates.

- Brand reports with customized logos. Branding is beneficial for distributing reports to different audiences.
- Set a schedule for generating both custom and default reports.
- Publish reports in various formats.

A report can consist of several data elements and can represent network and security data in a variety of styles, such as tables, line charts, pie charts, and bar charts.

Alerting in QRadar. As previously mentioned, QRadar rules generate alerts with a configurable action to assign to created rules. Both default and custom actions can be defined and QRadar. For instance, it is possible to configure an email server to distribute alerts, reports, notifications, and event messages.

3.3.10 Deployment and support

IBM Security QRadar architecture supports deployments of varying sizes and topologies, from a single host deployment, where all the software components run on a single system, to multiple hosts, where appliances such as Event Collectors, and Flow Collectors, Data Nodes, Event Processors, and Flow Processors, have specific roles.

The following diagram shows the QRadar components that can be used to collect, process, and store event and flow data. An All-in-One appliance includes the data collection, processing, storage, monitoring, searching, reporting, and offense management capabilities.

The Event Collector collects event data from log sources in a network, and then sends the event data to the Event Processor. The Flow Collector collects flow data from network devices such as a switch SPAN port, and then sends the data to the Flow Processor. Both processors process the data from the collectors and provide data to the QRadar Console. The processor appliances can store data but they can also use the Data Nodes to store data. The QRadar Console appliance is used for monitoring, data searches, reporting, offense management, and administration of the QRadar deployment.

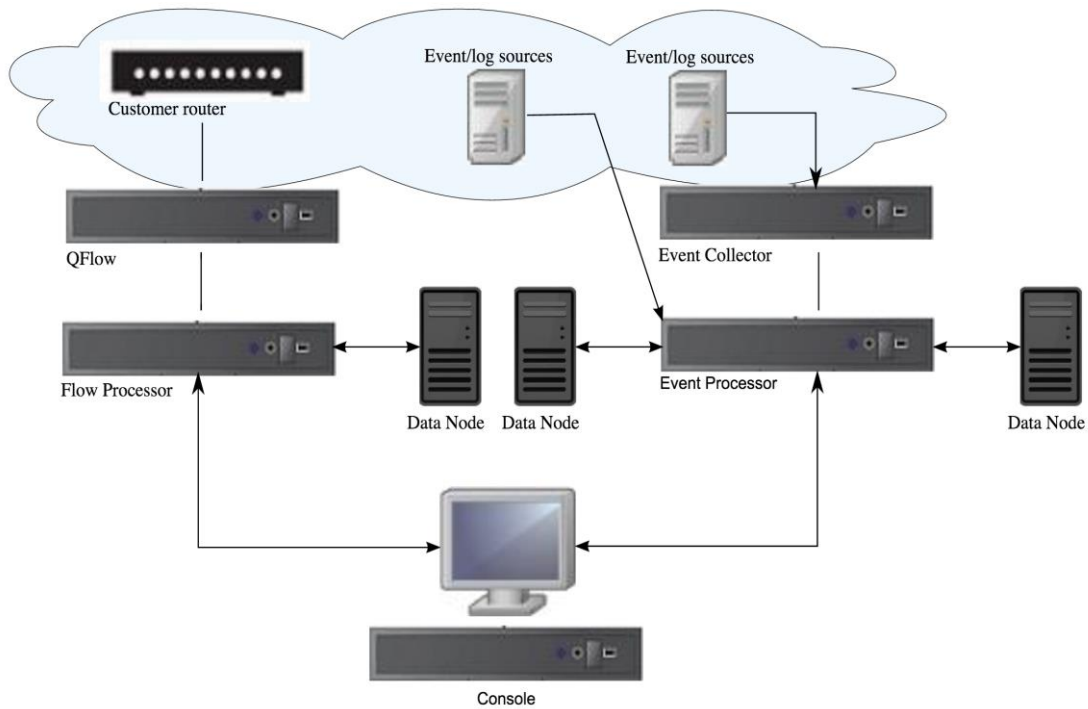


Figure 3-4: IBM QRadar event and flow components

QRadar console comes with a deployment helper “Deployment Editor” to add and configure deployment components.

The topology and composition of a QRadar deployment are influenced by the capability and capacity of that deployment to collect, process, and store all the data that to analyse.

A deployment can be geographically distributed but might be impacted by intermittent or poor connectivity to remote data centres. The figure below depicts a geographically distributed QRadar deployment.

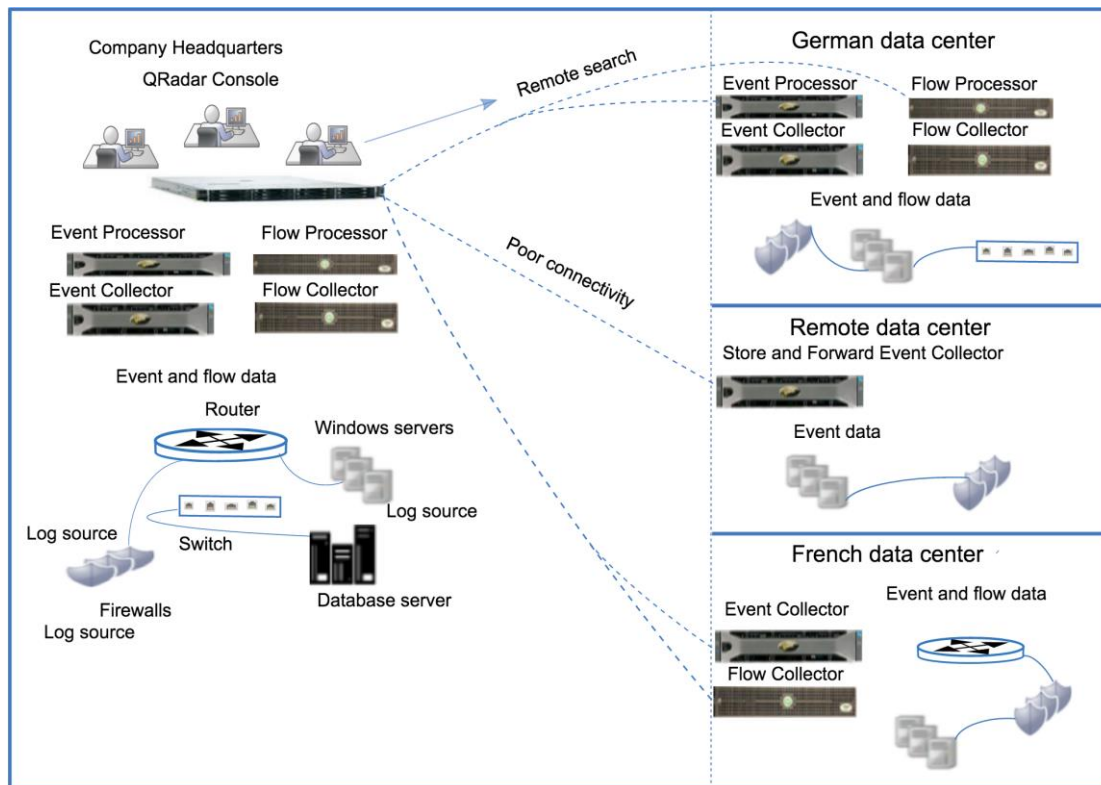


Figure 3-5: IBM QRadar Geographically Distributed Deployment

Many support resources are provided for QRadar customers. Depending on the issue/request customers can choose from the support channels below:

- QRadar Customer Forum (to ask support a question)
- QRadar Knowledge Center (for product documentation)
- Support Knowledgebase (QRadar Support - all published technical notes)
- QRadar Support Portal (Official Support Portal)
- IBM Fix Central for QRadar (Download software for QRadar)
- Master Software Version List (List of all software versions & release notes)
- IBM Security QRadar Request for Enhancement (Feature Requests & Enhancements)
- IBM Security QRadar software and appliances (Support lifecycle)
- IBM My Notifications (Information emails / RSS feeds)
- QRadar Support Videos (Review playlists for QRadar content)
- IBM Security Support Twitter (QRadar Twitter account)

3.3.11 Licensing

Because IBM QRadar SIEM is a modular product with multiple options per component, licensing and pricing is not publicly available and is usually depends on the agreement between IBM and its customers. However, the charge metric is generally based on usage such as log source events per second and network flows per minute. Organizations interested in better understanding the options can get the latest pricing information for all the available IBM QRadar SIEM licenses by contacting a QRadar commercial representative.

3.3.12 Position in Gartner Magic Quadrant

QRadar is present in Gartner's 2016 SIEM review (see Figure 3-1), and it is highly ranked in terms of Basic Security Monitoring (3rd position), Advanced Threat Detection (3rd position) and Forensic and Incident Response (2nd position). According to Gartner, QRadar is a multi-feature security-monitoring platform providing log management, NetFlow, SIEM and many more such as application monitoring, full packet capture, vulnerability scanning and risk analysis.

Midsized/large enterprises with general SIEM requirements as well as organisations looking for a single security event monitoring and response platform for their SOC should take QRadar into consideration. Also, if the midsized company is looking for a solution with flexible implementation, hosting and monitoring options, might also consider QRadar.

In Gartner's SIEM report, Gartner enumerates the strengths as well as weaknesses notes when dealing with QRadar, as follows:

- Strengths:
 - QRadar offers an integrated view of log and event data, network flow, vulnerability and asset data but also threat intelligence.
 - Network traffic behaviour analysis can also be correlated across log events and NetFlow.
 - The modular architecture supports security events and log monitoring in IaaS environments.
 - It is also relatively easy to deploy and maintain, no matter if it is an all-in-one application or a large-tiered, multisite environment.
- Weaknesses:
 - Endpoint monitoring for threat detection and response is rather a weakness
 - Gartner clients report mixed success with the integration of IBM vulnerability management add-on on QRadar.
 - Sales negotiations with IBM require persistence and can become very complex.

3.4 Intel McAfee Enterprise Security Manager

The McAfee SIEM solution is composed of several appliance-based platforms working in a coordinated way to offer fundamental information to enterprise security professionals within an enterprise. Different types of configurations allow for scalable and versatile SIEM architecture available, delivering real-time forensics, comprehensive application and database traffic/content monitoring, advanced rule- and risk-based correlation for real-time as well as historical incident detection and the most complete set of compliance features of any SIEM on the market. Every configuration is available in a range of physical and virtual models. Following image shows SIEM architecture¹:

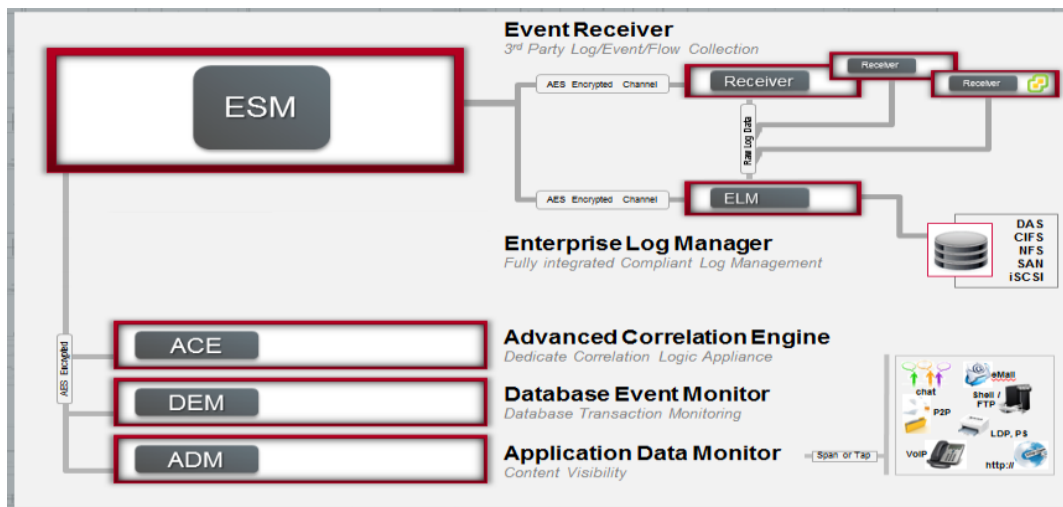


Figure 3-6: McAfee SIEM architecture

Basic McAfee SIEM includes the following components: Enterprise Security Manager (ESM), Event Receiver and Enterprise Log Manager (ELM). They can be deployed in a distributed way, each component in a different machine, for large systems or in an all-in-one installation for small systems.

Available SIEM components provided by McAfee are described as follows:

ESM – Enterprise Security Manager (sometimes referred to as ETM): This is the SIEM central console, the core of the McAfee SIEM solution. It includes the master database and the web interface. It is powered by the McAfee EDB proprietary embedded database.

REC – Event Receiver (sometimes referred to as ERC): The McAfee REC collects events, flows and logs from data sources. Receivers can also aggregate some information and perform rules-based event correlation.

ELM – Enterprise Log Manager: The McAfee ELM stores the raw event/log data collected from data sources configured on Event Receivers.

¹ <https://community.mcafee.com/docs/DOC-6207>

ACE – Advanced Correlation Engine: The ACE provides the SIEM rules-based and risk-based correlation capabilities. ACE also performs deviation, and historical correlation.

ADM – Application Data Monitor (sometimes referred to as APM): ADM analyses layer 7 (application) traffic flows in order to trace transmission of important data as well as detect malicious traffic, theft or misuse of credentials and application-layer attacks.

DEM – Database Event Monitor (sometimes referred to as DBM): DEM monitors in real-time network packages related to database transactions, so there is no need to install any component on databases to monitor them.

3.4.1 Data sources supported

Over 250 different parsers, each of them associated to a specific data source type, are provided with McAfee Event Receivers to integrate collected events in the SIEM. Additionally, the user has the possibility to create custom parsers with specific regular expressions to normalize events generated by new data sources.

The following common formats are supported by McAfee SIEM:

- Syslog (using UDP or TCP),
- Windows Management Instrumentation (WMI),
- Netflow: this includes generic Cisco Netflow format, sFlow (or “sampled flow”, an industry standard for exporting truncated packets at layer 2 of the OSIM model), IPFIX (Internet Protocol Flow Information Export) and Jflow (format used by Juniper manufactured routers and switches),
- Common Event Format (CEF) and
- Standard Event Format (SEF).

Apart from those formats, the own McAfee Event Format (MEF) is supported. This format is the one used by the McAfee SIEM Collector, a host-based software to send events from the McAfee Event Receiver to the McAfee ESM main component. The SIEM Collector can be also configured to collect and send events from a local or remote Window machine to the Event Receiver.

McAfee SIEM also supports traffic relayed through a Splunk server using Syslog format or through an ArcSight server in CEF format. In both cases, it is required to define data sources for each individual data source that is being relayed from these servers.

Additionally, McAfee ESM supports two-way integration with Hadoop to connect data with an existing Cloudera’s Hadoop cluster and support to integrate threat intelligence feeds such as the one provided by McAfee Global Threat Intelligence² (GTI) service.

3.4.2 Data storage capabilities

² <http://www.mcafee.com/us/resources/solution-briefs/sb-operationalizing-threat-intelligence.pdf>

Enterprise Log Manager is the McAfee SIEM component responsible of storage capabilities. It collects and stores log files compressing them for disk space optimization, although it also allows the storage in its original format to support chain of custody and non-repudiation for compliance needs. ELM also signs logs to ensure authenticity and integrity for regulatory compliance. Besides collection and storage, ELM also provides analysis and search functions.

Log storage can be done locally, using a Direct Attached Storage (DAS) attached to the Enterprise Security Manager, or via a managed storage area network (SAN). Storage in ELM is done using a concept called “**Storage Pools**” which are virtual groups of usable storage devices. Data is assigned to a specific pool based on the source device and considering different criteria (e.g. relevance to security, compliance, confidentiality...). Each pool can be composed by multiple physical storage devices (local storage, Network File System, SAN, etc.) and can be distributed in separate locations depending on the specific log management needs. For example, for critical data some redundant solution could be used, whereas when it is required a faster access it could be better a local storage.

In particular, McAfee offers its own high performance storage solution called McAfee Direct Attached Storage. There are different options depending on the array capacity (e.g. 100 TB). All of them include architecture with RAID controller, mirrored cache and Input/Output multi-pathing.

3.4.3 Processing capabilities

Though McAfee ESM has capabilities of event correlation, they are too basic and installation of McAfee Advanced Correlation Engine (ACE) component is almost mandatory.

The McAfee Advanced Correlation Engine complements McAfee Enterprise Security Manager with two dedicated correlation engines that can be run simultaneously:

1. Risk detection engine

This correlation engine provides a *rule-less* risk assessment for specific assets selected by the user and based on:

- Asset information and context
- Vulnerability information related to the specific asset
- Historical event activity detected involving the asset

2. Threat detection engine

This is the traditional type of event correlation engine to trigger alarms based on predefined rules.

The McAfee Advanced Correlation Engine solution supports not only correlation in real-time but also historical correlation. This last option can be useful for example to check after a zero-day attack detection, if the organization has been exposed to it in the past.

In general, production SIEM installations consist of two appliances, one for real-time risk and threat detection correlation and another for historical risk and threat detection correlation.

Usually, McAfee Advanced Correlation component is deployed stand-alone but it has the capacity to scale using its data engine.

In environments where no ACE component is available, it is also possible to process the events and perform a correlation based on rules directly on the Event Receiver components. However, in this case this correlation processing implies a reduction in the performance of about 20% on a typical receiver. Besides, not any type of correlation can be done in the receivers. In particular, it is not supported: flow-based correlation, risk-based correlation, deviation-based correlation and historical correlation.

On the other hand, the Enterprise Log Manager (ELM) component supports full-text index (FTI) for event detail information.

3.4.4 Flexibility in security directives

This SIEM includes hundreds of predefined event correlation rules that can be customized by means of embedded parameters and global variables (e.g. HOME_NET to represent the organization internal IP ranges) contained in the rules. An easy way to create rules is using a GUI event correlation rule editor included in ESM. This editor can also be used to customize an existing rule. For rule-less correlation a configuration editor is provided. These correlation rules can be applied for any supported data source.

To create a new rule, the user must first select the category it belongs to from a set of available major categories (such as IPS, for intrusion detection and prevention rules). There is not a custom language to be used by the user to define security directives. For each rule, different options and values can be selected from the list offered to the user depending on the category and the fields included in the different events. For example, to add a rule related to the firewall, the user could set up source and destination ports, source or destination ip addresses, the protocol, the severity of the event, etc.

3.4.5 Behavioural analysis at application-level

There are several ways to integrate McAfee Enterprise Security Manager with a User Behavior Analysis Solution. McAfee itself provides McAfee Network User Behavior Analysis Monitor. McAfee Network User Behavior Analysis (Securify) Monitor performs two types of real-time analysis of users' behaviour taken place in the network:

- analysis based on monitoring flow data provided by Cisco Netflow and Juniper J-Flow,
- use of native deep packet inspection (DPI) to analyse the traffic previously captured and decoded.

On the other hand, it is possible to integrate these capabilities to the McAfee ESM with the use of third-party tools. For instance, the company Niara³ has recently (October 2016) certified its developed machine learning-based user and entity behaviour analytics (UEBA) product to interoperate with McAfee Enterprise Security Manager.

3.4.6 Risk analysis capacity

McAfee Advanced Correlation Engine is the component that adds risk analysis capabilities to McAfee Enterprise Security Manager. It allows identification and assessment in real-time of incoming events from two points of view:

- **Rule-based logic**

This is the risk analysis traditional performed in SIEMs to correlate logs, events and network together with contextual information (e.g. identity, roles, vulnerabilities...) to detect threat patterns.

This type of analysis is the one already supported natively on McAfee Enterprise Security Manager. However, the use of McAfee Advanced Correlation Engine provides a dedicated processing to support larger volumes of data.

The main constraint to this type of analysis is that it allows only detection of known threat patterns and signatures, so to be effective it requires having them constantly updated. To solve this, McAfee provides the McAfee Global Threat Intelligence service that can be integrated in its products. This service allows receiving in real-time updated cloud-based threat vectors and reputation data.

- **Risk-based logic**

In this case, the risk analysis is done without the use of rules in a traditional sense. The user only needs to set up which users, groups applications, servers or networks are sensitive in the monitored infrastructure. The McAfee Advance Correlation Engine models the organization risks by scoring attributes for those selected assets. When a risk score exceeds a predefined normal threshold, it will be notified to the user. Furthermore, these risk scores can be also used within the rule-based logic and in the other sense, it can be incorporated a risk factor based on an event triggered by a traditional rule (e.g. to increase a 20% the risk score after a brute-force login attack).

3.4.7 Exposed APIs

Apart from the web interface, there are two ways of invoking functionality in the McAfee Enterprise Security Manager component: using the SOAP/XML based API or its corresponding RESTful form.

Functionalities provided by these APIs can be divided in five main groups:

³ www.niara.com

- 1) **Queries:** different commands are provided related to queries, e.g. to get all queries executed, customize some field and execute a specific query.
- 2) **Watchlists:** McAfee ESM uses what it is call a watchlist to integrate threat feeds in the SIEM workflow. The API includes commands to add, remove or edit values to a watchlist.
- 3) **Users:** to add, remove or edit users or user access groups.
- 4) **Alarms:** with different commands to manage alarms, e.g. to retrieve a list of alarms that have been triggered but have not been acknowledged, to delete an alarm or to mark a triggered alarm as acknowledged.
- 5) **Data Sources:** adding, removing and updating data source properties and listing them. It is also possible to get details on a specific data source or get a list with the devices defined in the system.

Additionally, without the use of APIs and thanks to McAfee Data Exchange Layer (DXL)⁴ application framework and the Open Data Exchange Layer (OpenDXL) SDK, it is also possible to connect and share bi-directional threat information in real-time between the McAfee SIEM components and third-party solutions. Some Innovation Alliance⁵ partners have already integrated their products with this DXL framework.

3.4.8 Resilience

No information available about resilience or fault tolerance provided by McAfee Enterprise Security Manager solution.

3.4.9 Security event management and visualization capabilities

Features about McAfee Enterprise Security Manager event management and visualization can be summarized in the following points:

- Reports: they include real time and historical view of the enterprise security threats.
- Dynamic filters and drill down: selected security data is shown through single platform dashboards via drill downs and easily customizable views. Filters can be established by data source, by time and even customizing the set of fields the filter applies.
- Asset, threat and risk views: it is possible to consolidate known vulnerabilities, external threats and counter measure data into single view for fast threat risk assessment.

Below it is shown an example of McAfee Enterprise Security Manager web graphical interface:

⁴ <http://www.mcafee.com/us/solutions/data-exchange-layer.aspx>

⁵ <http://innovationalliance.net/>

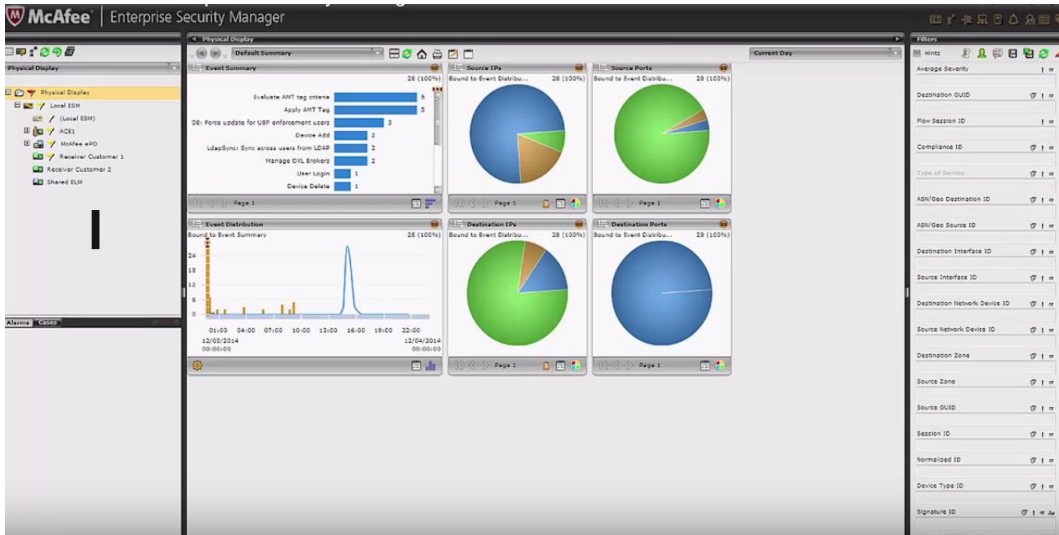


Figure 3-7: McAfee Enterprise Security Manager dashboard⁶

The ESM interface is composed of several distinct panes⁷:

- Event Summary: Shows most significant malicious events detected.
- Source IPs: Shows events source IP addresses.
- Total Events: Shows generic counts about events in the environment.
- Event Distribution: Shows graphs of event counts over time.
- Destination Geolocation: Shows event location in a concrete interval.

The dashboard is customizable in the sense that user can create a new view/dashboard element. User has also the possibility of create flexible filters to see and work with specific data from the view (right-hand side of the interface).

3.4.10 Reaction capabilities

McAfee Enterprise Security Manager allows execution of external actions by the invocation of an URL. It is known as **URL action** and it is used for example to perform lookups on data elements such as IP addresses, domain, file hashes, etc. For example, if you need to check a suspicious IP, i.e. 173.35.7.9, using service provided by the website IPVoid⁸, the following URL can be invoked <http://www.ipvoid.com/scan/173.35.7.9>

However, to provide automatic reaction when some alarm is detected it is necessary to use additional solutions. One example is McAfee Active Response⁹ (also included in the McAfee portfolio product McAfee ePolicy Orchestrator¹⁰), which provides preconfigured and customized actions to be triggered when a specific condition or set of conditions are fulfilled.

⁶ <http://www.mcafee.com/sg/resources/solution-briefs/sb-esm-and-threat-intelligence-exchange.pdf>

⁷ <https://www.sans.org/reading-room/whitepapers/analyst/security-intelligence-action-review-mcafee-enterprise-security-manager-esm-92-35095>

⁸ <http://www.ipvoid.com/>

⁹ <http://www.mcafee.com/es/resources/data-sheets/ds-active-response.pdf>

¹⁰ <http://www.mcafee.com/es/resources/data-sheets/ds-epolicy-orchestrator.pdf>

3.4.11 Deployment and support

McAfee assures that an ESM simple deployment can be set up in as little as one to two days to provide customers visibility into what is happening in their infrastructure.

McAfee provides through its web portal a Knowledge Center, support tools such as McAfee Virtual Technician and other diagnostic tools to help with problems, a downloading page to get patches and hotfixes, and a customer online community. They also offer a McAfee Customer Service where the user can call or write an email to get support. However, according to Gartner report, users highlight a lack of satisfaction with technical support.

3.4.12 Licensing

Commercial, where the user can choose between two main types of McAfee licensing: subscription licenses (for a specific period of time) and perpetual licenses (for customers preferring indefinite software license terms).

3.4.13 Position in Gartner Magic Quadrant

McAfee solution SIEM is placed among leaders in Garner Magic Quadrant. According to Gartner analysis the strengths of this SIEM are:

- Good integration with other McAfee security applications through McAfee ePolicy Orchestrator (ePO).
- Good coverage of operational technology and supervisory control and data acquisition (SCADA) devices.
- Intel Security's McAfee Data Exchange Layer (DXL) allows product integration without using APIs.

In the side of weaknesses, the report enumerates the following:

- Advanced capabilities in certain areas involve integration with other Intel portfolio products.
- No predictive analytics and other built-in features not as strong as those of leading competitors.
- Poor stability and performance is reported in the last 12 months.
- Customers are not satisfied with technical support.
- User comments about displacement of service are growing.

3.5 Alienvault OSSIM and USM

OSSIM, AlienVault's [11] Open Source Security Information and Event Management (SIEM), was one of the first open source SIEM products available in the market and has become one of the most widely used.

In this section, the open source version together with the improvements added by AlienVault to its Unified Security Management (USM) commercial version will be analysed.

OSSIM architecture is composed of two main elements: sensors and server (see Figure 3-8). The OSSIM server component includes not only the core functionality of the SIEM to receive and process the events but also the framework to manage the web graphical interface and the database management.

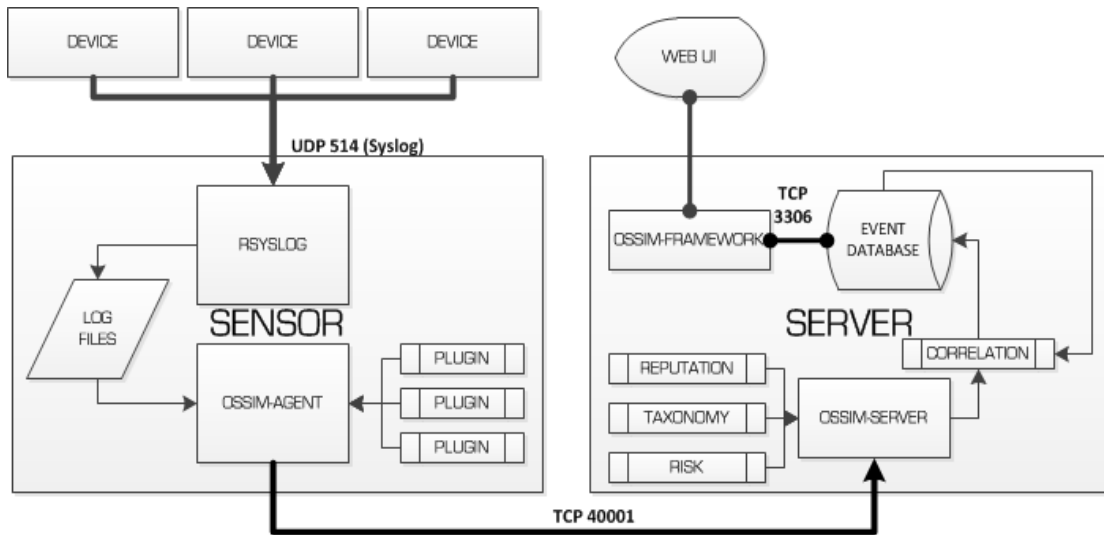


Figure 3-8 : Alienvault OSSIM Architecture [12]

USM solution follows a three-tier architecture integrating into a single system the following components [13]:

Alienvault USM sensor: One or several USM sensors are deployed throughout the monitored network infrastructure to gather information generated by different devices. Data collected is normalized to a common format and sent to Alienvault USM server for its processing.

Alienvault USM server: This is the core SIEM component in charge of aggregation and correlation of data collected by the different deployed USM sensors to detect potential threats and security incidents. It also includes a web-based graphical interface for administration, reporting and security event management.

Alienvault USM logger: This USM logger component is the one included in AlienVault commercial version to store raw logs collected by USM sensors to deal with forensic storage and compliance reporting.

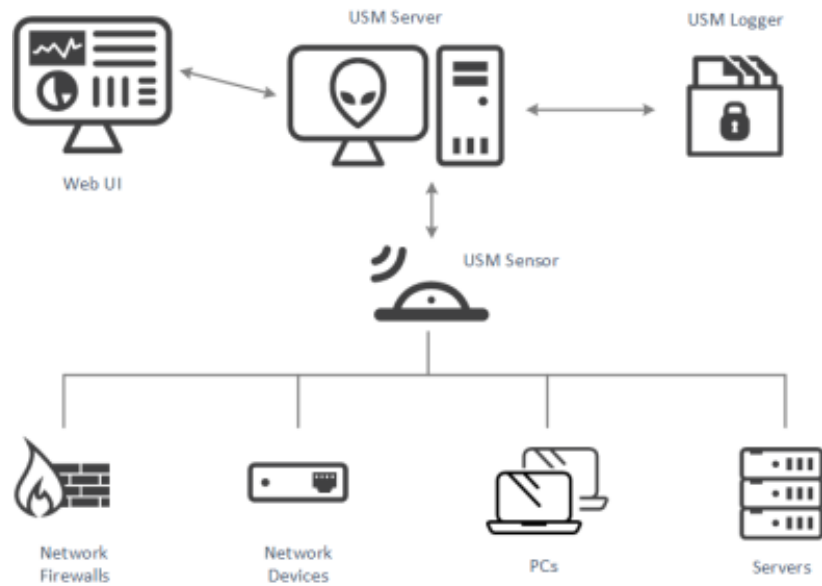


Figure 3-9 : Alienvault USM Architecture [13]

3.5.1 Data sources supported

Collection of data is done via agents running on the OSSIM/USM sensors installed throughout the monitored infrastructure. These agents include what it is called **plugins**, which define how to obtain and parse information generated by the different devices to transform it into security events in a normalized format. These plugins are based on regular expressions and make use of two different values to identify the data sources and event types or metrics by each data source which are supported:

- *plugin ID*: it identifies a type of data source. For instance, a Cisco firewall or a Snare Windows agent,
- *plugin SID*: it identifies specific event types for a data source (plugin ID)

Alienvault OSSIM/USM agents support two types of plugins:

- **Detector plugins:**
These plugins works in a passive way, reading and extracting events from log files which can be locally generated (using syslog or directly written by different devices or data sources) or received from remote systems.

This type of plugins is the most common and OSSIM/USM distribution includes a set of preconfigured plugins¹¹ which includes common security tools such as firewalls (e.g. Cisco or Juniper), network intrusion detection systems (NIDS) such as snort or suricata, host intrusion detection systems (HIDS) such as ossec, antivirus (e.g. panda), infrastructure monitoring tools (e.g. 48nalyz), etc. These plugins can be tuned by the user or new ones added.

¹¹ The complete list of available plugins for USM (as for 27 December 2016) can be found at <https://www.alienvault.com/docs/data-sheets/usm-plugins-list.pdf>

Data sources supported by OSSIM/USM using these detector plugins can be classified in the following groups that will be described below:

Data sources	Description
Database	Monitors the content of external databases.
Log	Monitors a file, usually receiving data through syslog.
Remote Logs	Monitors a file in a remote appliance.
Security Device Event Exchange (SDEE)	Monitors Cisco devices, using SDEE protocol.
Windows Management Instrumentation (WMI)	Remotely connects to Microsoft Windows events and data without an agent.

Table 3: Data Sources supported by Alienvault SIEMs

- **Monitor plugins:**

These plugins work in an active way, being invoked by the SIEM server to send a query to a specific tool or target monitored and collect the reply received. Some examples of tools that can be invoked are: nmap, ping, whois, and tcptrack. These monitor plugins also generate text logs to be consumed by detector plugins like standard logs.

3.5.1.1 External databases

Data can be extracted from external databases using queries included in detector plugins and the result is transformed into normalized events compatible with OSSIM/USM server. In the own plugin it can be also configured a duration between queries e.g. to read a value periodically. Supported databases are MySQL and MS SQL Server.

3.5.1.2 Logs

Contents included in log files generated by different data sources or by Syslog can be processed by detector plugins to generate events from them. In this case, events are extracted by matching each line in the log file according to a regular expression defined in the plugin. The original text in a log entry is then normalized to data fields in the security event.

3.5.1.3 Remote Logs

Log files generated by remote appliances can be also recovered using SSH or Syslog protocol, through specific detector plugins.

3.5.1.4 Security Device Event Exchange (SDEE)

AlienVault supports log collection using the Security Device Event Exchange (SDEE) protocol. This protocol specifies the format of messages used to communicate events generated by security devices.

USM supports the collection of events specifically from the following Cisco devices:

- Cisco Network Prevention Systems (IPS)
- Cisco Network Detection Systems (IDS)
- Cisco Switch IDS
- Cisco IOS routers with the Inline Intrusion Prevention System (IPS) functions
- Cisco IDS modules for routers
- Cisco PIX Firewalls
- Cisco Catalyst 6500 Series firewall service modules (FWSMs)
- Management Center for Cisco Security Agents
- CiscoWorks Monitoring Center for Security

3.5.1.5 Windows Management Instrumentation (WMI)

Alienvault supports collection of Microsoft Windows events and data remotely without an agent, using the Windows Management Instrumentation Command Line (WMIC) from a specific detector plugin.

However, currently WMIC does not support samba4/NTLMv2 authentication. This means that on case of using a WMI plugin in AlienVault to recover events from a Windows host using by default NTLMv2, it must be previously enabled manually NTLMv1 authentication.

3.5.2 Data storage capabilities

Both AlienVault's SIEM solutions include SQL database storage capabilities in the server. The incoming SIEM events (once normalized) are stored in a database, together with the generated alarms and other configuration data used at run-time. Through the graphical interface, the user can configure backup and storage thresholds and enable when to remove automatically alarms or logs from the database after a predefined expiration timeout. Besides, USM database add the capability to have asset inventory storage and continuous data (e.g. netflow) storage.

Apart from that, the USM commercial version also provides the possibility to forward raw logs (with the normalized log data) to an additional USM Logger server for remote storage and to reduce the load on the server. USM Logger is designed for long-term storage with the purpose to allow forensic analysis of events and support compliance requirements. Other features supported by USM Logger are:

- It includes indexing logs capabilities for full-text searches
- Support to sign cryptographically logs
- Data compression ratio 5:1
- Central retention policies to enforce corporate or regulatory data retention requirements.

In the case of Alienvault USM for AWS [14], it is also supported the use of Amazon S3 or Amazon CloudWatch logs to gather data for forensic storage.

3.5.3 Processing capabilities

OSSIM/USM supports three different types of correlation:

1. **Logical Correlation:** correlation between events from different sources based on the security directive rules specified. More details about these rules will be provided in next Section 3.5.4.
2. **Cross-Correlation:** in this type of correlation, it is checked if the IP destination address included in the incoming Network Intrusion Detection System (NIDS) events has some vulnerability defined and stored in the database. In case some vulnerability is found, the reliability of that event is increased to 10 (the maximum). For this correlation, it is required to have enabled some vulnerability scanner (such as Nessus or OpenVAS) and perform regular scans on the monitored hosts. Besides, it only is performed on events with destination IP address defined. More details about cross-correlation rules will be provided in next Section 3.5.4.
3. **Inventory Correlation:** correlation between events and destination characteristics to reduce the number of false positives modifying the reliability of the events. The features checked are operating system, port, protocol, service name and service version.

According to tests performed by [15] and the data found in AlienVault forums, the open source version OSSIM is not multithreaded and is able of processing around 200 EPS (events per second). However, 150 EPS is the average for typical mid-sized organization (500-1000 users) and typical average peak EPS is about 8000 (with 250 infected endpoints).

In the commercial version, the number of EPS that can be processed depends on the type of deployment. For example, according to Alienvault documentation¹², in the USM standard deployment, a sensor can collect around 2500 EPS and the server can correlate around 5000 EPS.

Horizontal scalability of AlienVault SIEM can be achieved, only in the commercial version, using several USM servers and configuring a sensor to send events to these servers (see Figure 3-10). When the USM servers are defined in the agent running on the USM sensor, the user must configure a priority value for each of them. The USM server will connect to every configured SIEM server which is alive but only will send events to the one with the highest priority (or those ones in case of several with same priority). Only when those servers are not working, the events will be sent to a lower priority server.

¹² <https://www.alienvault.com/resource-center/data-sheets/unified-security-management-data-sheet>

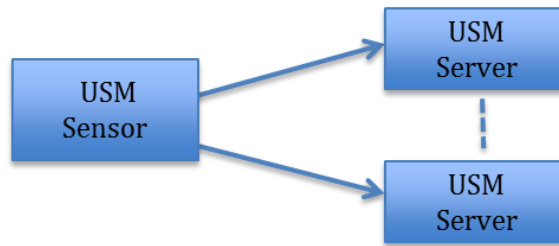


Figure 3-10: One sensor data to several USM

Another way to parallelize processing is to configure USM server to forward raw logs to several USM Loggers to have load distribution and improve long-term remote storage capabilities.

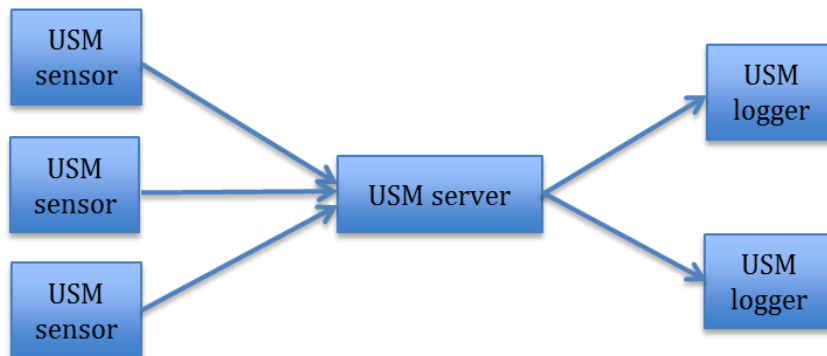


Figure 3-11: SIEMs data to several loggers

These two ways of parallel processing are only available in commercial version.

In case of having USM set up to receive raw pulse data and indicators of compromise (IoCs) from the AlienVault's open threat intelligence community called Open Threat Exchange (OTX) [16], USM server will also correlate it to alert in case some IoC affects assets monitored in the infrastructure.

3.5.4 Flexibility in security directives

OSSIM/USM security directives are based on XML files. Different categories are predefined (file categories.xml), each of them with an XML file assigned and a minimum and maximum values which define the set of security directives included in that category. For example, there is a category for denial of service (DoS) attacks:

```
<category name="DoS" xml_file="dos.xml" mini="14000" maxi="14999"/>
```

Each security directive XML file (e.g. directives-dos.xml for the category DoS) integrates a set of directives related to that category. Each directive is identified with an "id" and a "name" and includes a set of rules. These rules follow a hierarchical order which represents different correlation levels. The fields in the incoming events are checked against the conditions included in a <rule> (e.g. "from", "plugin_id"...)

predefined timeout (except for the first rule). Once it is reached the number of occurrences defined for a rule, the correlator will go to the next correlation level.

Below it is shown an example of Alienvault XML rule:

```
<directive id="34001" name=" Service attack, successful denial of service against web server on
DST_IP" priority="3">
  <rule type="detector" name="Apache mod_cache denial of service attempt detected"
reliability="0" occurrence="1" from="ANY" to="ANY" port_from="ANY" port_to="80"
plugin_id="1001" plugin_sid="12591">
    <rules>
      <rule type="monitor" name="Apache is down" reliability="8" from="1:DST_IP"
to="1:SRC_IP" sensor="1:SENSOR" port_from="80" port_to="80" plugin_id="2008" plugin_sid="2"
condition="eq" value="2" interval="15" time_out="30" absolute="true"/>
    </rules>
  </rule>
</directive>
```

Table 4: Example of Alienvault SIEMs security directive

The main advantage of these security directives is that the user can set up easily different correlation levels to trigger the same alarm with different risk values (based on the reliability and priority defined in each rule). However, only a predefined and fix set of attributes, most of them based on network properties (such as ip address or port) can be used in the rules. Besides, it is not possible to define complex conditions (e.g. comparison with "like" to match part of a text), only the logical ones are supported (e.g. equal, not equal...). This does not make it very flexible for a user to define new rules or understand them from a business perspective.

According to AlienVault web page¹³, USM provides over 3000 built-in directives and adds more every week through the AlienVault Labs Threat Intelligence Update service.

The following table summarizes the attributes allowed in the OSSIM/USM security directives¹⁴:

Name	Name of the rule with a description
Reliability	Reliability that the rule assigns to the event (from 0 to 10)
Timeout	The waiting time (in seconds) before the rule expires
Occurrence	Number of times an event has to occur in order for the rule to match
From	Source IP and port(s) that the rule tries to match.
To	Destination IP and port(s) that the rule tries to match.
Plugin_id	Data source ID that the rule tries to match.
Plugin_sid	Event type ID that the rule tries to match.
Sensor	The sensor that sends the events.

¹³ <https://www.alienvault.com/documentation/usm-v5/correlation/about-correlation-directives.htm>

¹⁴ <https://www.alienvault.com/documentation/usm-v5/correlation/about-correlation-rules.htm>

Protocol	The protocol specified in an event: Any, TCP, UDP or ICMP
Sticky Dif	When set, an arriving event needs to have a different value for a specific attribute than the previous one in order to be correlated. Accepted values for this attribute are None, Plugin_sid, SRC_IP, DST_IP, SRC_Port, DST_Port, Protocol, and Sensor.
Username	The username specified in the event.
Pass	The password specified in the event.
Userdata1-Userdata9	The user data fields specified in the event.

Table 5: Attributes in correlation rules supported by Alienvault SIEMs

Apart from those security directives, through the web graphical interface, the user can define cross-correlation rules. They establish the connections between NIDS events and vulnerabilities that will be used in the cross-correlation to increase the reliability of an event.

3.5.5 Behavioural analysis at application-level

As far as it is found in the documentation, current versions of AlienVault OSSIM/USM do not include User and Entity Behavior Analytics (UEBA) or machine learning capabilities.

Probably they could be provided by third-party tools and integrated in the SIEM with the development of new plugins, but OSSIM/USM solutions are more focused on network and systems monitoring.

3.5.6 Risk analysis capacity

AlienVault risk analysis takes into consideration the following aspects:

- **Reliability:** it represents how reliable it is that the attack will be successful. Reliability has a maximum value of 10 and a minimum of 0 (the event has almost no chance of being successful).
- **Priority:** it represents how important the attack is in case it is successful because it could cause an important damage. It has a maximum value of 5 and a minimum of 0 (the event has no interest and no alarm will be generated).
- **Asset relevance:** it represents how important an asset (a machine or device in the monitored network) is for the user. It has a maximum of 5 and a minimum of 0.

Once the normalized events arrive to the SIEM server, the following risk analysis operations are performed:

- **Event priority and reliability:** it is initially assigned priority and reliability scores to each incoming event considering the type of event and if that event belongs to a policy defined by the user. Policies are used in Alienvault to determine how to proceed with the incoming events arriving to a SIEM server (or to the logger, in the case of USM). A policy is defined by a set of filtering conditions and a set of consequences. Filtering conditions can be the source

or destination IP addresses, the source or destination port, the type of event, the sensor where the event were collected or the event timestamp. When an event matches the conditions specified in a policy, the consequences (what it is called “actions” in AlienVault) of that policy will be executed. One of these available consequences is to qualify the event with a specific priority. In this case, it will be used the priority specified in the policy instead of the one stored in the SIEM database for that event type matching the policy conditions.

- **Cross-correlation:** in case of an incoming event with defined destination IP address, it will be checked against destination IPs of vulnerabilities detected through scanning jobs performed from the SIEM server and stored in SIEM database. In USM v5, vulnerability scans can be also done directly from the assets.

The relationship between type of events and vulnerabilities to be checked is done through the cross-correlation rules defined. If it is found a coincidence, the reliability value is changed to 10.

- **Risk assessment:** it is calculated a risk score to be stored with the alarm generated. That score will be a value between 0 (minimum) and 10 (maximum) because of the following formula:

$$\text{Risk} = (\text{priority} * \text{reliability} * \text{asset_relevance}) / 25$$

The asset relevance value is retrieved from the inventory of known assets on the network stored in the SIEM database.

Additionally, the USM server also performs a cross-checking reputation data. IP addresses included in each event are checked against a reputation database of IP addresses provided and updated using the own AlienVault's Open Threat Exchange (OTX) service.

3.5.7 Exposed APIs

Incoming events to OSSIM/USM only can be received from agents running on SIEM sensors located in the architecture with the data sources supported described in a previous section. New compatible plugins can be added to the agents to enable new data sources or event types.

OSSIM/USM does not expose specific APIs to export alarms generated. This information is stored in a SQL database or sent to another OSSIM/USM server through the own server process. However, by using the **AlienVault Open Threat Exchange (OTX)**, it is possible to incorporate in the SIEM and share information about threats or malicious hosts in real-time. This data is shared using what they call an AlienVault *OTX Pulse*, which provides a summary of the threat and the related indicators of compromise (IOC).

3.5.8 Resilience

There is no reference in the OSSIM documentation to specific mechanisms to provide fault tolerance capabilities and resilience.

However, the use in the USM commercial version of deployments with multiple servers, loggers and sensors could be considered as a measure to provide resilience to the system.

In particular, in its USM Standard and USM Enterprise products, it is possible to configure USM for High Availability (HA). In a HA USM deployment, one node is the primary instance (master) which is active whereas there is a secondary slave instance (a mirror USM node where data is replicated) passive. When the primary node fails, the slave is automatically started to replace it. There are some restrictions to this HA feature, such as both nodes must be on the same subnet and they must be connected through a dedicated network cable.

3.5.9 Security event management and visualization capabilities

OSSIM/USM dashboard includes the following capabilities:

- Dashboards: different graphical charts are shown to the user as an overview of the monitored system status.
- Analysis: alarms, security events and raw logs can be visualized by the user.
- Environment: additional information provided by open source tools included with the OSSIM/USM distribution is integrated in the graphical interface. For example: Netflow traffic detected with the use of Fprobe tool, vulnerabilities detected with the use of Nessus or OpenVAS tools or the asset inventory provided by .
- Reports: it is possible to generate PDF reports with a summary of the SIEM analysis.
- Configuration: the SIEM administration and the configuration of the threat intelligence can also be done from the graphical interface. Additionally, in the commercial version the user has also the possibility of managing the sensors from the graphical interface.

Below it is shown an example of the open source OSSIM web graphical interface:



Figure 3-12: OSSIM Graphical Interface¹⁵

3.5.10 Reaction capabilities

Reaction capabilities are supported in OSSIM/USM through the definition of an **action** as consequence of a **policy** defined by the user (see Section 3.5.6). This means that when it is received an event that matches the conditions defined in a policy, it is triggered the action associated to that policy (if any).

There are three different available types of actions to be selected by the users in the OSSIM/USM graphical interface:

1. **Send an email:** It can be configured one or more email addresses and a text to be sent in case an alarm is triggered. This action can be also used to implement notifications by phone or with external ticketing system, but in that case, it would require an external messaging gateway capable of translating them.
2. **Execute a script:** It can be indicated a script to be launched once an alarm is triggered. This script will be executed in the same machine where the SIEM server is installed.
3. **Open a ticket:** OSSIM/USM includes a ticketing system where each user can have one or more tickets registered. When it is defined the action to open a ticket, it needs to be configured which user that ticket will be assigned to and a text to indicate the action to be taken by that user.

In any of them, the fields included in the alarm generated can be used in the different actions in the way of action keywords. Those keywords will be substituted by the value included in the alarm when they are triggered.

¹⁵ As can be seen in <https://www.alienvault.com/products>

3.5.11 Deployment and support

Alienvault OSSIM SIEM can be easily installed in a raw machine or using virtual resources, i.e. VMWare or Virtual Box, by means of an ISO file that Alien Vault provides. This ISO file contains an OS based on Debian Linux with all required for a full operative SIEM application running.

AlienVault USM can be deployed in a simple deployment called “USM All-in-one” (as the one provided by OSSIM) or in a complex/distributed deployment. In this latter, there are two different versions to be chosen depending on the user type and organization requirements: USM Standard and USM Enterprise (solution suitable for large organizations although it is not available as a virtual appliance). Besides, AlienVault USM can also be integrated in Amazon Web Services (AWS) environments.

Differences between OSSIM (open source) and USM (commercial) arise in support provided. Whereas OSSIM support is exclusively done by community in forums and online documentation, dedicated phone and email support are also available for USM.

3.5.12 Licensing

Alien Vault provides OSSIM as open source solution and USM (Unified Security Management) as commercial product.

3.5.13 Position in Gartner Magic Quadrant

Gartner report places this SIEM in visionaries due to lack of ability to execute compared to leaders.

Strengths of this SIEM according Gartner report are:

- AlienVault USM platform provides SIEM and a sort of other integrated security capabilities.
- USM provides a user-friendly interface.
- Lower cost compared to other solutions.
- License based on appliances instead of event volume.

Weaknesses of this SIEM according Gartner report are:

- Though it provides NetFlow capture, cannot generate alerts from NetFlow.
- Integration of unsupported data sources is not as easy as other competitors.
- Identity and access management to be used for linking with assets is restricted to Active Directory and LDAP.

3.6 XL-SIEM

Atos' XL-SIEM (Cross-Layer SIEM) tool was initially developed in the context of the European initiative FIWARE [17] as part of the Security Monitoring Generic Enabler (component in the FIWARE Platform) with the aim of overcoming some limitations detected in the open source SIEMs available in the market. In particular, to enhance their performance and scalability capabilities allowing the processing of increasing amounts of data and to add the possibility of correlation of events at different layers with more complex rules. Later, the XL-SIEM has been subsequently improved and validated throughout different projects such as ACDC (Advanced Cyber Defence Centre) [18], FI-XIFI [19], SAGA (Secured Grid metering Architecture), RERUM (Reliable, Resilient and secUre IoT for sMART city applications) [20] or WISER (Wide-Impact cyber Security Risk framework) [21].

The XL-SIEM is built on top of the open source SIEM OSSIM developed by Alienvault (see Section 3.5) and integrates a set of Java processes, including the high-performance correlation engine **Esper library** [22], packaged into a topology to be deployed in an Apache Storm cluster. **Apache Storm** [23] is an open source distributed real-time computation system for processing large volumes of data.

Figure 3-13 depicts the XL-SIEM architecture with its main components. The collection of data is done on the monitored infrastructure by SIEM Agents (see more details in Section 3.6.1) and the events are sent to the XL-SIEM core running on Storm where they are processed and correlated (see more details in Section 3.6.3). The events gathered as well as the alarms generated and configuration used, are integrated with the OSSIM deployment integrated in the XL-SIEM for its storage and visualization.

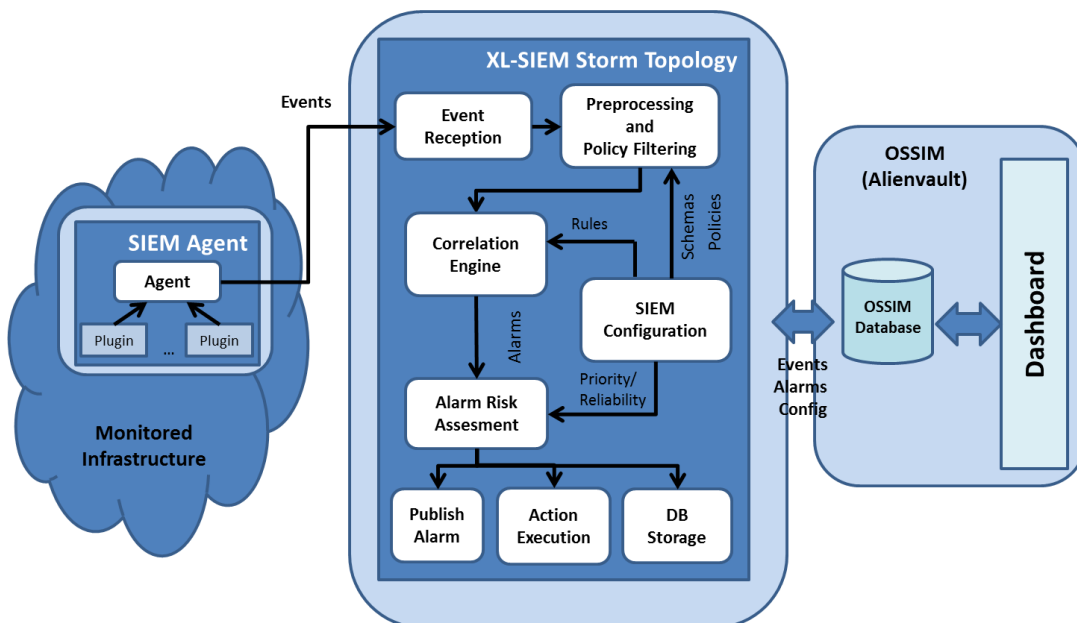


Figure 3-13: XL-SIEM Architecture

3.6.1 Data sources supported

The SIEM agents, which are an extension of the ones included in the open source OSSIM, are deployed for the collection of data from the monitored infrastructure that feed the processing done in the XL-SIEM. For this reason, data sources supported by OSSIM are also supported by XL-SIEM (see more details in Section 3.5.1).

Apart from those data sources, the following ones have been added to Atos XL-SIEM:

- STIX (Structured Threat Information eXpression) format data [24]: Cyber-threat observations, represented using this type of structured language for cyber threat intelligence, are supported through a STIX plug-in developed by Atos in the project ACDC [18] and added to the SIEM agents. This plug-in parses the STIX data and generates its representation in the OSSIM normalized event format used in the XL-SIEM.
- JSON format data: It is also supported data in JSON format received from the open source message broker RabbitMQ [25] that implements the Advanced Message Queuing Protocol (AMQP).

Additional features added to the SIEM agents included in OSSIM related to the data sources are:

- Secure transmission: the normalized events generated by the agents can be sent from the monitored infrastructure to the XL-SIEM server encrypted using the TLS (Transport Layer Security) protocol.
- Data anonymization: the user has the possibility of defining which fields in an event type are sensitive and need to be transmitted and stored anonymized. In the case of IP address, it is done a pseudo-anonymization where the original IP addresses are stored in a database together with the anonymized IP in order to be able of recovering the original one if required. For the rest of data, it is added a salt to the hashing process.

3.6.2 Data storage capabilities

The XL-SIEM takes advantage of the OSSIM data storage capabilities in order to allow the integration between the detection done by this open source SIEM and the one done by the XL-SIEM correlation processes. For this reason, the data storage capabilities included in the OSSIM version (see more details in Section 3.5.2) are also available in XL-SIEM. And the format used by the XL-SIEM for the events and alerts storage is consequently the same defined in OSSIM. They can be summarized in the following points:

- Data is stored in MySQL relational databases.
- There is a separate database for historical data.
- Currently, it is not supported integration with cloud storage services.

- The data storage can be in a different machine from the one where the event processing takes place to improve the performance or in case it is required more storage capacity. However, this does not mean that it is horizontally scalable since the storage needs to be done on a single server.

Besides, the XL-SIEM includes the capability to store both, events gathered by the agents and alarms generated by the server, for example in an external data warehouse using a RabbitMQ server. The format supported to send events and alarms in this case is JSON.

3.6.3 Processing capabilities

One of the advantages of XL-SIEM architecture is the use of a high-performance correlation engine running in an Apache Storm cluster for the processing of the incoming security events.

Scalability and distributed real-time processing of events in XL-SIEM are achieved by **Apache Storm** running together with **Apache ZooKeeper** [26] and **ZeroMQ** [27]. Apache Storm adds the capacity to perform the event processing throughout different servers distributing the load between them and supporting to increase the number of machines in the cluster. Apache ZooKeeper provides distributed synchronization across the Storm cluster maintaining centralized the configuration information. Communication among concurrent processes in the Storm cluster is done using the asynchronous distributed messaging ZeroMQ, without dedicated message broker and reducing the latency.

Through the XL-SIEM graphical interface it is possible to define multiple correlation processes to run in the Storm cluster, that is in different servers. Each of them can have a different set of rules and a different parallelism set up.

XL-SIEM makes use of Esper for event correlation and generation of alarms. According to the performance values described in Espertech website [28] using their own benchmarking kit provided with the source code, this complex event processing engine is able of processing 500000 events per second on a dual 2GHz CPU Intel based hardware with latency below 10 microseconds average with more than 99% predictability. These values are slightly reduced to a throughput of 120000 events per second in the benchmarking done in [29] with more complex queries, but in any case, they give an idea of its capability for processing large volume of data.

Besides, taking into account that Esper requires knowing in advance the type of incoming data to be processed, the XL-SIEM graphical interface also includes the possibility of defining different data schemas. When a new correlation process is configured, it can be used one of the existing data schemas or created a new one according to the expected incoming events. Furthermore, XL-SIEM also supports the usage of filtering policies as the ones defined in OSSIM (see more details in Section 3.5.3). These filters can be selected when the correlation processes are defined and used to reduce the number of events arriving to a specific correlation engine. These filters also enable XL-SIEM to support multi-tenant

processing capabilities. In this case, each client will have separated policies based on the agents belonging to a specific client or organization where the data were collected.

Consequently, this architecture provides real-time distribution through different machines not only of the correlation processes but also the support for different filtering policies, different rules and data schemas associated to each correlation process. This allows more flexibility in the processing and improves the processing capabilities and the usage of the available resources.

However, one weakness inherited from OSSIM that appears currently in XL-SIEM for data processing is the usage of MySQL relational databases located in a single server for the data storage and the lack of event indexing capabilities. As a first step to overcome this, XL-SIEM has enabled the reception of incoming data and the emission of alarms detected in the correlation, using the protocol AMQP (Advanced Message Queuing Protocol).

3.6.4 Flexibility in security directives

As it has been described in the previous section, XL-SIEM supports the configuration of different correlation processes. For each correlation process, the user has flexibility to configure:

- The rules or security directives that will be enabled in that correlation engine to trigger alarms.
- Filtering policies to be applied before the incoming events arrive to the correlation engine.
- Actions to be executed when an alarm is generated by a specific correlation engine. Supported actions will be described later in Section 3.6.10.
- Fields grouping to be used with the events arriving to a specific correlation process. Apache Storm includes a feature called “stream grouping” which allows configuring how the incoming stream to a process (in this case, the set of fields included the normalized XL-SIEM event arriving to the correlation process) will be partitioned among its different tasks (in this case, if the parallelism of the correlation process is higher than one, there will be a task by each processing thread). This is an advanced feature that probably only will be used by expert users but can improve the performance in case the rules are based on some specific field.

Concerning the definition of security directives, XL-SIEM supports two means:

- Pre-configured categories of rules:
For each correlation process, the user can select one or several directive categories to be included: scans behaviours, malware detection, denial of service attacks, brute force attacks or network attacks.
Each of these categories includes a pre-configured set of rules or security directives related to that different topic that will be enabled in the correlation engine.
- User custom rules:

The user also has the possibility to define his/her own rules or security directives and select some of them in the configuration of each correlation process.

In order to correlate events and generate alarms, the open-source Java-based Esper is used in the XL-SIEM. It implies that security rules are expressed in Event Processing Language (EPL).

EPL [30] is a declarative programming language similar to SQL (Structured Query Language) which allow expressing security directives with rich event conditions and patterns in a simple way. The usage of SQL clauses, such as *select*, *order by*, *group by* or *where*, in the definition of the rules that will trigger the alarms as well as SQL concepts such as joins or filtering through sub-queries, add a business perspective to the definition of security directives.

Another relevant feature of the EPL language is the capacity of using sliding or batched windows for processing the incoming events as well as the usage of filtering expressions in patterns, such as *every* or *every-distinct*. The window size can be defined considering a maximum number of events (e.g. “*win:length(10)*”) or with a time-based expression (e.g. “*timer:within(5 min)*”). And, although it is not included in the current XL-SIEM implementation, EPL has support to allow that external data accessible through JDBC may be queried and joined with the incoming stream data.

Esper API interfaces are used in XL-SIEM not only to send events into the correlation engine but also to create and manage EPL and pattern statements. Through the XL-SIEM graphical interface, the user can set up three levels in the definition of the security directives:

- **EPL variables:** These are values to be used in different EPL statements or directives that can change in the time (for example, to define a timeout). In this way, the user does not need to modify each security directive to change the value of a variable used in several of them.
- **EPL statements:** Sometime, it can be necessary to forward events from the original incoming stream to other stream for further processing. For example, to provide a first filtering in the incoming events to use the filtered stream in patterns included in the security directives. These EPL statements will not trigger alarms; they only will be inputs to another EPL statements or directives. They make use of *INSERT INTO* clauses to provide a first filtering in the correlation and to allow additionally a better and easier understanding of the security directives.
- **EPL directives:** These are the rules or security directives that will trigger alarms in the XL-SIEM. By each of these rules, a listener mechanism is automatically added to the Esper correlation engine instance configuration for notifying the pattern occurrence as soon as it happens. In the definition of each security directive, the user has flexibility to select the reliability and priority values associated to that alarm, in the same way that the ones used in OSSIM for each data source (see more details in Section 3.5.3). These EPL directives can use EPL variables and EPL statements to make them clearer

from a business perspective or for non-expert users. Another feature included in the XL-SIEM is the possibility of generating alarms based on previous alarms (cross-alarms). That is, to define security directives based on previous security directives.

Although it is not included in the current implementation of XL-SIEM, Esper API also includes event and event type interfaces to define java *EventType* objects with support to encapsulate metadata and inheritance hierarchy for event types that could be integrated in future XL-SIEM developments.

3.6.5 Behavioural analysis at application-level

XL-SIEM only includes behavioural analysis at application-level through the definition of EPL directives taking into account data provided by applications. In this case, it is required the implementation of specific plugins to parse the logs or data generated by the applications and make them compatible with the XL-SIEM format for its processing.

Current version does not include UEBA integration or machine learning capabilities.

3.6.6 Risk analysis capacity

XL-SIEM includes a risk assessment procedure analogous to the one provided by AlieVault SIEM solutions. This risk analysis takes into consideration the following aspects: Reliability, Priority and Asset relevance (see more details in Section 3.5.6).

The final risk assessment to be stored with the alarm generated will be the result of the following formula:

$$\text{Risk} = (\text{priority} * \text{reliability} * \text{asset}) / 25$$

3.6.7 Exposed APIs

Apart from the APIs inherited from the fact that XL-SIEM deployment integrates the open source OSSIM (see more details in Section 3.5.6), the following ones are available in XL-SIEM:

- **Events via RabbitMQ:** The events generated by the SIEM agent from the data collected by the sensors can be sent in JSON format to a RabbitMQ server. This output supports the use of TLS protocol.
- **Alarms via RabbitMQ:** Alarms generated by XL-SIEM can be also sent in JSON format to a RabbitMQ server. This output supports the use of TLS protocol.
- **Alarms via DDS:** XL-SIEM supports to send the alarms generated using the Data Distribution Service (DDS) [31], commonly used for real-time systems.
- **DRPC service to provide network topology JSON:** XL-SIEM includes a Distributed Remote Procedure Call (DRPC) [32] service that can be invoked to get a JSON which includes the monitored network topology with the

alarms associated to each node. This JSON can be used e.g. to generate a graphical representation with the status of the network.

3.6.8 Resilience

The resilience capabilities included in the XL-SIEM are the ones provided by the setup of a supervisory process called *daemontools* [33] (although any other supervisor such as *supervisord* [34] could be used) to monitor Apache Storm (where the XL-SIEM processes are running) and Apache Zookeeper processes.

Apache Storm is fault-tolerant [35] in the sense that if it is detected that one of the worker processes running in a node is not alive, it will be automatically restarted by the storm daemon called *supervisor*. In case it is detected that a node in the Storm cluster is not available, the workers running on it will be automatically restarted on another node in the cluster. This reassignment is done by the storm daemon called *nimbus* thanks to a heartbeat system and the coordination service provided by Apache Zookeeper.

However, Storm is also a fail-fast system and therefore in case of any unexpected error, the processes will automatically halt. Consequently, if you have only one nimbus instance running on the cluster, this means that the fault tolerance provided by Apache Storm is not enough. Running the Storm daemons under supervision using *daemontools*, they will be automatically restarted in case of failure. And for the same reason, Apache Zookeeper also needs to be run under supervision.

3.6.9 Security event management and visualization capabilities

XL-SIEM web graphical interface is built on top of OSSIM dashboard and consequently it integrates its visualization capabilities (see more details in Section 3.5.9).

Additionally, it includes high-level charts and diagrams in different dashboards to provide valuable information about incidents to non-security expert users. These diagrams can be adapted based on the client necessities and requirements. Some examples of available dashboards are the following ones:

- **Executive Dashboard:** It shows at glance high-level information relevant for a C-level administrator. For example, the current threat level of the monitored system with a colour code (green, yellow and red).
- **Operational Dashboard:** This dashboard shows information relevant for system administrators to be able of taking decisions. For example, the top 5 incidents detected, the hosts identified as source of security incidents or alarms or the destination TCP/UDP ports of the attacks.
- **Situational Awareness Dashboard:** This dashboard takes advantage of the DRPC service provided with the XL-SIEM to show a graph with the monitored network topology including the number of alerts detected in each component as well as representing with a colour code (green-yellow-red) the risk level of each node.

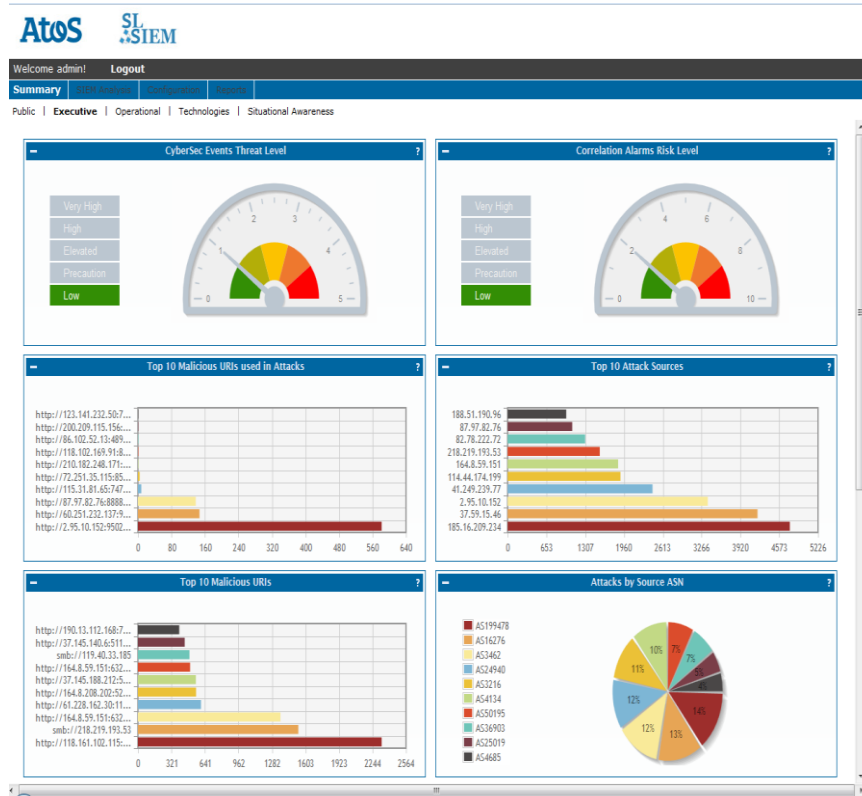


Figure 3-14: XL-SIEM Web Graphical Interface

3.6.10 Reaction capabilities

The processes running in the XL-SIEM topology include support to provide the same reaction capabilities offered in the open source version of the AlienVault SIEM: execute a script, send an email or create a ticket (see more details in Section 3.5.10) but with an enhanced performance.

On one hand, the different actions can be associated to each correlation process which allows its execution for specific alerts and not only associated to a defined policy. On the other hand, since there is a specific process defined in the Storm topology for the application of these reactions, it is possible to increase its parallelism to reduce the reaction time or even distribute in which node of the cluster this process will be launched. This can be interesting in case e.g. the script to be invoked or the email server is only available in a specific node.

Additionally, the XL-SIEM provides a Decision Support System (DSS) to help the user to analyse the risks detected and select suitable mitigation measures. Based on business data provided by the user and a set of mitigation measures associated to the different risks, this component offers an analysis of the societal impact of the risks as well as an analysis of costs and benefits of the mitigation measures proposed.

3.6.11 Deployment and support

XL-SIEM is provided as a self-contained Virtual Machine which includes the open source OSSIM distribution (operating system based on Debian), the Apache

Storm already installed with all the required dependencies (Apache Zookeeper, ZeroMQ) and the XL-SIEM topology deployed.

Documentation is also available with the installation and user guide manuals.

3.6.12 Licensing

Part of the software associated to the XL-SIEM product (such as the agents and a basic functionality of the server) is delivered under GPL license, but non-free. Atos can be contacted to obtain further information about pricing of complete distribution and other commercial use terms and conditions.

3.6.13 Position in Gartner Magic Quadrant

XL-SIEM is a SIEM only used internally in Atos and in different European projects so it does not appear in Gartner Magic Quadrant.

3.7 Splunk

3.7.1 Introduction

Splunk is the market-leading platform that powers Operational Intelligence. Splunk Core products include Splunk Enterprise (data collection, indexing, and visualisation engine for operational intelligence), Splunk Light, Splunk Cloud (the cloud version of Splunk Enterprise) and Splunk Universal Forwarder (streamlined version of Splunk Enterprise that forwards data to other Splunk Instances).

Premium solutions include Splunk Enterprise Security, Splunk User Behaviour Analytics and Splunk IT Service Intelligence, as follows:

- **Splunk Enterprise Security** is a security practitioner built on the Splunk Operational Intelligence platform also using search and correlation capabilities. Its main capabilities are providing customers the ability to capture, monitor and report on data from systems, applications and security devices but also giving admins the possibility to quickly investigate raised issues and resolve security threats across different domains (e.g. network protection domain, access domain etc.).
- **Splunk User Behaviour Analytics** is a Splunk premium solution meant to find known, unknown and hidden threats. Its main capabilities include: using unsupervised machine learning algorithms to detect internal/external threats, providing context around these threats, retrieving rank-ordered threats as well as their supporting evidence but also supporting incident scoping, workflow management, bi-directional integration with Splunk Enterprise and automated responses.
- **Splunk IT Service Intelligence** whose principal capability is to provide monitoring using a machine-driven approach for analytics driven IT, giving the user the possibility to quickly react to alerts.

In the following sections, the focus will be on Splunk Enterprise Security.

3.7.2 Data sources supported

3.7.2.1 Default data sources

It collects data from different environments:

- Local machine (e.g. on-the-premises in a server room)
- Remote machine (e.g. off-the-premises in a datacentre)
- Cloud (e.g. corporate network)
- Hybrid (e.g. on premise and in the cloud)

It also benefits from having more advanced data sources like 'Forwarders' (more information about Forwarders in Section 3.7.2.2).

The different types of native data sources, accepted by Splunk, are the following:

- Files and directories

- Network events (i.e. any network port).
- Windows sources: Splunk Enterprise accepts a wide range of Windows-specific inputs. Splunk Web lets you configure the following Windows-specific input types:
 - Windows Event Log data
 - Windows Registry data
 - WMI data
 - Active Directory data
 - Performance monitoring data
- Other sources:
 - First-in, first-out (FIFO) queues
 - Scripted inputs: data from APIs and other remote data interfaces and message queues.
 - Modular inputs: Splunk offers the possibility to define a custom input capability to extend the Splunk Enterprise framework.
 - The HTTP Event Collector endpoint: Splunk uses the HTTP Event Collector to get data directly from a source with the HTTP or HTTPS protocols.

Several Splunk apps and add-ons simplify the process of getting data into your Splunk deployment. Splunk also offers a multitude of apps/plugins to deal with more complex data types such as Google Spread Sheet, Protocol Data Input, Website input and so on¹⁶.

3.7.2.2 Forwarders

A forwarder is a Splunk Enterprise instance that forwards data to another Splunk Enterprise instance or a third-party system. Forwarders are meant to take the data and send it to the Splunk deployment for indexing. Most forwarders are lightweight instances (i.e. having minimal utilisation) which allow them to be located on the machine generating the data. Forwarders are similar to Beats in Elastic Stack.

Forwarders generally have the following capabilities:

1. Tagging of metadata (source, source type, and host)
2. Configurable buffering
3. Data compression
4. SSL security
5. Use of any available network ports

The types of forwarders provided by Splunk are:

Forwarder	Description
Universal forwarder	Dedicated, streamlined version of Splunk Enterprise that contains only the essential components needed to send data. In most cases, the universal forwarder is enough to send the data to indexers, though its main limitation is that it can only forward unparsed data.

¹⁶ https://splunkbase.splunk.com/apps/#/app_content/inputs

Heavy forwarder	Full Splunk Enterprise instance, having some features disabled thus achieving a smaller footprint.
Light forwarder	Full Splunk Enterprise instance, with most features disabled to achieve a small footprint. The universal forwarder supersedes the light forwarder for nearly all purposes. The light forwarder has been deprecated as of Splunk Enterprise version 6.0.0.

Table 6: Types of forwarders provided by Splunk

3.7.3 Data storage capabilities

Splunk stores the user's data through indexes (for more information about indexers, see Section 3.7.4.1). As the indexer indexes the data, it creates two types of data, which together constitute the Splunk Enterprise index:

1. The raw data in compressed form.
2. Indexes that point to the raw data and some metadata files i.e. associated index files.

The **files** reside in sets of **directories** organized by age. After a long period of time (several years usually), the indexer removes old data from the system. Each of the index directories is known as a **bucket**. In other words, data in Splunk is stored in buckets. A bucket moves through many stages as it ages. As buckets age, they may "roll" from one stage to the next.

Bucket stage	Description	Searchable?
Hot	Contains newly indexed data. Open for writing. One or more hot buckets for each index.	Yes
Warm	Data rolled from hot. There are many warm buckets.	Yes
Cold	Data rolled from warm. There are many cold buckets.	Yes
Frozen	Data rolled from cold. The indexer deletes frozen data by default, but you can also archive it. Archived data can later be thawed.	No
Thawed	Data restored from an archive. If you archive frozen data, you can later return it to the index by thawing it.	Yes

Table 7: Bucket stages provided by Splunk

Hence, an index can reside across many aged-designated directories. Splunk offers the possibility to create a **retirement and archiving policy/ log rotation** by configuring the size of the indexes or their age. When indexed data (which reside in directories called buckets, as previously mentioned) reach the final, frozen state, the indexer removes the data from the index by default. But the user is able to configure the indexer to archive the data when it freezes, instead of deleting it completely.

As previously mentioned, the moments in time when the user can decide to archive the data with respect to its freezing time:

1. The moment the index becomes too large: Attributes such as *maxTotalDataSizeMB* (which monitors when/if an index grows larger than its maximum specified size) control when a data changes from “cold” to “frozen”, thus freezes data when it gets too large.
2. The instance when the data becomes too old: Similar to the previous point, by setting the attribute *frozenTimePeriodInSecs*.

Splunk Enterprise supports archive signing, whose configuration allows the verification of the data by checking the hash signature of all the data in the archived bucket. Splunk has a built-in signing feature.

Splunk also offers the possibility for the user to create its own archiving script with data signature. To restore the archived indexed data, it is enough for the user to move the archived bucket into the thawed directory, which is the system’s directory not subject to the aging processing (hot > warm > cold > frozen) thus allowing the user to manipulate its data for as long as he/she wants. When the data is no longer needed, the user can simply delete or move it from the thawed.

It is also worth noting that **scalability** is easy in Splunk. It is enough to just add another server. Incoming data is automatically distributed evenly and searches are directed to all Splunk instances so that speed increases with the number of machines holding data. Optionally redundancy can be enabled, so that each event is stored on two or more Splunk servers.

In terms of **cloud storage services**, Splunk offers Splunk Cloud, which is a hybrid solution (centralized visibility across Splunk Cloud SaaS technology and Splunk Enterprise). Basically, it offers all the features of the Splunk Enterprise as a cloud-based service. This platform grants access to datacentres, private clouds but also public clouds. For public ones, the users can access apps such as Splunk App for AWS (Amazon Web Services). Other apps such as End-to-End AWS visibility help the user gain visibility across AWS and hybrid environment. Other apps for cloud services include Splunk App for Service Now (providing insights to changes, incidents and event management processes) as well as Splunk App for Akamai (which deals with performance monitoring, user adoption metrics, dashboards and real-time application security). Users can also use Splunk Light (a ‘light’ version of Splunk Cloud).

3.7.4 Processing capabilities

3.7.4.1 Data Indexing

Splunk indexes standard data types but also has the option to index custom data by offering additional configurations. Splunk Indexer typically receives data from a group of forwarders. The indexer consolidates the data and transforms it into

events and stores the events into an index. The indexer also has the role of searching the indexed data in response to search requests.

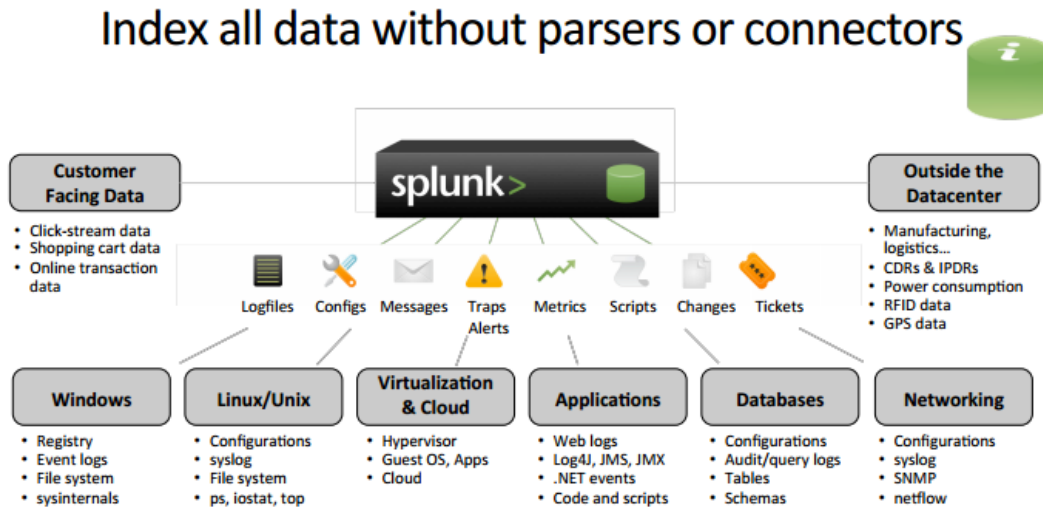


Figure 3-15: Splunk Data Indexing

The steps to index data in Splunk are the following:

1. Retrieve data from different data sources
2. Create a parsing queue
3. Define Parsing Pipeline:
 - Extract a set of default fields for each event, including host, source and source type.
 - Configuring character set encoding.
 - Identifying line termination using line breaking rules.
 - Identifying or creating timestamps.
 - *Optional*: anonymizing data based on configuration and applying custom metadata.
4. Create index queue
5. Define Index Pipeline:
 - Breaking all events into segments that can then be searched (the level of segmentation can be predefined which can affect the indexing searching speed, search capability, and efficiency of disk compression).
 - Building the index data structures.
6. Writing the raw data and index files to disk, where post-indexing compression occurs.
7. Check maximum indexing volume (subject to Splunk Enterprise license).

The full process is shown in Figure 3-16.

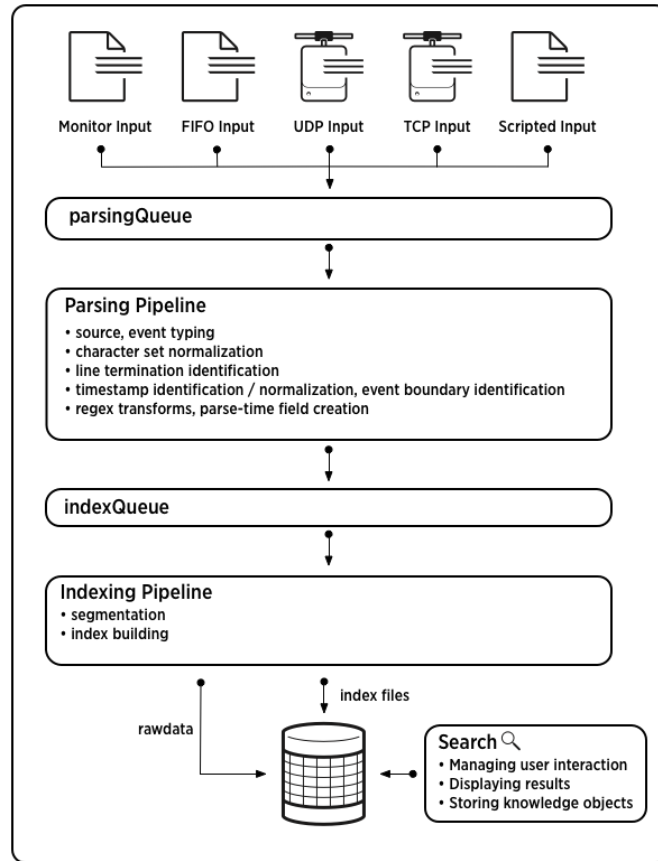


Figure 3-16 : Indexing process in Splunk

3.7.4.2 Data Model

Splunk uses data models, which are hierarchically structured search-time mapping of knowledge about one or more datasets. They encode the domain knowledge necessary to build a variety of specialized searches of those datasets which eventually help generate reports and dashboards without designing the searches to generate them.

In building a typical data model, knowledge managers use knowledge object types such as lookups (which provides data enrichment by mapping a selected value in an event to a field in another data source and appending the matched results to the original value, hence data correlation), transactions (which are a group of conceptually related events within a time span), search-time field extractions (the process by which Splunk extracts fields from the event data as well as the result of the process), and calculated fields (the output of an expression).

A Splunk user can select the data model that represents the category of event data that he wants to work with. Then he selects a dataset within that data model that represents the specific dataset on which she wants to report or create dashboards on. Data models are composed of datasets, which can be arranged in hierarchical structures of parent and child datasets. In other words, the child dataset will contain a subset of the parent's dataset. More precisely, data models are similar to relational database schemas.

A data model can be derived from:

1. A homogenous system such as .csv, which will result in a flat model with a single top-level root.
2. A heterogeneous system which might have several root datasets (e.g. events, searches and transactions). Any of these root datasets can be the first dataset in a hierarchy of datasets having the parent-child relationship. Also, each child dataset can incorporate new fields in addition to the fields they inherit.

In other words, data models are composed of one or multiple datasets with the following attributes:

1. Each data model dataset corresponds to an index.
2. Datasets are of four types: event dataset, search dataset, transaction dataset and child dataset (hence, root dataset can either be event, search or transaction dataset).
3. Datasets tend to be hierarchical: parent-child relationship.
4. Child datasets have inheritance: they inherit constraints and fields from their parents.
5. Child dataset is a subset of parent dataset.

Splunk offers a possibility to accelerate the results of data models. Data model acceleration enables faster results such as tables and charts. To accelerate a data model, it must contain at least one root event dataset, or one root search dataset that only uses streaming commands. Acceleration only affects these dataset types and datasets that are children of those root datasets. But overall, data models can contain a mixture of accelerated and un-accelerated datasets.

3.7.4.3 Search

Splunk benefits from having its own Search Process Language (SPL)¹⁷, which is similar to Elastic Query DSL. Using the “*search*” command, the user is able to use keywords, phrases, fields, boolean expressions, wildcards, key/value expressions, syntax completion, comparison expressions, etc. to specify exactly which events you want to retrieve from Splunk indexes. It is also possible to use pipeline “|” in order to use the output of the previous command as input in the next command.

From the implementation point of view, Splunk uses “*search heads*”, “*search peers*” and “*search head clusters*” which work together to generate search results as follows:

1. Search heads are Splunk Enterprise instance that are meant to handle search management functions, directing search requests to a set of search peers and then merging the results back to the user.
2. Search peers perform indexing and respond to search requests from the search head.
3. A group of Splunk Enterprise search heads that serve as a central resource for searching form a search head cluster.

¹⁷ <https://docs.splunk.com/Documentation/Splunk/6.5.0/Search/Aboutthesearchlanguage>

Using Splunk searching capabilities, the user can perform statistical searches, calculate metrics and look for specific conditions within a sliding window, reveals spikes, patterns and trends, drill down possibility etc.

3.7.4.4 Reporting

Reports can be created in Splunk by saving a **search** or a **pivot**. Ad hoc or scheduled, reports can be added to dashboards, with the possibility to accelerate reports (i.e. if your report has many events and is slow to complete when you run it, the user can accelerate it. Splunk software can run a background process that builds a data summary based on the results returned by the report. When you next run the search, it runs against this summary rather than the full index.) As previously mentioned, Splunk offers the possibility to manually create reports (by adding reports to the Report listing page from either Search or Pivot, or configuring a report manually or converting a dashboard panel into a report.

Splunk is also able to set up schedules report (i.e. reports that run on a regular interval and which triggers an action each time it runs i.e. send email or run script), as well as do more complicated task such as configure the priority of scheduled reports and generate PDFs of reports, dashboards, searches and pivots.

Once a report is created, the user (besides viewing and editing the report) can also change the permission of the report, embed it in external website and/or add report to dashboard.

3.7.4.5 Pivot / Pivot Editor

The Splunk pivot tool allows users to quickly design reports with tables and data visualizations that present different aspects of a selected Data Model. Pivots permit the user to generate these kinds of reports with a UI interface instead of having to use the SPL.

The pivot works in a very simple manner: it uses data models to define the category of events interested in and then utilizes data model datasets to subdivide the original data and define the fields that the user wants the pivot to return results on. The Pivot Editor allows users to simply point and click their way to creating reports/charts/graphs that provide great insight of the different selected datasets.

The Pivot editor contains multiple options such as:

- Visualization types: Statistics Table (default), Column Chart, Bar Chart, Scatter Chart, Bubble Chart, Area Chart, Line Chart, Pie Chart, Single Value Display, Radial Gauge, Marker Gauge, and Filler Gauge.
- Document Actions: Data Model (select Data Model, edit or change Data Set) and other basic options such as Save as and Clear.
- Job Actions: providing Pause and Stop buttons control the progress of the Pivot job. Other actions include: Share, Export, Print, and Search.

3.7.5 Flexibility in security directives

Splunk flexibility in security directives can be seen on different axes:

- Incident Review and Classification: comprehensive incident review capability, bulk event reassignment option, changes in status and event criticality classification and auditing possibilities.
- Reports and Security Metrics: reports, dashboards and metrics, possibility to convert search results into a graphics, dashboards or tables with analytics capabilities and export raw data as PDF or CSV.
- Risk-Based Analysis: possibility to apply risk score to any data and exposing the score's contributing factors.
- Threat Intelligence Framework: integrate, de-duplicate and assign weights to any number of open, proprietary or local threat intelligence feeds.
- Access, endpoint and network protection: simplifies access control monitoring, increases the effectiveness of endpoint security products such as Symantec Endpoint Protection, IBM Proventia Desktop or McAfee Endpoint Protection and it can discover anomalies in network data such as protocol data, firewalls, routers, DHCP, wireless access points etc.
- Asset Center/Identity Center: it allows the possibility to understand assets criticality. It leverages Splunk's ability to perform real-time 'lookups' of data stored in an asset database, active directory, spreadsheets or CSV files and use information as context for security events in reports and dashboards.

It is also worth mentioning the Splunk software is an easy-to-use tool for every user. The software was created with the non-technical user in mind, making for a more intuitive user experience. Features in the Splunk Enterprise Security include "point and click" data capture, an easy to use threat intelligence framework, and self-modifying correlations and thresholds.

3.7.6 Behavioural analysis at application-level

Splunk User Behaviour analytics app relies on data science and machine learning methodologies that require no signatures or human analysis, enabling multi-entity behaviour profiling and peer group analytics – for users, devices, service accounts and applications – and extended dormant timeline analysis. Splunk UBA addresses security analyst and hunter workflows, requires minimal administration and integrates with existing infrastructure to locate hidden threats.

Splunk UBA has the following capabilities:

- Streamlined Threat Workflow: It leverage security-semantics-aware machine learning algorithms, statistics and custom machine learning driven anomaly correlations to identify hidden threats without human analysis.
- Threat Review and Exploration: It visualizes threats over a kill chain to gain context.
- Kill Chain Detection and Attack Vector Discovery: It detects lateral movement of malware or malicious insider proliferation and responds to

real-time detection of anomalous activity (e.g. dynamically generated domain name). It detects behaviour based irregularities (e.g., unusual machine access, unusual network activity) and pinpoints botnet or CnC activity (e.g., malware beaconing, etc.) etc.

Splunk Enterprise, Splunk Enterprise Security (Splunk ES) and Splunk UBA work together to:

- Extend the search/pattern/expression (rule) based approaches in Splunk Enterprise (Security).
- Provide security teams with machine learning, statistical profiling and other anomaly detection techniques.
- Combine machine learning methods and advanced analytics capabilities to enable organizations to monitor, alert, analyse, investigate, respond, share and detect threats.

3.7.7 Risk analysis capacity

Splunk uses the Risk Analysis dashboard to display recent changes to risk scores and objects that have the highest risk scores. The users can use this to track changes in risk scores and analyse them. The risk score represents a single metric for the relative risk of a device or user object in the network environment over time. The object can be a system, user or other system.

Based on pattern matches, alerts can be raised and a risk modified. A risk modifier can be considered a number for the score computation. By using the computation and the dashboard, the user can review changes to an object's risk score, determine the source of a risk increase and decide if any action is needed. There are multiple filters on the risk analysis framework to search and filter the results. Those are:

- Source: Filters by the correlation search that have risk modifiers
- Risk object: Selects a risk object type and type a string to filter by risk object

This capacity of risk analysis offers multiple views for the user:

- Key Indicators: Displays the metrics relevant to the dashboard sources
- Risk Modifiers Over Time: Displays the changes made to risk modifiers over time
- Risk Score By Object: Displays the objects with the highest risk score
- Most Active Sources: Displays the correlation searches that contribute the highest amount of risk to any object
- Recent Risk Modifiers: Displays a table of the most recent changes in a risk score

The enterprise security version of Splunk uses risk analysis to find about and calculate the risk of events and suspicious behaviour over time with respect to

the user’s environment. In this case, the indexes all risk as events in the risk index.

By using the risk scoring, a user has a mean to capture and aggregate the activities of an asset or identity into a single metric using risk modifiers. The risk modifiers are associated with risk objects. There are three types of risk objects:

- system: Network device or technology
- user: Network user, credential, or role
- other: undefined object that is represented as a field in a data source

As examples, Splunk presents several security and fraud use cases:

- Detect and Investigate Malware
- Detect and Stop Data Exfiltration
- Privileged User Monitoring
- Using DNS Data to Identify Patient Zero Malware
- Detect Zero-Day Attacks
- Fraud: Detect Account Takeovers
- Compliance: Detect When a Critical System Stops Sending Logs to Splunk

Splunk comes with the risk analysis framework. It provides the mean to identify actions that raise the risk profile of individuals or assets. The risk is accumulated to allow the identification of entities that perform an unusual amount of risky activities. In their documentation, Splunk presents the diagram as an overview of the Risk Analysis framework:

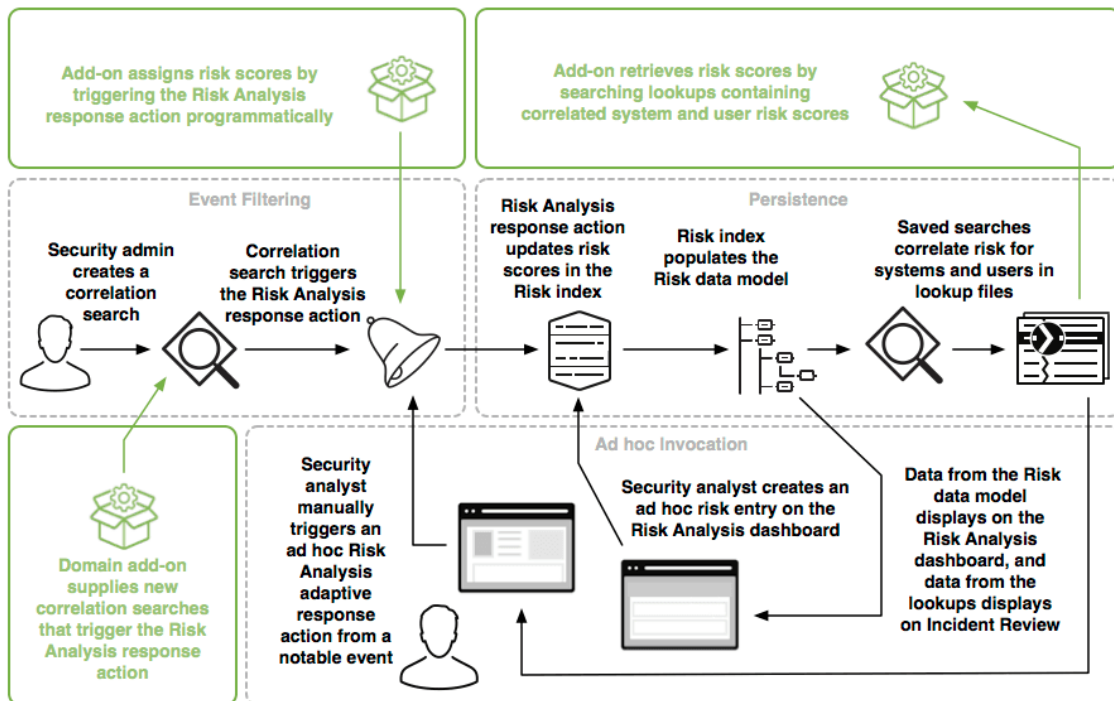


Figure 3-17 Splunk Risk Analysis Framework

3.7.8 Exposed APIs

The Splunk Enterprise REST API provides methods for accessing every feature. In particular, Splunk offers different documentation related to REST API:

- **REST API Reference Manual:** to learn about available endpoints and operations for accessing, creating, updating, or deleting resources. The resources are grouped into categories:

Category	Description
Access control	Authorize and authenticate users.
Applications	Authorize and authenticate users.
Clusters	Configure and manage cluster master and peer nodes.
Configuration	Manage configuration files and settings.
Deployment	Manage deployment servers and clients.
Input	Manage data input.
Introspection	Access system properties.
Knowledge	Define indexed and searched data configurations.
KV store	Manage app key-value store.
Licensing	Manage licensing configurations.
Outputs	Manage forwarder data configuration.
Search	Manage searches and search-generated alerts and view objects.
System	Manage server configuration.

Table 8: REST API Reference Manual Categories provided by Splunk

- **REST API User Manual:** users can find information about the APIs. API functions fall into one of the following categories, which have different interface behaviour:
 - Run searches.
 - Manage objects and configurations.

The REST API is organized around object and configuration resources. The architectural properties align with a REST API implementation that accesses domain resources with corresponding endpoints, using the HTTP protocol. The access methods provided by the Splunk REST API are limited to the following operations:

- DELETE: Delete a resource.
- GET: Get current state data associated with the resource or list child resources.
- POST: Create or update resource data, including enabling and disabling resources functionality.

Username and password authentication is required for most endpoints and REST operations. Additional capability or role-based authorization may also be

required, particularly for POST or DELETE operations. The API supports token-based authentication using the standard HTTP Authorization header.

Splunk REST API typically follows the standard API methodology (standard HTTP status codes, Atom Feed format responses, encoding schemas etc.)

Splunk comes with a REST API defining mechanisms to access and manipulate features from products. The API is built in a way to ensure functionalities encapsulation and transparency to the user.

Splunk offers an API more oriented on input, configuration management of files and settings. The API can be divided into the following categories:

- Search API. Manages search resources, such as alerts triggered, command information, results, scheduled view objects. There are different subcategories: alerts/fired_alerts, data/commands, saved/searches, scheduled/views, search/jobs, search/scheduler.
- Introspection API. This API is used for monitoring and to access server and instance information. The user can find information about data/index or server/introspection for example.
- Output API _cat. This category is used to manage data from forwarders, such as data/outputs/tcp.
- Cluster API. This category of API manages master and peer cluster nodes via API under the forms of cluster/config, cluster/master, cluster/searchhead, cluster/slave.
- Indexes and data management. This API is one of the mostly used for the user. It defines data configurations indexed by the Splunk platform. It is formed from methods used to manage indexes, index settings and data tables and properties. It provides means to handle the data by using look-ups, field extractions, field aliases, source types, and transforms. Some methods fall into categories such as data/lookup-table-files, data/props or search/fields. In addition, it manages saved event types and search field configurations and search time tags.
- Deployment. Splunk has an API oriented on nodes and server-client, like deployment/client, deployment/server or search/distributed.

3.7.9 Resilience

Splunk usually handles crash recovery without any intervention. If an indexer goes down unexpectedly, some recently received data might not be searchable. When the user restarts Splunk, it will automatically run a specific command that will diagnose the health of the buckets and rebuilds search data as necessary. The user will rarely need to run this command manually, which may take several hours (during which, the data will be inaccessible). If so, there is a full Splunk support for this use case.

On the other hand, if the index and metadata files in a bucket (version 4.2 and later) somehow get corrupted, the user has the possibility to rebuild the bucket

from the raw data file solely. Also, Splunk automatically deletes the old index and metadata files and rebuilds them.

Another use case is invalid hot buckets. A hot bucket becomes an invalid one when Splunk detects that the metadata files are corrupt or incorrect. Splunk ignores invalid hot buckets. Data is not get added to such buckets, and they cannot be searched. Invalid buckets also do not count when determining bucket limit values. This means that invalid buckets do not negatively affect the flow of data through the system. Splunk also offers the possibility to recover invalid hot buckets.

Also, it will be rarely the need to rebuild index-level manifests, but if the user is obliged to, Splunk provides a few commands that do just that.

It is also worth mentioning that in Splunk there is no single point of failure. In many environments, overloaded database server slows down half the applications in the data centre. In Splunk, this will not happen.

3.7.10 Security event management and visualization capabilities

For security event management in Splunk, the user has the possibility to create alerts. There are different alert types and triggering options. The real-time alert system continuously searches for results in real time. The user can configure real-time alerts to trigger every time there is a result or if results match the trigger conditions within a particular time window. The user has the option to configure an alert action by editing *savedsearches.conf* or he can configure a script.

Splunk also provides advanced threat detection methods by using a combination of approaches (machine learning, statistics and rules). Additionally, advanced analytics and threat modelling is used to detect advanced threats across different vectors, such as email, malware and web-based attacks. Splunk UBA provides advanced threat-specific behaviour detection algorithms aligned with the Cyber Kill Chain.

Integration with popular sandbox vendors, as well as native threat intelligence feeds, is also used. Splunk App for Stream analyses wire data including HTTP, DNS communications in real time to provide network visibility, which can be correlated with additional data.

Splunk also offers the possibility to create dashboards to investigate security events. Splunk web user interface allows the user to create and edit dashboard in a simple manner. Dashboards use Simple XML source code to define their content and behaviour.

3.7.11 Reaction capabilities

The scope of the Splunk User Behaviour analytics app, as well as the Enterprise Security, is to produce actionable results with risk ratings plus supporting evidence so that SOC analysts can respond and investigate threats.

Send an email notification to specified recipients when an alert is triggered. Email notifications can include information from search results, the search job, and alert triggering. In addition to alerting, there are other email notification contexts.

Splunk allows the user to use tokens when creating mail notifications. Tokens represent data that a search generates. They work as placeholders or variables for data values that populate when the search completes. The user is able to use tokens in the fields of the email notification (e.g. To, Cc, Bcc, Subject etc.)

Splunk also offers incident management features such as the Investigation Journal and the Investigation Timeline. The Splunk Custom Alert Actions allows task and workflow automation, which includes integration with third-party applications. Integration with popular third-party service desk solutions and services is provided.

3.7.12 Deployment and support

In order to deploy Splunk, the user needs to configure a deployment server and the deployment clients (but note that the user is also able to use third-party tool, such as Chef, Puppet, Salt, or one of the Windows configuration tools instead of a deployment server).

The deployment server (which is the recommended tool) is the utility for distributing configurations, apps, and content updates to groups of Splunk Enterprise instances. The Splunk admin can use it to distribute updates to most types of Splunk Enterprise components (e.g. forwarders, non-clustered indexers, search heads etc). The deployment server instance can also be dedicated exclusively to managing updates, if needed.

The advantage of using a deployment server is that the admin can group diverse Splunk instances into different groups, identify the configurations and apps needed by each group, and then use the deployment server to update their apps and configurations when needed.

To make things simpler, the Forward Manger, built on top of the deployment server, provides an easy UI to configure the deployment server but also to monitor the status of deployment updates. As a primary use, the Forward Management is meant to manage large groups of forwarders, but it can also be used, as secondary use, to configure the deployment server and manage updates. Key elements of a deployment are:

- Deployment server (i.e. centralized server configuration)
- Deployment client (i.e. Splunk instance remotely controlled by the server)
- Deployment app (i.e. content including configuration files that are maintained on the deployment server and that are also deployed as a unit to deployment clients or server classes). Note that the term 'app' is somehow different than the usual 'app' meaning
- Server class (i.e. a group of deployment clients sharing some characteristics)

See how these deployment elements fit together in Figure 3-18.

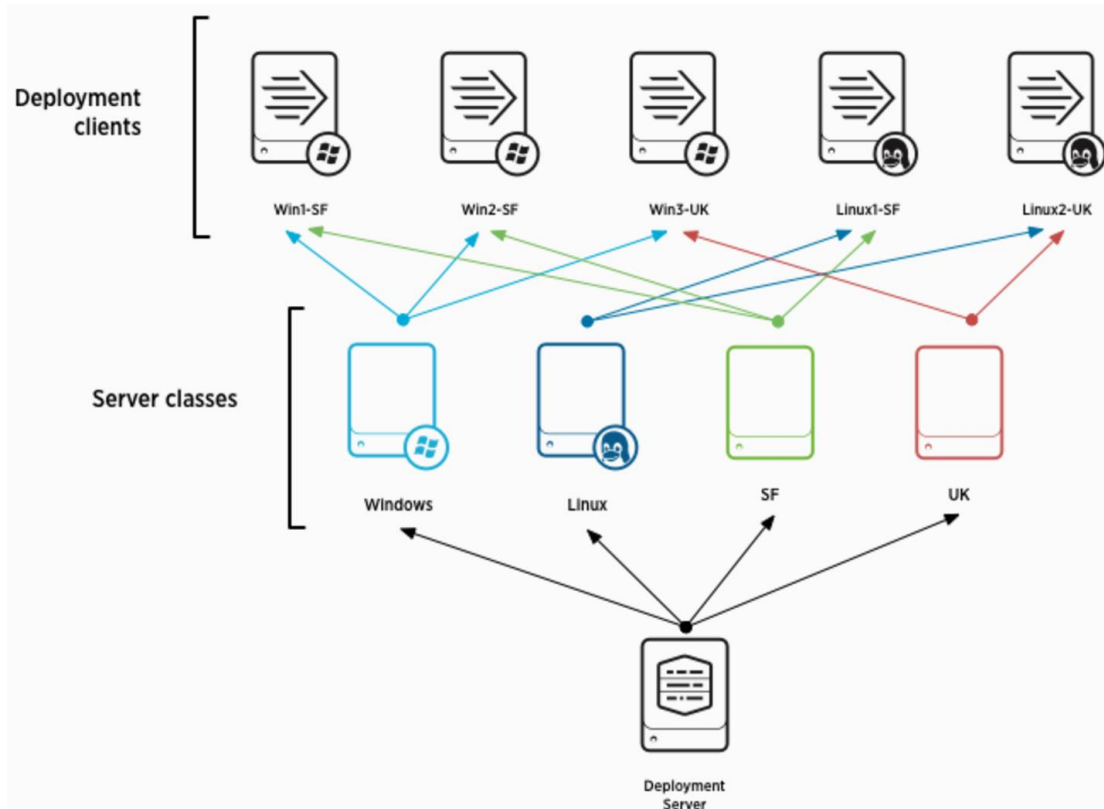


Figure 3-18 Splunk Deployment units - Dependencies

3.7.13 Licensing

Splunk uses a licensing system based on indexing. Indexing is the process of taking in data from the sources that the user designs and processes for analyse. Splunk uses calendar days licensing system. Therefore, the license specifies how much data the user can index per day.

The license system Splunk offers is formed from a license master and a pool formed from license slaves, each license coming from other indexers.

The Splunk license specifies the amount of data a Splunk platform instance can index and what features the user has access to. It grants the Free Software, Content Subscription, Splunk Extensions, Customer Extensions, Purchased Software, Evaluation Software, Test and Development Software, Open Source Software.

Each Splunk software instance requires a license. Splunk licenses specify how much data a given Splunk platform instance can index and what features you have access to.

The Splunk licences are:

1. **The Enterprise license:** enables all Enterprise features (e.g. authentication, distributed search, deployment management, scheduling of alerts, and role-based access controls). Enterprise licenses are available for purchase and can be any indexing volume. Splunk enterprise licenses include:

- No-enforcement license: allows users to keep searching even if you acquire five warnings in a 30 day window.
 - Enterprise trial license: allows a maximum indexing volume of 500 MB/day. The Enterprise trial license expires 60 days after you start using Splunk.
 - Sales trial license: varying size and duration. The Enterprise trial license expires 60 days after you start using Splunk.
 - Dev/Test licenses: operate Splunk software in a non-production environment.
2. **The Free license:** includes 500 MB/day of indexing volume, is free, and has no expiration date, but has many disabled features (e.g. multiple user account, distributed search, forwarding TCP/HTTP formats, alerting/monitoring and authentication and user management).
 3. **The Forwarder license:** allows forwarding (but not indexing) of unlimited data, and enables security on the instance so that users must supply username and password to access it.
 4. **The Beta license:** require a different license that is not compatible with other Splunk releases. Beta licenses typically enable Enterprise features, they are just restricted to Beta releases of Splunk products for evaluation purposes.
 5. A license for a premium app is used in conjunction with an Enterprise or Cloud license to access an app's functionality.

It is also worth mentioning that, to use search heads, the users must acquire an Enterprise license. There are also some restrictions regarding indexer cluster nodes (for index replication), most importantly, the user must have the Enterprise license and should configure cluster nodes with the same licensing configuration.

There are warnings and violations that occur when the user exceeds the maximum indexing volume allowed for the license. The pricing is based on the number of GB of data per number of days. Several examples: 1 GB/day - \$5,175 (Perpetual License per GB) and \$2,070 (Annual Term License per GB) or 50 GB/day - \$2,185 (Perpetual License per GB) and \$874 (Annual Term License per GB).

However, the Splunk license does not grant to copy, modify, adapt or create derivative works of any Splunk Materials. Furthermore, no rent, lease, loan, resell, transfer, sublicense, distribute, disclose or otherwise provide any Splunk Materials to any third party. Lastly, does not grant to decompile, disassemble or reverse-engineer. Splunk has the ownership worldwide right, title and interest in and to the Splunk Materials and all related Intellectual Property Rights.

For Splunk Enterprise, there is no limit to the number of users, searches, alerts, correlations, reports, dashboards or automated remedial actions. Its license price is based on maximum daily aggregate volume of uncompressed data indexed, expressed in gigabytes per day, as can be seen in Figure 3-19.

The More Data You Index, the Less You Pay

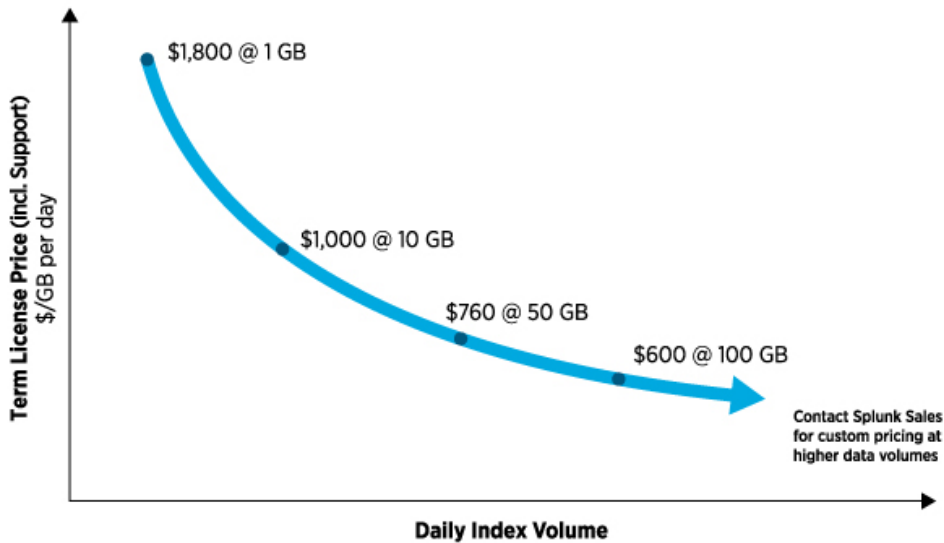


Figure 3-19: Splunk Enterprise Price

3.7.14 Position in Gartner Magic Quadrant

It is important to say that in both Gartner reports, Splunk was highly evaluated (Figure 3-1). It is worth mentioning that Splunk advanced security capabilities are due to its native machine learning functionality of the Enterprise Security version, but also from the integration with Splunk User Behaviour Analytics (UBA).

In Gartner's SIEM report, Gartner enumerates the strengths as well as caution notes when dealing with Splunk, as follows:

- Strengths:
 - Splunk's investment in security monitoring use cases. This helps Splunk visibility in client basis.
 - Advanced security analytics capabilities. This includes both native machine learning functionality and integration with Splunk UBA.
 - Splunk's presence, and investment, in IT operations monitoring solutions.
- Weaknesses:
 - Splunk Enterprise Security provides only basic predefined correlations for user monitoring and reporting requirements, while leading competitors offer richer content for use cases.
 - Splunk license models are based on data volume in gigabytes indexed per day. This results in costly solutions since customers deal with Big Data, as well as planning and prioritization of data.
 - Potential buyers of Splunk UBA must plan appropriately, since Splunk UBA requires a separate infrastructure and license model than the one for Splunk Enterprise and Enterprise Security.

The position of Splunk in Gartner SIEM ranking for Advanced Threat Detection can be seen in Figure 3-20.

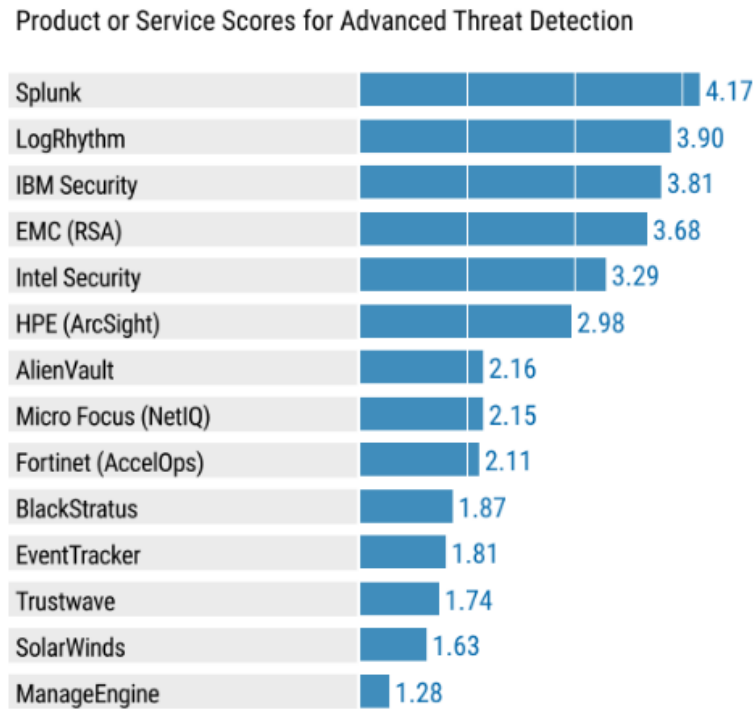


Figure 3-20 Advanced Threat Detection – Gartner

3.8 Elastic Stack

Elastic Stack is not a SIEM, at least not in the same manner than others. The partners of DiSIEM project has decided to analyse the functionalities and capacities of Elastic Stack that could improve the security of actual SIEMs used, or perhaps, became a type of SIEM.

Below is described the most relevant features of Elasticsearch and other tools that together form the Elastic Stack. The sections below are largely based on the Elastic products documentation from the official Elastic company website. [36]

3.8.1 Introduction

Elasticsearch is a product that is a readily-scalable, open-source, broadly-distributable, enterprise-grade search engine. This product is accessible using an extensive and elaborate API. It can also power very fast searches that are able to support customer data discovery applications. Elasticsearch represents the core of the ELK Stack (or BELK which stands for Elasticsearch, Logstash, Kibana and Beats).

Logstash is also a very useful tool. It is a server-side data processing pipeline which is also open source. One of its main scopes is to ingest data from a variety of sources in the same time, transforms them, and then send them to customer's favorite "stash." Elasticsearch is usually the chosen stash.

Kibana (which is the K in the ELK Stack) works as the visualization module that queries Elasticsearch and that is meant to render the query results by means of visualizations and dashboards. Kibana core uses classic visualizations such as histograms, line graphs, sunbursts, pie charts, histograms etc. They add value the full aggregation capabilities of Elasticsearch.

The ELK stack is the most popular product of the Elastic company, but more products (free/paid and commercial/noncommercial) have been developed by the company and that complete the ELK Stack (Monitoring: Marvel, Security: X-Pack, Logs forwarders: Beats, etc.)

Briefly, Elastic Stack core is made of:

- Elasticsearch – Index, correlation and search engine.
- Kibana – Analytics and visualization platform.
- Logstash – Engine to collect and enrich data.
- Beats – Data shippers that collect and send data to Logstash (or directly to Elasticsearch). Standard Beats include:
 - Filebeat – Collects logs and files (apache logs, syslogs, etc).
 - Metricbeat – Collects metrics from systems and services (MEM, CPU, etc).
 - Packetbeat – Lightweight network packet analyzer.
 - Winlogbeat – Collects and sends events from Windows systems.
 - Customized Beat – Possibility to create custom Beat with Libbeat.

The Figure below shows the components that form the Elastic Stack and the standard data flow.

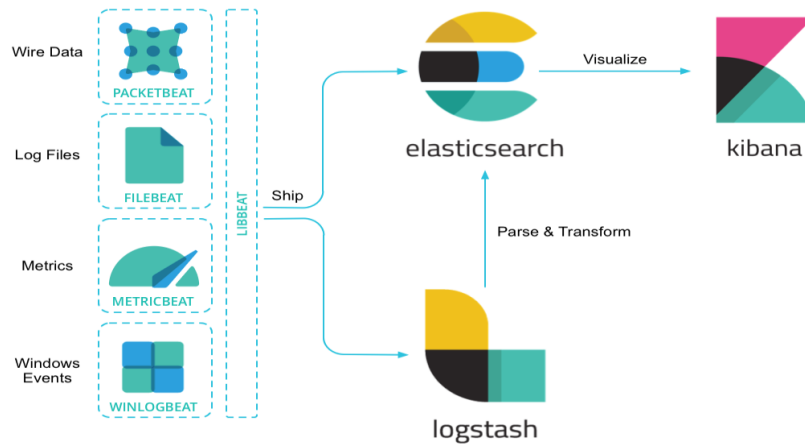


Figure 3-21: The standard components of Elastic Stack.

3.8.2 Data sources supported

The ELK Stack offers different data ingestion capabilities. It also has a focus on flexibility and support for a wide range of log data formats/sources.

The user has the possibility, when indexing data into Elasticsearch, to perform it either using one of the Elastic products or either by means of custom software/application. When an Elastic product or a custom connector is used, Elasticsearch will use JSON format to store documents (see example below) therefore the output of the data sources should always be in a JSON format.

```

{
  "@timestamp": "2016-04-22T22:09:09.631Z",
  "beat": {
    "hostname": "DeepThought",
    "name": "DeepThought"
  },
  "event": "logged_in",
  "fields": {
    "planet": "Magrathea",
    "service": "Answerer"
  },
  "id": 42,
  "input_type": "log",
  "offset": 0,
  "session_id": "91e5b9d",
  "source": "test/structured.log",
  "tags": [
    "i",
    "heart",
    "json"
  ],
  "type": "log",
  "user": "arthur",
  "verified": false
}

```

Figure 3-22: A JSON file example

Even though this can be seen as a restriction, Elasticsearch is the search engine component of the ELK Stack, hence Elasticsearch is not meant to support multiple input formats and. In fact, this is the job of the data forwarders, parsers and connectors.

In *Appendix I: Elasticsearch Products for Log Ingestion*, there is a list some of the widely-used data input sources for Elasticsearch.

3.8.3 Data storage capabilities

There are a few core concepts of Elasticsearch very important to understand data storage:

- *Near Realtime (NRT)*: Elasticsearch is a near real time search platform. Hence there is a slight latency (approximately one second) from the time a document is indexed until the time the user can search for the document.
- *Cluster*: This is a collection of n nodes (servers) that together holds the entire data. They also provide federated indexing and search capabilities across all nodes.
- *Node*: This is a single server that is part of a cluster. It is meant to store data, and to participate in the cluster's indexing and search capabilities.
- *Index*: It is a collection of documents that have somewhat similar properties e.g. an index for customer data and another index products. An index can be identified by a name. This name is used to refer to the index when performing different operations such as indexing, search, update, and delete operations on the documents.
- *Type*: Inside an index, one or several types can be defined. A type can be seen as a logical partition/category of index whose semantics depends on the user.
- *Document*: This is basic unit of information which the user can index and it is rendered in JSON format (JavaScript Object Notation). Within an index/type, the user can store as many documents as he wants.

Shards and Replicas. An index can store a large amount of data which can exceed the hardware limits of one single node. Shards in Elasticsearch offer the option to subdivide the index into multiple pieces which are called shards. When the user creates an index, he can simply define the number of shards that he wants. At a closer look, each shard is a fully-functional and independent "index" which can be kept on any node in the cluster.

Sharding is important for two main motives:

- It permits the distribution and parallelization operations across shards (and maybe on multiple nodes) hence increasing throughput/performance.
- It allows horizontally splitting /scaling content volume.

The mechanics of how a shard is distributed and how its documents are aggregated back into search requests are completely managed by Elasticsearch. They are also transparent to the user.

Replication is important for two main motives:

- It permits the scaling of search volume/throughput (because searches can be executed on all replicas in parallel).
- It offers high availability in case a shard/node fails. For this reason, it is important to note that a replica shard is never allocated on the same node as the original/primary shard that it was copied from.

Data storage in Elasticsearch. Very important is the fact that there are very important configuration parameters (i.e. deploying Elasticsearch is the data directory on each node of the Elasticsearch cluster). The user knows Elasticsearch uses Lucene (which handles the indexing and querying on the shard level). The files in the data directory are written by both, Elasticsearch and Lucene.

Lucene is responsible for writing and maintaining the Lucene index files. Elasticsearch writes metadata related to features on top of Lucene (e.g. field mappings, index settings and other cluster metadata). End user and supporting features that do not exist in the low-level Lucene but are provided by Elasticsearch.

Lucene index files: The actual data indexed in Elasticsearch is presented by a variety of local files in Lucene. It should be mentioned the fact that Lucene is based on the notion of inverted index (i.e. a very versatile data structure that is based on computing Terms frequency and mapping terms to documents).

Lucene creates different files, which can be seen in the table below:

Name	Extension	Brief Description
Segments File	segments_N	Stores information about a commit point
Lock File	write.lock	The Write lock prevents multiple Index Writers from writing to the same file.
Segment Info	.si	Stores metadata about a segment
Compound File	.cfs, .cfe	An optional “virtual” file consisting of all the other index files for systems that frequently run out of file handles.
Fields	.fnm	Stores information about the fields.
Field Index	.fdx	Contains pointers to field data.
Field Data	.fdt	The stored fields for documents.
Term Dictionary	.tim	The term dictionary, stores term info.
Term Index	.tip	The index into the Term Dictionary.
Frequencies	.doc	Contains the list of docs which contain each term along with frequency.
Positions	.pos	Stores position information about where a term occurs in the index
Payloads	.pay	Stores additional per-position metadata information such as character offsets and user payloads.
Norms	.nvd, .dvm	Encodes length and boost factors for docs and fields
Per-Document Values	.dvd, .dvm	Encodes additional scoring factors or other per-document information.
Term Vector Index	.tvx	Stores offset into the document data file.

Term Vector Documents	.tvd	Contains information about each document that has term vectors.
Term Vector Fields	.tvf	The field level info about term vectors
Live Documents	.liv	Info about what files are liv.

Table 9 : Lucene files

Retiring Data: When data ages, it tends to become less relevant. As Elasticsearch works at an index per time frame, it enables the user to easily delete old data.

If the user wants to keep them around, these indices can be closed. They will still exist in the cluster, but they won't consume resources other than disk space. Reopening an index is much quicker than restoring it from backup.

Finally, very old indices can be archived off to some long-term storage (e.g. shared disk or Amazon's S3 using the snapshot-restore API). This is in case the user needs to access them in the future. Once a backup exists, the index can be deleted from the cluster.

Flat Data: We also can store a copy of data, sending raw data to a file in a raw form or formatted (e.g. csv).

Elasticsearch Curator: Elasticsearch Curator helps the user curate, manage, the Elasticsearch indices and snapshots. Curator features can be seen below:

- Delete snapshots
- Change the number of replicas per shard for indices
- Take a snapshot of indices
- Add or remove indices from an alias
- Open closed indices
- Force merging indices
- Change shard routing allocation
- Close indices
- Create index
- Restore snapshots
- Delete indices

Storage Scalability: The user can increase the storage capability of an Elasticsearch cluster quite simple by adding one or many data nodes. Adding new data nodes will trigger a load balancing process to equally partition data between nodes to stay under the storage limits for each node. It is also meant to improve search performance.

3.8.4 Processing capabilities

In this section, information about data indexing, search, filtering, aggregation and reporting will be provided.

Searching. Tutorials are provided by the company on how to the user can manipulate the search API. This is done by:

- sending parameters through REST request body. It allows the user to be more expressive and to define the searches in a more readable JSON format,
- or by sending these parameters to REST request URI.

In terms of the query language, ELK provides a JSON-style domain-specific one, which bears the name Query DSL, which at first it is not necessary intuitive.

Filtering. Lucene calculates a field called “document score” (i.e. a numeric value which is a relative measure that measures how well a document matches a query). If a score is high, it means the document matches well the query. Hence, if the score is low, it means the document does not correspond to the query.

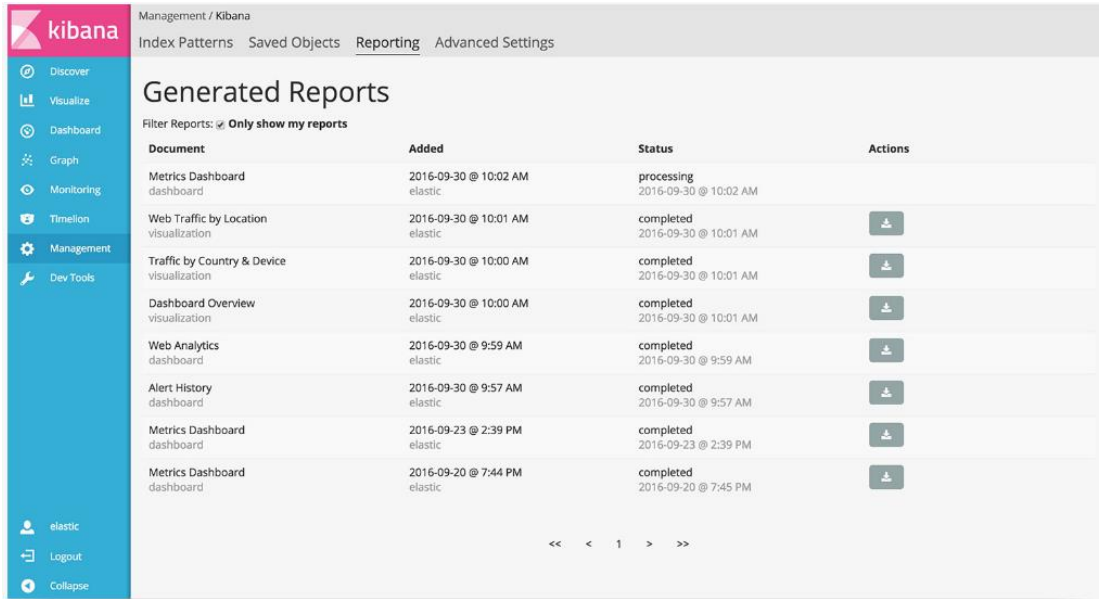
It is worth mentioning that the queries do not always need to compute the score. The scores are only used for filtering reasons. ELK detects these situations and hence it does not compute useless scores. This also help with the optimization of the query execution.

Aggregations. Aggregations are meant to help the user group and extract relevant statistics from the data. ELK offers the possibility to execute aggregations and return hits as two separate results. The types of aggregations are:








- Pipeline: compute statistics over the output of other aggregations and the associated metrics.
- Matrix: family of aggregations whose scope is to operate on multiple fields and produce a matrix result. This matrix result is based on the values that are extracted from the requested document fields. This aggregation family does not support scripting.
- Bucketing: family of aggregations where each bucket has associated a document and a key. The moment the aggregation is executed, all the bucket criteria will normally be evaluated against every document. If the document meets the criteria, the document is considered to ‘fall’ in the bucket.
- Metric: aggregations that keep track and compute the metrics of a document collection.

Reporting. It is very useful the fact that ELK offers the possibility to generate, schedule and email reports. Kibana is particularly used this case. The user has the option to create a report from any Kibana visualization or dashboard. Every report is also print-optimized, customizable and PDF-formatted. The user also has the possibility to automatically share the reports with other people.

Processing Scalability. Elasticsearch is meant to scale. Elasticsearch permits scaling out flexibly and rapidly. It can run easily on a single node cluster or in a cluster containing hundreds of nodes with almost identical experience. Increasing from a large cluster to a very large cluster requires more planning and design, but it is still rather easy.



The screenshot shows the Kibana Reporting interface. The left sidebar contains navigation options: Discover, Visualize, Dashboard, Graph, Monitoring, Timelion, Management (selected), and Dev Tools. The main content area is titled 'Generated Reports' and includes a filter 'Only show my reports'. Below this is a table with the following data:

Document	Added	Status	Actions
Metrics Dashboard dashboard	2016-09-30 @ 10:02 AM elastic	processing 2016-09-30 @ 10:02 AM	
Web Traffic by Location visualization	2016-09-30 @ 10:01 AM elastic	completed 2016-09-30 @ 10:01 AM	
Traffic by Country & Device visualization	2016-09-30 @ 10:00 AM elastic	completed 2016-09-30 @ 10:01 AM	
Dashboard Overview visualization	2016-09-30 @ 10:00 AM elastic	completed 2016-09-30 @ 10:01 AM	
Web Analytics dashboard	2016-09-30 @ 9:59 AM elastic	completed 2016-09-30 @ 9:59 AM	
Alert History dashboard	2016-09-30 @ 9:57 AM elastic	completed 2016-09-30 @ 9:57 AM	
Metrics Dashboard dashboard	2016-09-23 @ 2:39 PM elastic	completed 2016-09-23 @ 2:39 PM	
Metrics Dashboard dashboard	2016-09-20 @ 7:44 PM elastic	completed 2016-09-20 @ 7:45 PM	

At the bottom of the table, there is a pagination control showing '<< < 1 > >>'.

Figure 3-23: Kibana reporting

Unit of Scale: A shard is the unit of scale in the Elasticsearch world (in fact, the smallest index the user can have is one with a single shard).

Shard Overallocation: A shard lives on a single node, but on the other hand, a node can hold multiple shards. Applications interacting with Elasticsearch communicate with indices (not with the shards). Having multiple shards for a single index improves the search and index performance (i.e. parallelism).

Replica shards: These are used in case of failovers (i.e. if the node holding a primary shard dies, a replica is promoted to the role of primary shard). But replica shards can serve read requests. If, as is often the case, the index is search heavy, the user has the option to increase search performance by augmenting the number of replicas.

Multiple indices: Using multiple indices when interacting with Elasticsearch data can boost the search and index performance. More explicitly, when a user issues a search request, it is forwarded to a copy (a primary or a replica) of all the shards in an index.

Event Correlation. The capability of correlating events of different types is an essential building block of a good SIEM. Even though Elasticsearch is not a SIEM, it offers event correlation capability that has its limitations in some cases. Event correlation is based on handling relationships.

Elasticsearch (similar to most NoSQL databases) treats the world as though it were flat. An index is a flat collection of independent documents. A single document should contain all the information that is required to decide whether it matches a search request.

This flat view of events or entities has its advantages (e.g. fast indexing and search, a large amount of data which can be spread across multiple nodes since each document is independent from the other documents).

In order to keep the advantages of the “SQL” and “NoSQL” features, Elasticsearch renders four common techniques to manage relational data:

1) Application-side joins: The user has the possibility to simulate a relational database by implementing joins in applications. After populating two different indices (with a primary key of an index as a field in the second index), the user can create joins by running two queries: one the first and one on the second index with the result of the first query as search term for the second.

It is worth saying that there are advantages to this method, most important one of application-side joins is that the data is normalized.

2) Data denormalization: This is the way to get the best search performance out of Elasticsearch is to use it as it is intended. In other words, by denormalizing the user data at index time. Redundant copies of data in each document eliminates the need to create joins. Instead of having useful data in different indices, the user can add all the useful fields in one index and (hence not worry about joins since all data fields reside in the same index). This method also has its advantages: speed.

3) Nested objects: Creating, updating and deleting, a single document in Elasticsearch is atomic, hence it makes sense to store closely related entities in the same document.

Similar to data denormalization (since all of the content is in the same document), the user has no need to create joins.

4) Parent/child relationships (similar in nature to the nested model): It allows the user to associate one entity with another. There is one difference though: in nested objects, all entities live within the same document but in the parent-child, the parent and children represent separate documents.

The parent-child option allows the user to associate one document type with another, more precisely in a one-to-many relationship (e.g. one parent to many children). The advantages that parent-child has (w.r.t. nested objects) are:

- Child documents can be added, changed, or deleted without affecting either the parent or other children. This is especially useful when child documents are large in number and need to be added or changed frequently.
- Child documents can be returned as the results of a search request.
- The parent document can be updated without reindexing the children.

Usually a good solution requires a mixture of these techniques. Moreover, a lot of efforts have been carried by the community to improve Elasticsearch joins capabilities. For instance, an Open Source plugin for Elasticsearch “SIREn Join” allows for fast joins across Elasticsearch indices.

3.8.5 Flexibility in security directives

Since the Elastic Stack is not a SIEM by itself, it does not support security rules natively. Therefore, the Elastic Stack does not have pre-configured set of rules or actions and it is up to the user to define security rules. However, Elasticsearch provides a full Query DSL (which relies on JSON) to define queries and therefore rules.

Querydsl [37] is an extensive Java framework, allowing the generation of type-safe queries in a syntax similar to the SQL one. It currently has a wide range of support for various backends using separate modules e.g. MongoDB, SQL, Java collections, RDF, JPA, JDO, Lucene, Hibernate Search.

As mentioned in previous sections, the Elastic Stack supports integration with most used programming language today but also a native integration with the popular Hadoop ecosystem. This allows creating simple advanced security rules.

As previously mentioned, Elasticsearch is based on the Apache Lucene search engine, hence it supports the Lucene Query Syntax [38], which is a rich query language based on operators and terms. This is quite useful for defining simple security rules but also to carry out investigations.

3.8.6 UEBA integration

ELK is NOT shipped with a UEBA plugins, but it has its flexible REST API, hence the UEBA solution integration to the Elastic Stack is possible. One (commercial) solutions meant to leverage ML anomaly detection and other behavioral analytics capabilities is Prelert.

Prelert Behavioral Analytics is designed to analyze log data residing in Elasticsearch, find the anomalies within that data, and links these anomalies together, giving the user the possibility to create a timeline of events.

Prelert is capable of faster root cause discovery and early detection of incidents (e.g. data exfiltration) but also controls communication in near real-time.

Prelert gathers more context than the usual monitoring tools that naively rely on a single source. The additional context greatly improves reducing the false-positive rate.

3.8.7 Risk analysis capacity

Elastic Stack does not provide any risk analysis capability natively.

3.8.8 Exposed APIs

Rest API. The ELK REST APIs are exposed using JSON over HTTP. This API presents a mechanism to manipulate and access data features from a specific product.

The indices can be easily manipulated by using the ELK API. The related methods are used to manage individual indices, index settings and templates, aliases, mappings and warmers. The indexes can be created, deleted, retrieved, opened, and closed. Similar methods can be used for indexes mapping.

A category of API is given by the search methods based on given queries. There are more search subcategories, such as:

- Shards search.
- Suggestions, counting or validation for the search functionality.
- Search based on specific request, under the forms of query, sort, filtering, fields, preferences.
- Search based on templates.
- Aggregation searches, used more for real-time data analytics providing standard methods such as min, max, sum, avg, percentiles, range or histogram.

Additional functionalities are provided by the API, used for cache cleaning, optimize indexes, refresh indexes or upgrading them to the latest format.

Creating New Connectors

Creating Logstash Plugins: Logstash plugins can insert, query, enrich, update and delete data into Elasticsearch.

The normal user can create his own Logstash plugin in a matter of seconds. The generate subcommand of bin/logstash-plugin creates the foundation for a new Logstash plugin with templated files. It makes the correct directory structure, gemspec files, and dependencies so that the user can start adding custom code to process data with Logstash.

Creating a Beat: The Beats can be defined as a collection of lightweight daemons that are able to collect operational data from the user servers and ship it to Elasticsearch/Logstash. The common parts for all Beats are placed in the libbeat library, (this contains packages for sending data to Elasticsearch and Logstash, for configuration file handling, logging, signal handling, etc). All Beats are written in Go (quite new language).

Very largely said, a simple Beat has two principle components:

- A publisher that sends the data to the specified output, such as Elasticsearch or Logstash.
- A component that collects the actual data, and

Elasticsearch Clients: Important about Elasticsearch, is that it is programming language independent. All the APIs for indexing, searching and monitoring can be accessed using HTTP and JSON so it can be integrated in any language that has those capabilities. Nevertheless, Java (the language Elasticsearch and Lucene are implemented in) is very dominant.

The user can manipulate the Java client to perform standard index, get, delete and search operations on an existing cluster, but also to make administrative tasks on a running cluster. Other clients like Python, Perl, Ruby, PHP, .NET and Groovy, can be found at the official Elastic website.

3.8.9 Resilience

As Elasticsearch clusters grow bigger, their resilience to hardware and network failure becomes very important. A lot of effort has been invested into making Elasticsearch and Apache Lucene both detect as well as cope with increasingly difficult failures.

The failures can be due different factors: network disconnections, long garbage collection, dead master, corrupted disk, timeouts, etc. In a distributed environment, failures are the rule and by far not the exception. Hence, Elasticsearch offers a set of features to ensure data resiliency:

- **Snapshot/Restore API.** The snapshot and restore module allows to create snapshots of individual indices/an entire cluster into a remote repository e.g. shared file system, S3, or HDFS. It is worth mentioning that these snapshots are perfect in case of backups sine they can be restored relatively quickly but they are not archival because they can only to be restored to versions of Elasticsearch that are able to read the index.

In some cases, closing an old index can be a viable alternative: indices get old, they reach a point where they are almost never accessed. We could delete them at this stage, but perhaps we want to keep them around just in case somebody asks for later.

These indices can be closed. They will still exist in the cluster, but they won't consume resources other than disk space. Reopening an index is much quicker than restoring it from backup.

ELK also has the option to use a shared file system repository (i.e. a shared file system to store snapshots). This as well as a snapshot repository (where both are storage containers) need to be registered. There is also an alternative to shared file system which is a URL repository which support the following protocols: ftp, http and https. The setting also supports wildcards, paths, queries and fragments. Certain repository plugins exist such as:

- repository-hdfs for HDFS repository support in Hadoop environments
- repository-gcs for Google Cloud Storage repositories

- repository-azure for Azure storage repositories
- repository-s3 for S3 repository support

ELK has certain limitations: only one snapshot process can be executed within a given cluster at any point in time. Also, while the snapshot is being created, the shard cannot be moved to any node. The downside of this process is that this can interfere with the allocation filtering process as well as with the rebalancing process.

- **Sequence Numbers.** Elasticsearch assigns a sequence number to operations that occur on primary shards. The most obvious one is speeding up the replica recovery process when a node is restarted. Previously, every segment has to be copied. Sequence numbers will allow us to copy over only the data that has really changed.
- **Multiple Data Paths.** It is possible to set multiple data paths when configuring Elasticsearch. The goal is to spread data across multiple disks or locations, thus limiting disk failures impacts.
- **Checksums.** Elasticsearch and Lucene under the hood is performing checksums on different data stages to ensure that data is not corrupted. The snapshot process verifies checksums for each file that is being snapshotted to make sure that created snapshot does not contain corrupted files.
- **About Elasticsearch Cluster.** A node is a running instance of Elasticsearch. But a cluster is a group of nodes with the same cluster.name that are working together to share data and are also meant to provide failover and scale, even though a single node can form a cluster all by itself. The distributed aspect of Elasticsearch is greatly transparent.

One node in the cluster is selected to be the master node. The master node is in charge of managing cluster-wide changes like creating/deleting an index/adding/removing a node from the cluster. The master node is not required to be involved in document-level changes or searches, which means that having just one master node will not become a bottleneck as traffic grows, since any node can become the master.

Elasticsearch can scale out to hundreds (or even thousands) of servers and it is able to handle petabytes of data. Here are some of the operations happening automatically under the hood:

- Partitioning the documents into different containers or shards. They can be stored on a single node or on multiple nodes.
- Balancing these shards across the nodes into the cluster to spread the indexing and search load.
- Duplicating each shard to provide redundant copies of data but also to prevent data loss in case of hardware failure.
- Routing requests from any node in the cluster to the nodes that hold the data the user is interested in.
- Seamlessly integrating new nodes as the cluster grows or redistributing shards to recover from node loss.

3.8.10 Event management and Visualization capabilities

Kibana is the visualization component of the Elastic Stack. Using Kibana, users can add line/ bar/scatter plots/charts/maps, to visualize large amounts of data. Kibana can be used not only for creating dashboards and visualisations but also for data discovery and interactive exploration by means of the “Discover” tab. It is worth mentioning that dashboards in Kibana can be shared and embedded as HTML documents. Furthermore, all Kibana objects are stored as Elasticsearch documents which is useful when it comes to cloning and updating visualisations using the Elasticsearch API.

The X-Pack is easily integrated with Kibana. It offers reporting and Elastic cluster health monitoring. Kibana also offers “Dev Tools” box for managing Elasticsearch indices and interacting with data using the Elasticsearch API.

3.8.10.1 Kibana visualisations

Kibana visualisations are the building block of dashboards. These visualizations are based on Elasticsearch queries. By means of a series of Elasticsearch aggregations to extract and process the user data, the user can create charts that show the user the trends, spikes, and dips the user needs to know about.

Kibana comes with nine types of visualizations: Data Table, Area Chart, Line Chart, Metric, Pie Chart, Markdown Widget, Timeseries, Tile Map, and Bar Charts. Possibility to add a new custom visualization to Kibana by means of JavaScript is offered to the user.

Advanced Kibana visualisations are based on the aggregation feature in Elasticsearch. The aggregations include all the faceting functionality and in the same time it provides powerful capabilities. The user can use different types of aggregations like bucketing and metric. The bucketing aggregations have the scope to create a list of buckets, each one containing a set of documents. Such aggregations are terms/data ranges/histograms. The metric aggregations are used to compute metrics over a set of documents, by means of mathematical functional and stats. Using faceted search, users may calculate aggregations of their data.

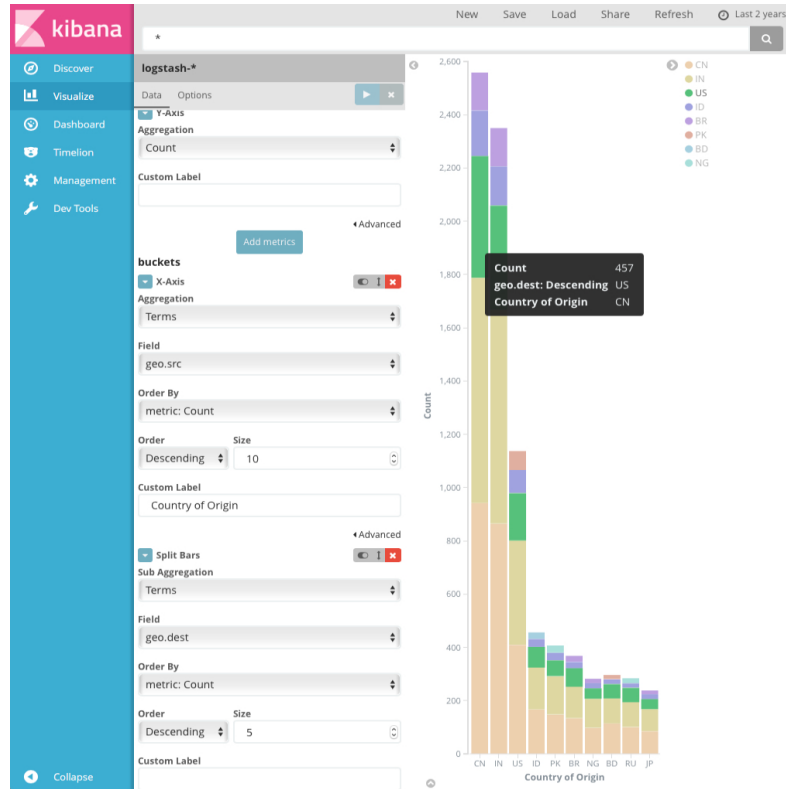


Figure 3-24: Kibana Visualize tab

3.8.10.2 Kibana Dashboards

All types of visualization can be merged in a very customizable dashboard. The user may arrange and resize the visualizations as needed and save dashboards so to be reloaded and shared (See Figure below).

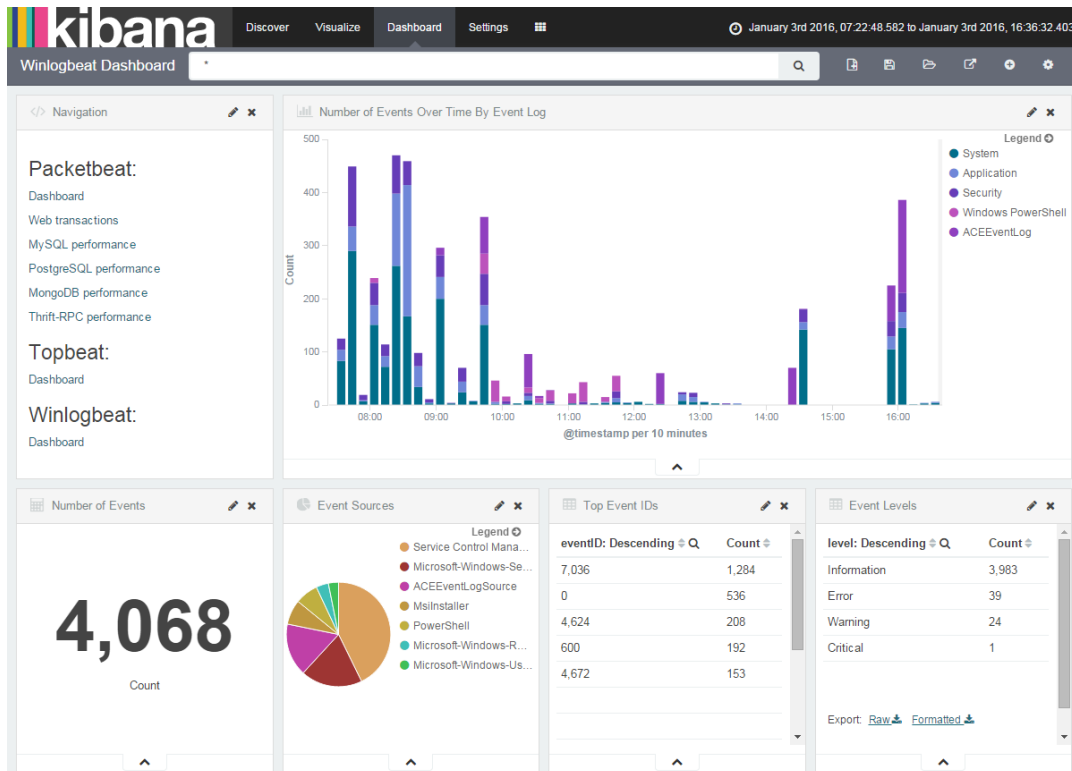


Figure 3-25: Kibana Dashboard example

3.8.10.3 Timelion

The Timelion is a time series data visualizer that enables the user to join totally independent data sources inside a single visualization (Figure 3-26). It's driven by a simple expression language the user can make use of to retrieve time series data, perform calculations, and visualize results.

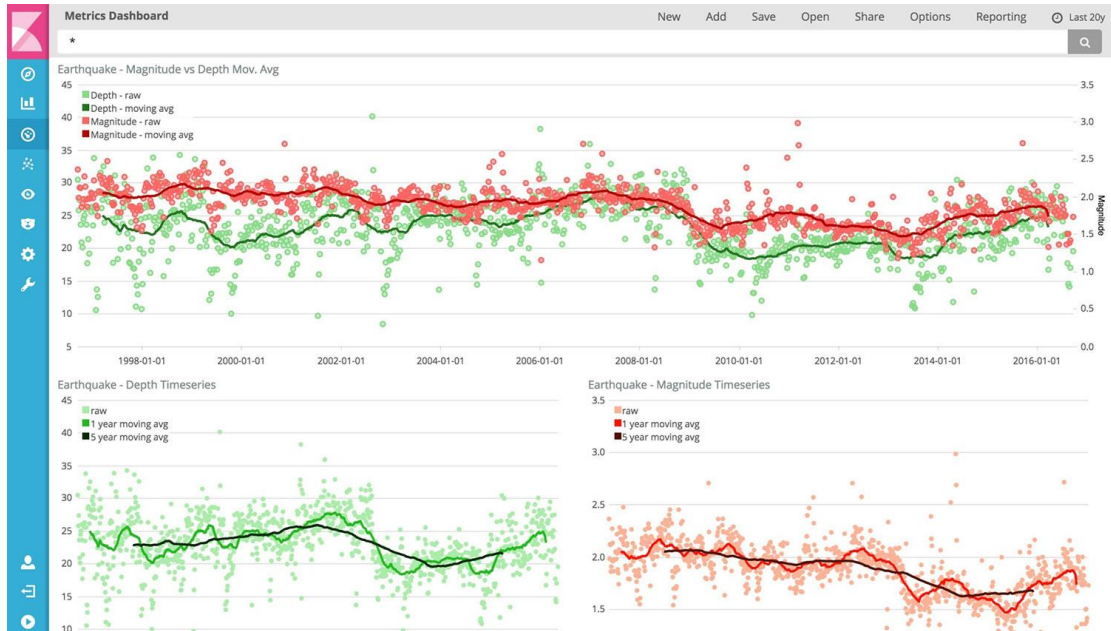


Figure 3-26: Timelion, a time series data visualizer

3.8.11 Reaction capabilities

An alerting extension can be added to the ELK stack called Watcher. This alerting module used to be a separate component but starting from the version 5.0.0 is part of the X-Pack [39] . X-Pack is an Elastic Stack extension that bundles security, alerting, monitoring, reporting, and graph capabilities. The X-Pack can alert about common cases, such as CPU usage increase or application response time spiking. The target of the X-Pack is to identify the changes in the data that the user is interested in. This feature is customizable for the user, proving great flexibility. Different alerting channels are supported by the X-Pack, such as email, media chats or other such environments. X-Pack provides also an alert history, so the user can track and visualize them in Kibana, as seen in the figure below:

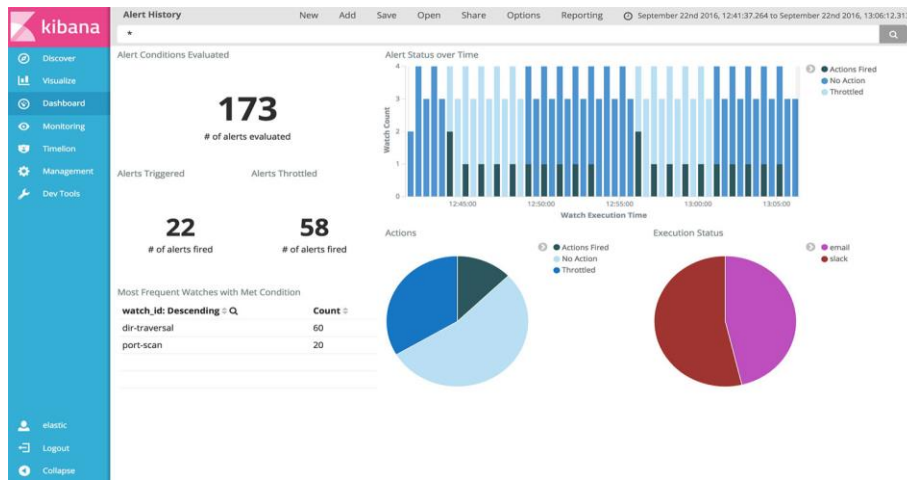


Figure 3-27: Kibana Alerting

Another Open Source alerting extension that can be added to the Elastic Stack is ElastAlert [40], developed by Yelp. ElastAlert offers great flexibility in terms of event trends by querying Elasticsearch data and leveraging the aggregation framework.

ElastAlert was initially a comprehensive log management system for the user's data. However, a rule engine was added to match patterns in the data. Based on the patterns, specific alert action will be triggered. For each rule from the set, the ElastAlert tool queries periodically Elasticsearch for getting relevant data in real time.

ElastAlert works by combining ELK with the rules engine and alerts. Based on the data pattern match from the rule, alerts are sent to the user. ElastAlert used the following rule types:

- Frequency type (events on time).
- Spike type (rate of events).
- Flatline type (event rate on time).
- Blacklist / whitelist type.
- Customizable filter type.
- Change type (data value changes on time).
- New term type (unusual value for a field).

Based on these rules, ElastAlert supports the following outputs: Command, Email, JIRA, OpsGenie, SNS, HipChat, Slack, Telegram, Debug and Stomp. Users can also visualize alerts in a Kibana dashboard.

3.8.12 Deployment and Support

In terms of deployment, Elasticsearch is easy to install immediately start indexing data. Depending on how big the cluster is, the complexity increase to deploy a cluster depends on it. Elasticsearch can both run on a simple laptop or on a small cluster of machines lying around. When deploying Elasticsearch to production, a few recommendations must be made. See in *Appendix II: Elasticsearch deployment* these recommendations.

Elastic provides consulting, support, and training. Types of support may be seen in Figure 3-28.

	OPEN SOURCE	BASIC	GOLD	PLATINUM
	Free Download	Free License	Request Info	Request Info
Support				
Support Coverage			Business hours	24/7/365
Response Times			Critical: 4 hrs L2: 1 day L3: 2 days	Critical: 1 hr L2: 4 hrs L3: 1 day
Unlimited # of Incidents			✓	✓
# of Support Contacts			6	8
Web and Phone Support			✓	✓
Emergency Patches				✓

Figure 3-28: Elastic Stack types of support

3.8.13 Licensing

All open source projects of Elastic are under Apache 2.0 License, like Elasticsearch, Logstash, Kibana and Beats. However, X-Pack is a commercial package.

The X-Pack has a 30-day trial license, which allows the user to have access to all the Elastic Stack features. The user has the possibility to purchase a subscription at the end of the trial period.

A summary of the Elastic Stack license subscription options and enabled features can be found in the official Elastic website [41].

All Elastic Cloud customers get access to basic authentication, encryption and role-based access control, in addition to monitoring capability.

Elastic Cloud Gold and Platinum customers get complete access to all the capabilities in X-Pack.

3.8.14 Position in Gartner Magic Quadrant

Although ES is not obviously visible in Gartner's SIEM Magic Quadrant Report, its technology is the basis for multiple SIEMs present in quadrant like LogRhythm, whose position is rather 3rd in Gartner's ranking, as can be seen in Figure 3-1.

LogRhythm has chosen in the last year to separate its log processing capabilities from its indexing capabilities and also to add a storage back end based on Elasticsearch in order to have unstructured search capabilities. This was, clustered full data replication was added. Elasticsearch incorporation allowed LogRhythm to become a very good solution for organisation that need integrated advanced threat monitoring capabilities in combination with its SIEM foundation.

In fact, one of LogRhythm strengths, as depicted by Gartner, is combining SIEM capabilities with UEBA, endpoint monitoring, incident management capabilities, endpoint monitoring and advanced threat monitoring use cases (partly thanks to Elasticsearch). See Figure 3-20 and Figure 3-29.

Gartner also comments on SIEM alternatives. It specifically points out ES in combination with OpenSoc and Apache Metron that use big data platform like Hadoop offer good analytics capabilities. Gartner's opinion is that an enterprise having sufficient resources to deploy and manage such a combination, would most likely develop a good solution for the company's use cases also having a lower price than existing SIEM UEBA solutions.

Elasticsearch is mentioned a second time by Gartner in his 'Critical Capabilities for Security Information and Event Management' for its Advanced Analytics capabilities as being one of the features that helps LogRhythm to be a successful SIEM (since Elasticsearch is handling the backend storage of this tool).

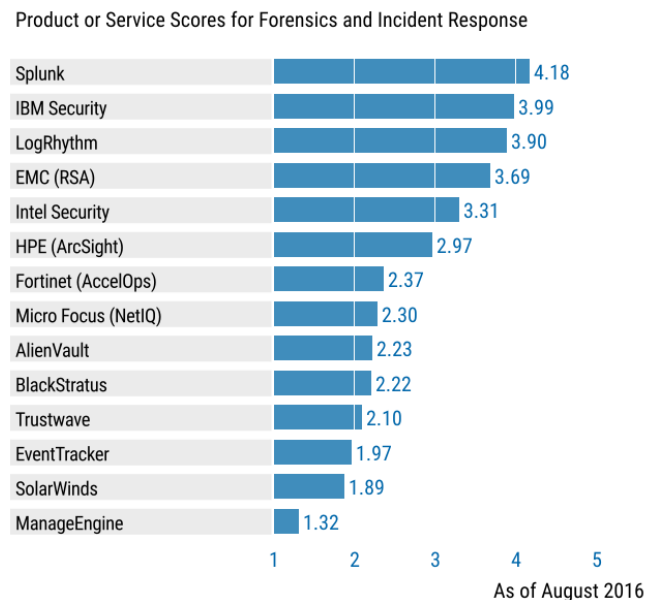


Figure 3-29: Gartner – Incident Response

4 Analysis of potential extensions to be developed in DiSIEM

4.1 Introduction

Most SIEMs support the integration of new connectors or parsers to collect events or data and provide APIs or RESTful interfaces to collect the events at a later date, as detailed in Sections 2 and 3 above. These extensibility mechanisms allow creating add-ons and extensions to existing systems. The DiSIEM project aims at exploiting this feature to enhance the quality of the events fed to the system (e.g., using new monitoring systems or collecting external data from open source intelligence) – through custom connectors, and provide new visualisation tools, collecting data from the SIEM data repository. Initial implementations may be in R or MatLab, or indeed any programming language, so we must address how these can be integrated into a SIEM, and what data they will use as input.

The set of components developed in DiSIEM will be validated through three pilot deployments in test and production environments, considering three different organizations, with diverse requirements, environments, infrastructures and technology. By the end of the project, we expect the DiSIEM technology to be ready to be supported either by the consortium members directly, startup initiatives spun out of the consortium, or third party associates.

This section assesses potential SIEM extensions, including metrics and modelling of security and risk using probabilistic methods and predictive models, the use of open source intelligence data and visualisation enhancements. Finally, the specifics of custom connectors, allowing diverse data sources to be used, along with cloud usage for long-term storage, will be investigated. The project proposal promised to:

1. enhance the quality of events collected using a diverse set of sensors and novel anomaly detectors;
2. add support for collecting infrastructure-related information from open-source intelligence data available on diverse sources from the internet;
3. create new ways for visualising the information collected in the SIEM and provide high-level security metrics and models for improving security-related decision project;
4. allow the use of multiple storage clouds for secure long-term archival of the raw events feed to the SIEM.

Since the purpose of this report is to present an in-depth analysis of the state of the art in SIEM systems, with particular focus on how such systems can be extended with custom connectors and new event visualisation tools, each section will propose ways to integrate the extensions into any SIEM, bearing in mind the various APIs and connectors exposed. This is informed by the details discovered and presented in Section 2.

4.2 Security metrics and probabilistic modelling

4.2.1 Risk assessment and security metrics

In work package 3, multi-level risk assessment and security metrics will be developed. The defined security metrics will assess security characteristics that are of interest for the operational and managerial security decision making, and will apply quantitative, probabilistic methods to support decisions on how best to combine multiple defences given a threat environment. This involves understanding how the strengths and weaknesses of diverse defences add up to the total strength of the system. Enhancing SIEMS with diversity-related technologies provides a major innovation of this project. Special attention will also be paid to “diversity” measures – i.e., how similar or different security protection systems, vulnerabilities, attacks etc., are from each other. These types of diversity metrics are less studied in the literature compared with metrics for individual components.

There is no common agreement on the definition of security metrics. Jaquit, 2007 [42] defines “metric” as “a consistent standard for measurement” and proposes five requirements of a “good metric”. The Center for Internet Security (CIS) report, 2016 [43] describes security metrics as “unambiguous definitions for security professionals to measure some of the most important aspects of the information security status”. The CIS twenty-eight metrics cover seven business functions: Incident Management, Vulnerability Management, Patch Management, Application Security, Configuration Management, Change Management and Financial Metrics. Recently, Yasasin and Schryen in 2015 [44] survey proposals of security metrics, observing a lack of consistent requirements against which these can be evaluated and also derive requirements for “good metrics”. A common fallacy about metrics is confusion between what they measure and what they may allow to be predicted, and failure to assess their predictive value before prescribing them to decision makers. DiSIEM will address this issue by validating predictions in real-world environments.

Organizations positioned at the highest maturity level in cyber defence and SIEM approaches operate with whole organization’s IT security mobilized with enterprise risk focus [45]. Linking security with business risks is a factor of differentiation of mature organizations, in addition to the monitoring of critical applications, the promotion of people awareness, and the adoption of strong governance strategies. In particular, top management is mainly concerned with monitoring high-level security metrics, especially for risk analysis: “Potential loss of critical assets” and “Current level of risk by critical area” (in monetary units) [46].

The true value of a SIEM platform will be in terms of Mean Time To Remediate (MTTR) or other metrics that can show the ability of rapid incident response to mitigate risk and minimize operational and financial impact [47]. Many of the metrics currently presented by the SIEMs (e.g., number of failed logins per day, number of access denied requests) are targeted for a technical SOC’s audience, and hence may not be adequate to support management-level strategic decisions

within an organisation. Therefore, there is a need to integrate effective high-level security metrics in SIEMs. Although some proposals already contemplate aggregating metrics considering conditional probabilities of attack in the network [48], state of the art SIEMs do not include these capabilities.

Commercial products that integrate with SIEM (e.g, AlienVault products [49]) provide dashboards for monitoring several indicators in categories such as: Security events (most frequent components); Accountability and statistics on tickets; Taxonomy (distribution of types of events); Accountability of vulnerabilities, etc. Although some risk analysis features are provided, the risk model is simple. Risk values are assigned to events depending on the asset value (manually tuned using a high-granularity scale (1 to 5)), the event priority (importance) and probability of an attack. The risk score of an asset is the sum of the risk of each event on that asset

IBM Security Qradar Risk Manager [50] and IBM Security Qradar Vulnerability Manager [51] integrate with IBM Security QRadar SIEM [52]. Qradar Vulnerability Manager is a network scanning platform that detects vulnerabilities within applications, systems, and devices on the network, and, additionally, provides asset discovery. Qradar Risk Manager obtains events, context and flow data, and offers integrated risk management. The risk score of an asset is the sum of the risk scores of all vulnerabilities on that asset [53] where the risk score of a vulnerability provides specific network context by using the Common Vulnerability Scoring System (CVSS) base, temporal, and environmental metrics [54]. The network risk score is also computed by an additive function that considers the network topology. IBM QRadar has a considerable dependency of third-party technologies, especially to Endpoint monitoring for threat detection and response, or basic file integrity. Experience has reported some problems related with the integration of the Vulnerability Manager module with the rest of the system.

ArcSight [55] does not provide a specific module for risk assessment or vulnerability scanning. Therefore, these procedures must be performed by external applications. Vulnerabilities externally discovered can be uploaded (as events). Risk evaluation is made at the event level and is given by the event “priority”, based on threat detection [56]. The purpose is to determine if a threat is trying to exploit a vulnerability. The priority formula uses four distinct variables: Model Confidence, Relevance, Severity and Asset Criticality. The Model Confidence variable assesses the information that the system has about the target asset (asset under evaluation). The Relevance variable considers if a target asset has an exploitable vulnerability, which may be exploited by the event action, and the accessibility to the target (port opened or not.) To a better insight of this variable, if the action on the event is a scan, on a port or vulnerability, the importance or relevance for the system is low, but if the port is open and there is an open vulnerability, the relevance score is high. The Severity variable considers not just if the target has already been compromised, but also, if prior activity from this same event source has been recognized. Finally, the Asset Criticality is responsible for measuring the importance of the target asset in the organization context.

A major advance of the DiSIEM project when compared with the state of the art will be the development of useful operational metrics that allow SOCs to make decisions supported by quantitative evidence, where uncertainty in the measures is explicitly stated, and with better visualisation support to enable better communication of these decisions to the relevant stakeholders in the organization (e.g., SOC operators, CISOs, CIO, CEO and board members). We expect to support measurements on several layers of defence (e.g., Firewalls, IDSs, AntiVirus products, Operating Systems, applications) and different products of each type: the measurements would be provided by the consortium partners as well as from open source and experimental datasets obtained within DiSIEM and related projects.

Kotenko et al. in 2013 [57] propose an ontology including metrics related with the network topology. Here, the Attack Impact can be calculated taking into account a relative importance of assets, a static cost of assets (when the cost is assigned statically to every asset) or a dynamic cost (when the cost is propagated via service dependencies). In a simple static case to define the Attack Impact, the Attack Propagation metric is suggested, which attempts to capture the damage that may result from any host in the network that attacker can reach. In addition, to calculate the Attack Impact we also consider such metrics as Response Effectiveness, Benefit of the Response and Response Collateral Damage, which form a sub-group Response metrics. Other component of the ontology concerns cost metrics, which include the expected annual loss, the total gain and the return on investment from reaction to an attack.

The reviewing of existing SIEMs allowed us to conclude that these systems do not provide high-level security risk metrics. DiSIEM will pursue the development of risk-based metrics considering several layers of dependencies such as hosts, applications, middleware, and services. These will allow scoring risk for the different operational and functional areas. The Attack Propagation and Attack Impact metrics [57] will be extended to consider different hierarchical operational layers. Though cost metrics can be hard to compute due to the difficulty of organizations in estimating security costs, one of our purposes is to approach this category of metrics using high granularity estimation of costs.

Requirements for the extension

For the development of high level security metrics and of multi-level risk assessment it is necessary to have access to the following information: open vulnerabilities and incidents; and for each vulnerability/incident: the type of vulnerability/incident, target and destination information, open and close date of the vulnerability/incident, severity and impact, host, department or structure associated with the case, and priority of the event, whenever available. Other requirements include: classification of vulnerabilities, information about the assets (including owners and value classification), the network topology, all the information about the layer of applications and dependencies between applications and hosts, as well as services dependencies.

Figure 4-1 displays the data collection and workflow with ArcSight by EDP SOC. All information regarding incidents and vulnerabilities is collected in a Master Database. Information about vulnerabilities is produced by sources that scan and classify vulnerabilities, while incidents are recorded from ArcSight ESM. This Master Database is the input for an application (produced in-house) that computes several indicators and displays those using specially designed dashboards. The data collection and workflow structure in Figure 4-1 can be used with other SIEMS and therefore will be considered for the advances in work package 3.

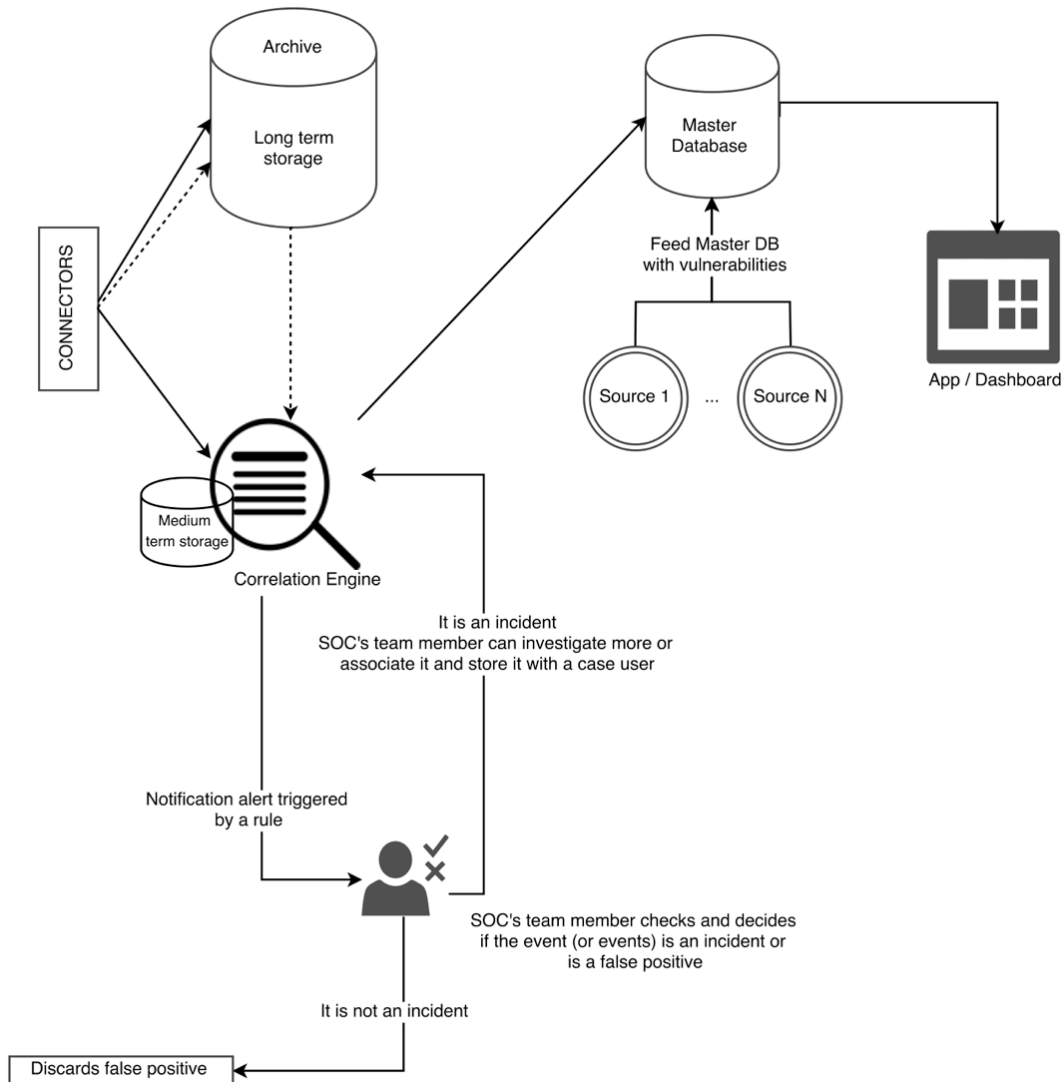


Figure 4-1: Data collection and workflow with ArcSight in EDP

4.2.2 Probabilistic modelling for diversity and defence in depth

4.2.2.1 Modelling to support on-line operational decisions

An important part of design for security is defence in depth: “layers” of defence that reduce the probability of successful attack or at least try to contain its effects. For example, an attack needs to pass one (or more) firewalls and go undetected by one or more Intrusion Detection Systems (IDS) before it can exploit a vulnerability in a protected host to attack an asset. Guidance documents

now advocate defence in depth as an obvious need, and SIEM deployments also use diverse tools (e.g. OSSIM uses both Snort and Suricata IDSs). But their qualitative guidance ignores decision problems, e.g. justifications like “Even the best available Information Assurance products have inherent weaknesses...” it is only a matter of time before an adversary will find an exploitable vulnerability”¹⁸ ignore the probabilistic nature of security. What matters is not that adversaries will eventually break through – this applies to layered defences too. It is how soon they are likely to break through. Added layers of defence postpone that moment, directly by requiring extra steps, and indirectly by allowing for detection and defensive steps. Quantifying and assessing by how much requires some thought.

Specifically we will address the use of diversity, including but not limited to the use of multiple intrusion detection systems (IDSs) and disparate open source intelligence data. There has been only sparse research (e.g. [58] [59]) on how to choose among alternative layered defences; occasionally, unsuitable models appear relying on the naive assumption of independent failures between the diverse components [60]. Security engineers have little or no theory to guide their decisions about diversity, although unaided intuition can be very misleading here (e.g. Littlewood and Wright [61]). Previously, DiSIEM partners have studied modelling, assessment and design issues in diversity and software fault tolerance including the use of diverse anti-virus software in malware detection.¹⁹ We will investigate ways to send diverse inputs to SIEMs and consider approaches to generating rules, for example voting systems, and assess their performance. The literature touches on the use of ensemble methods to assess the results of classification systems for security [62] however our focus is diverse inputs rather than the aggregation of diverse machine learning techniques. One paper considers possible combinations of two IDSs [63] and presents ways to evaluate their performance either in series or in parallel. More generally, the security community is aware of diversity as potentially valuable (earlier references [64] (section 4); more recent ones in [65] (section 2)). Discussion papers argue the general desirability of diversity among network elements, like communication media, network protocols, operating systems etc. Research projects studied distributed systems using diverse off-the-shelf products for intrusion tolerance (e.g., the U.S. projects Cactus, HACQIT [66] and SITAR; the EU MAFTIA project²⁰). New uses of diversity appear every now and then, e.g., diverse anti-virus software supplied “in the cloud” by VirusTotal²¹, Metadefender²² and CloudAV [67], metrics for effectiveness of defence in depth [68], etc.

SIEMs already provide the functionality for reading logs from multiple different security monitors and detection tools at different layers (cf. Sections 2 and 3 of this deliverable). In DiSIEM we will build tools that allow consolidation of

¹⁸ <http://www.nsa.gov/ia/files/support/defenseindepth.pdf>

¹⁹ See <http://www.city.ac.uk/centre-for-software-reliability/research/diversity-for-further-details>

²⁰ <http://research.cs.ncl.ac.uk/cabernet/www.laas.research.ec.org/maftia/>

²¹ <https://www.virustotal.com/>

²² <https://www.opswat.com/metadefender-core>

outputs from multiple diverse monitors of similar type, which may be monitoring similar types of assets. This will help in improving the accuracy of the detections, and reducing the false alarm rates that are reported back to the SOCs.

4.2.2.2 Support for off-line decisions

Additionally, at present there is a lack of proper support to enable decision makers to effectively assess the benefits they get from using multiple monitors and protection tools within their organisations. Most organisations do not use a single tool to implement their security defence, but rather use multiple tools in a defence in depth strategy. The important question facing a decision maker (such as CIO) who needs to make procurement decisions is about whether, for example, these specific three layers would improve security more than those two; and – if possible – how to quantify these security gains. Providing answers to these questions, and implementing solutions that allow existing SIEMs to give this kind of information to analysts is an important contribution of DiSIEM. Crucially, these questions concern the evaluation of diversity: layered defences should be diverse in their weaknesses. Any attack that happens to defeat one defence should (with high probability) be stopped or detected by some other one²³. Diversity and defence in depth are two facets of the same defensive design approach, which this contribution means to support and integrate to SIEMs.

We will also produce methods to answer questions about which diverse combinations of monitors and protection tools should be deployed in a given environment. These questions will be answered (a) in probabilistic terms, as is inevitable since many vulnerabilities, and the future behaviour of attackers, are unknown and (b) with clarity about the uncertainties involved and thus how much confidence can be put in these answers. While there has been a strong demand for quantitative support for decisions in security for a while²⁴, many claim that this is impossible. To summarise arguments we developed elsewhere [64], blanket objections ignore that quantitative probabilistic reasoning is a means for reasoning rationally about uncertainty, not for eliminating it. Overall, we aim at developing both:

- Conceptual probabilistic models, for insight – showing the relative importance of different factors in a diverse, layered design, theoretical bounds on benefits of diversity, comparing the effectiveness of different designs, etc. Generally, such models do not yield numeric predictions, since their parameters cannot usually be estimated for a specific system;
- Operational models, typically covering less detail but such that it is feasible to estimate their parameters from observation and to use them in operation for security assessment and prediction.

²³ Some authors use the term “defence in breadth” for defences that are complementary in being meant for different types of attacks (“covering a broader front”), rather than “happening to” differ in how well they cope with them. In practice there is no sharp boundary but a range of “weaknesses” of defences, from intentionally not covering certain attacks, to accidentally doing so, with deterministic or random effects. We will study effects of diversity on any combinations of these issues.

²⁴ See for example:

<http://webhost.laas.fr/TSF/IFIPWG/Workshops&Meetings/55/workshop/04.Sanders.pdf> and references therein

Operational models extrapolate to the future the patterns and trends observed over a period of previous observation. They help to configure defences against “normal” threats – the “known unknowns” of the evolving malware market used by ordinary attackers. They will not predict disruptive, first-of-their-kind events (e.g., massive surprise attacks by nation states). For these, the conceptual models allow at least “what if” reasoning about a system’s resilience given attack scenarios.

We will pursue two directions in probabilistic modelling of security through defence-in-depth:

1. Probabilistic models of efficacy of diversity involving diverse layers of defences (e.g., AntiVirus products, Intrusion Detection Systems, Firewalls, operating systems, applications etc). This would allow combining and analysing trade-offs in the security tools feeding data to the SIEM (e.g., “Can a combination of AntiVirus A, with IDS B and Firewall C give me a better protection compared with AntiVirus D, IDS E and two firewalls G and H?”).
2. Extending from measurement to prediction of security of a system that employs diverse layers of defences. There has been much work on measuring and estimating various security parameters and attributes, but not enough emphasis on prediction. This would allow us to enhance the SIEMs capabilities towards making predictions on the level of security organisations can expect in the future (hence being able to answer questions such as “with this combination of protection tools, the probability of observing a security incident of type X is Y with confidence Z”). Validation of predictions is hard, but within DiSIEM partners there has been much success with validation of predictions of reliability and safety which make us confident in producing innovative contributions for security also. By collecting all the data and holding it for longer time periods, DiSIEM will be able to assess and validate the metrics and models produces and find ways to gain many insights from the data, models and metrics. This if further expanded in Section 4.4 “Visualisation capabilities” below.

In summary, a major advance compared with state of the art will be the development of operational tools that allow these models to be used in practice by SOCs integrated with SIEMs that they are already familiar with, hence eliminating the need to learn a new tool or platform.

4.2.3 Statistical analysis for anomaly detection

DiSIEM will explore and implement novel unsupervised techniques that combine statistical analysis and multi-criteria decision analysis to automatically model applications and users’ behaviours and subsequently identify anomalies and deviations from known good behaviours that are statistically relevant. This will lead to the deployment of enhanced application monitoring sensors, which will feed SIEM systems with diverse types of events that can be correlated with more traditional security events collected from host and network-based appliances.

These new sensors will be based on an unsupervised statistical learning approach to automatically model users' behaviors and highlight anomalies or significant deviations from the previously learnt behavioural patterns. These sensors will then generate events that will be sent to and analysed by the SIEMs. By combining the anomaly-based events with those provided by more traditional heuristic and signature based tools and the data from OSINT (see WP3), we expect that we can improve the false positive rates of these components, which have traditionally been the main stumbling block of their wide adoption in real operation.

4.3 OSINT data fusion and analysis

The use of OSINT for feeding a SIEM with information about security threats and about the identification of trends over time, considering a monitored IT infrastructure, requires the development of the following SIEM extensions:

4.3.1 Data Collection

An information extraction extension is required to manage information sources and the data collection process. In terms of collecting information from the various social media sources, the related works that have been done so far have used crawlers in conjunction with parsers to extract information from the web pages of blogs, forums, marketplaces and other relevant sites [69] [70] [71]. Nunes et al. [69] included a noise filtering step which used a classifier for filtering out irrelevant data and used specialized parsers and crawlers for forums and marketplaces on the dark web as the structure of these sources of information are both different. In some cases, such as twitter, the data could be collected through their API. As such, there will be a need for specialized crawlers and parsers for the various sources of information that do not have APIs (see Figure 4-2).

DigitalMR collects information from a variety of social media including twitter, facebook, blogs, forums, news and other various other openly available data on the Internet for market research. These are typically tagged with information pertaining to relevance, sentiment, and emotion. This data is also tagged with information that classifies these data in terms of a taxonomy which gives an idea relating to the topic of the data's content. For example, a tweet about two people drinking Pepsi while watching a football match will be classified to be an 'occasion', and specifically a 'sport' occasion. Essentially, this is a form of hierarchical clustering which can also be done for cyber-threats. There is already a taxonomy of the types of cyber threats with some of the main types including asset exposure and vulnerability exploitation [72] [73]. Such information could also be tagged with the training data meant for the cyber threat detection model which will make the reports of cyber threats more specific. The tags will also allow the cyber threat detector to learn and identify future events with more specificity. Furthermore, another advantage of tagging in relation to a taxonomy is that it can also be used to identify specific trends, e.g. seasons when some threats are more prominent than others, and any other related trends.

4.3.2 Data storage

One way to implement data gathering from OSINT sources is to run the crawlers at intervals to capture information over time and dump all these diversely structured data into an Amazon S3 bucket or into alternative cloud-based storage systems as proposed in DiSIEM. This gives a temporal dimension, which keeps track of content over time. The sampling could be done in such a way that previous samples of the sources slowly get irrelevant and archived or removed from the data store, thereby giving the effect of sliding window. On the other hand, data from sources with an API like twitter can also be streamed into an S3 bucket with the help of Amazon's Kinesis. One approach to aggregate the data could be to aggregate in relation to time such that each data source within the same time interval ends up in the same time slice. These slices of data can be built by a lambda function – on intervals, or upon receiving new data – which will aggregate the various forms of data. This not only gives the data a context of what is happening in the various sources of OSINT data sharing the same time slice, it also gives a context of how all the content is changing over time. Other approaches that aggregate over some relevant property of data could also be explored in the before implementation – if necessary. For example lagging some of the data sources that generate data faster (i.e. has more velocity, such as twitter) so that the same information is not split up in different time slices, which can affect the performance of models that capture the temporal dimension with sliding windows.

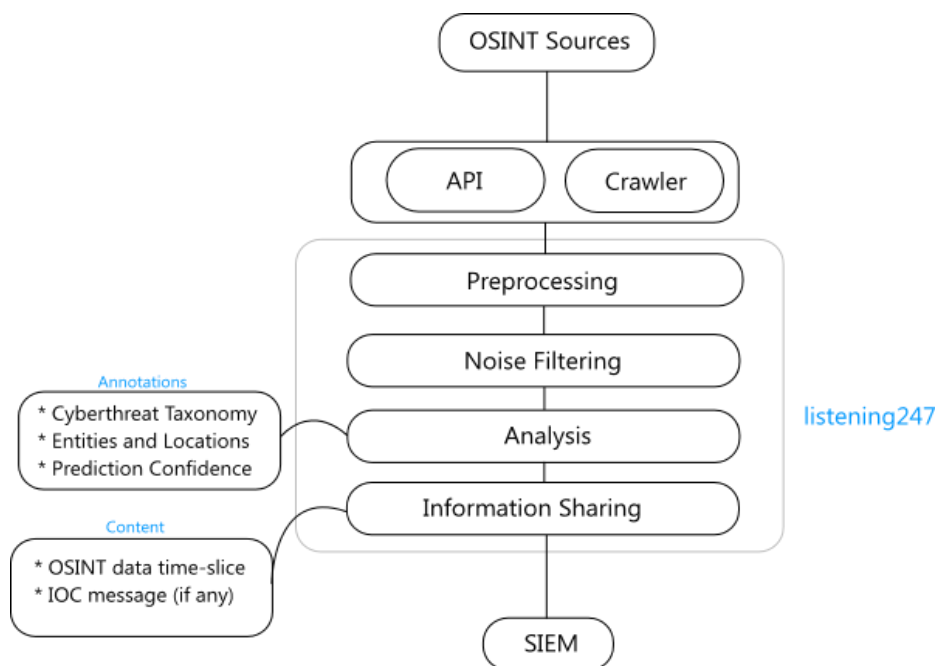


Figure 4-2: OSINT data processing using listening247

4.3.3 Monitored IT infrastructure specification

An extension is required to gather and manage information about the hardware/software assets in the monitored IT infrastructure. The information could be input by SIEM operators but could also be gathered automatically by

specialized sensors deployed in the IT infrastructure. The validity of the information gathered should be reviewed periodically as this information would be correlated with the OSINT data extracted from the various sources in order to filter relevant security threat information into the SIEM.

4.3.4 Threat detection

This extension has the purpose of classifying OSINT information as a threat to the monitored infrastructure or not. As set out in DiSIEM proposal two approaches will be followed: one more experimental where deep learning algorithms and models will be developed; another on the basis of DigitalMR listening247 platform.

DigitalMR's listening247 platform can be used as the model used to classify threats, as well as filter out noise from the information from the various sources of data with the help of relevance tags (see Figure 4-2). A potential restriction that will affect the accuracy of the classifier in predicting threats will be the availability sufficient OSINT data that relating to cyber threats. There is a need for getting data in major languages that relate to cyber threats which can then be used as training data for the threat predictor. A potential solution could be the use of language processing to identify threats from the use of keywords that will typically indicate a threat in major languages; such as 'ddos', 'security breach', 'leak' and more. This information can be used to tag OSINT data as relevant or irrelevant, which can then be used as training data for the threat predictor. In addition to the type of threat, other information from the OSINT sources such as location and entities involved could also be extracted to provide a more comprehensive description of the threat. The prediction confidence of the classifier can be included in the data sent to SIEMS, which will help avoid the issue of false alarms.

Predictions for possible threats based on data could be made using a recurrent neural network, or support vector machines with the lagged version of the dataset to predict the likely of threat. The predictions could include information about the likely nature of the threat, location and entities that could be evolved can also be included. Manning et al. [74] used support vector machine for classifying whether or not there is an exploit. This was accompanied by feature selection was applied to help increase the prediction accuracy. However, this was just doing binary prediction.

In our case, there will be a need for more training data to be able to predict with more specificity. One approach to overcome this is the use of reinforcement learning where a threat prediction message is sent to the SIEMS and sparingly requiring that a feedback be given of the prediction. These feedbacks are then used for training the model to make more accurate predictions. Reinforcement learning is frequently used in scenarios where there is not a lot of training data, and there is a need for quickly adapting to a dynamic environment, such as in real time systems. It is currently being used by the University of Southampton to optimize performance and energy efficiency of multicore ARM processors.

Listening to data from OSINT sources also shares similar characteristics, which makes it an ideal candidate for this application.

4.3.5 Incidence reporting and analysis

On the topic of incidence reporting, one or more of the upcoming frameworks for describing threats, such as STIX or OpenIOC can be used as a threat description framework for listening²⁴⁷ to report threats to the SIEMs. Information about the nature of the threat (i.e. type of threat in the threat taxonomy), and the prediction confidence of the classifier could be included in the description (see Figure 4-2). Further information about entities and locations relating to the threat can also be included (when available) to add valuable information necessary for an effective response.

The threat detection results and related information should be analysed over time-based sliding windows of OSINT data with the purpose of unifying similar detections and of building a trend analysis of security threats in the analysed period.

A possible use case for the predictions could be to adjust the level of scrutiny if there is the danger of an imminent cyber threat. It also helps in putting in place measures to prevent or manage the attack before it starts. In general, there is a lot of potential for more informed channeling of security resources with cyber threat prediction. OSINT data captured in various languages, and trained using a mixture of natural language processing and reinforcement learning techniques provides a wide coverage of information for more effective handling of cyber threats.

4.4 Visualization capabilities

During the reviewing of the state-of-the-art of existing SIEMs, we observed that the reporting and data visualisation capabilities are limited in terms of supporting the effective extraction of actionable insights from the huge amount of data being collected by the systems. Although all the systems offer data visualisation capacities to their users, most often, the visual representations are generic, not designed with particular user needs in mind, or even highly rudimentary to have any significant effect on how the generated data is utilised.

To add to that limited capacity, existing systems do not have the capacity to utilise the diverse data modalities that DiSIEM will be generating, such as statistical modelling outputs, OSINT data collections, or comprehensive models of user behaviour. These novel data facets, when combined with the data that is already being gathered, offer challenges and opportunities that we want to build upon and investigate within DiSIEM. To enhance the visualisation capability of existing systems, we identify the following goals:

- Design and develop a rich set of specialised visualisations that can handle the diverse types of data that we will be analysing within DiSIEM. The types of visualisations we anticipate will be able to visualise multi-faceted data that is:

- high-dimensional (i.e., each data object has several attributes associated, both from raw data and from the modelling process)
- temporal (i.e., data objects (events) often come with time information crucial within analysis)
- textual (i.e., certain facets will involve unstructured information, in particular the OSINT data)
- relational (i.e., relations within users, sessions or actions that are forming (probabilistic) network structures)
- spatial (i.e., physical locations attached to the events)

In addition to being able to accommodate these multiple aspects concurrently, we will aim to design visualisation solutions that provide: effective overviews, interactive capabilities to focus on details, and mechanisms to compare individual and/or groups of data instances.

- Design and develop visualisations that can be capable of handling the dynamic nature of the data (e.g., streaming system activity logs, OSINT data, etc.) to support real-time analysis and decision-making.
- Develop accessible, informative and holistic visual representations of the computational infrastructure that will enable stakeholders to evaluate the performance of their system and evaluate it in comparison to other diverse configurations. One very critical and important question for users of SIEMs is to evaluate whether their investments in the infrastructure is paying off. We plan to offer such capability through forming “what-if” scenarios where analysts might be able to concurrently investigate diverse models of the systems. The goal is to help SOCs and C-level managers to make effective decisions in building resilient systems.
- Develop novel visual analysis methods to enable SOCs to evaluate the spatio-temporal characteristics of security related events and how they relate to the spatial configuration of their infrastructure. Both the system deployments themselves and the attacks at these systems often exhibit a geographical dispersion and this aspect becomes a critical component whilst evaluating a system’s status [75]. We aim to make sure that this inherent spatial variation, together with the temporal changes in these events are represented to enable spatio-temporal analysis of the data.
- Develop interactive methods to elicit experts’ knowledge to optimise the predictive and probabilistic models (as discussed in the two sub-sections above). Although algorithmic models are highly powerful and are extremely critical in discovering relations and finding patterns in a system, there are problems where users’ expertise and knowledge about the system can improve the computational models’ performance significantly. We will include intuitive interaction mechanisms to provide experts to communicate their prior knowledge and preferences to the algorithms.
- Develop methods to visually perform a “forensic investigation” (a task that we have identified during our consortium meetings) of an attack and trace the “provenance” of the incident together with the context how, where, and when it occurs, e.g., which parts of the system are involved, which users/sessions are involved, when the incident have occurred along with the path of actions that are taken within the system as a whole. The visualisation team will work along with SIEM users (EDP, Amadeus, ATOS) to identify the

relevant sources of information and the types of events during the requirement elicitation process for visualisation design.

- Develop techniques to communicate the uncertainties within the estimations and modelling. Since DiSIEM will involve several computational tools and mechanisms, such as probabilistic modelling and anomaly detection, user-behavior modelling, the results often come with a particular level of uncertainty [76] and these uncertainties may impact significantly the subsequent decisions. We aim to design visual representations that are aware of these uncertainties and communicate them to users to avoid potential pitfalls.
- User behaviour models provide knowledge of common user behaviours and support detecting suspicious activities from users. We aim to apply visualisations to allow analysts gaining deeper understanding of the models before validating and improving them. One approach is to develop visual summary of user activities that quickly reveals common/abnormal patterns in a large set of user sessions, compares multiple sessions of interest, and investigate in depth of individual sessions. The visualisation will be closely linked with the model to be able to identify correlation between the model output and visual patterns.

Overall, the major enhancement will be a visualisation system that is flexible enough to work with several data sources that carry heterogeneous characteristics and with data that is under constant change, i.e., real-time streaming data. Moreover, we aim to develop novel representations to communicate the provenance of an attack and the characterisation of sessions/users that will enable an analyst to better profile the system, ongoing activities and its vulnerabilities. We aim to achieve these with close collaboration with domain experts and tailor the design of the methods towards the particular needs and tasks of SIEM users.

Because many existing SIEMs are web-based, we also plan to develop our visualisation system in a web-based environment. More technically, we will build our visualisation using the powerful d3.js²⁵ library. Many SIEMs base on the Elastic stack²⁶, which use kibana for their visualisation. Internally, kibana also uses d3.js to implement its visual components. Therefore, at the tightest level, our visualisation system can be directly integrated into Elastic-based SIEMs as kibana plugins with little programming effort, enabling high adoption. For other web-based and desktop-based SIEMs, it is still possible to integrate our visualisation system with them by transferring data/event through the internet. The visualisation can read streaming data from SIEMs and provide real-time monitoring and situation awareness, as well send back to SIEMs any needed information so that they can promptly act on the insight revealed by the visualisation.

²⁵ <https://d3js.org/>

²⁶ <https://www.elastic.co/>

4.5 Infrastructure enhancements

4.5.1 UBA

Future improvements include creating sensors that would rely on unsupervised statistical learning approaches to firstly create a baseline for normal entity behaviour (users and applications alike). The scope is to be able to highlight anomalies and/or deviations from this pattern by using a SIEM scoring-alerting system.

At first, the solutions will be application based, later to be developed as more generic. Also in terms of UBA, a set outlier detectors or classifiers as well as other unsupervised machine learning algorithms will be used in order to manage user/application profile.

4.5.2 Event consolidation

With respect to previously defined metrics and models within the DISIEM project, the scope of the event consolidation is, not only to implement them, but also to operationalise them in a smart way. Due to the extent of implicit uncertainty of the random variable prediction, we want to have an “as-clear-as-possible” assessment of them.

As part of this scope, we have the integration of multiple sources as a single SIEM (correlated) event source. By multiple data sources, we mean: AntiViruses, Intrusion detection systems, Vulnerability scanners, Firewalls and so on. This will serve as complimentary (enriching) data. This will later serve as prototype in the test, then real environment.

The prototype will be delivered with documentation and setup instructions.

4.5.3 Storage capabilities enhancement

SIEM solutions are designed to collect, process and correlate large amounts of data for continuous monitoring and incident response. Analysing large volumes of logs is important to gain more insights from the data collected and to have a better monitoring, however keeping large volumes of collected data in a live SIEM system is impractical, costly and can impact the effectiveness of incident investigations. To mitigate those issues, SIEM solutions have different strategies when it comes to treating aging data.

Splunk for example offers an automatic archiving of data as it ages, i.e.: data rolls to “frozen” state. By default, indexers in Splunk will automatically delete “frozen” data, but indexers can be configured to archive the data before removing it. Splunk offers two ways of handling the archiving: indexers perform the archiving automatically or by specifying an archiving script. The first option will copy frozen buckets to a specified location while the second option offers more flexibility both in terms of actions to apply to frozen data before archiving and different archiving locations that can be specified: HDFS, Amazon S3, shared storage, local disk, etc. Of course, archived data can be restored on demand by

copying archived buckets to the data directory in Splunk. However, OS versions and Splunk indexer versions can make some restrictions on archive restoring.

For IBM QRadar, backup and recovery feature can be used to back up event and flow data. However, restoring event and flow data can only be done manually. By default, IBM Security QRadar creates a backup archive of configuration information daily at midnight. The backup archive includes configuration information, data, or both from the previous day. This nightly backup can be customized and an on-demand configuration backup can be created as required. QRadar backups can be stored locally and to a shared storage system. QRadar does not offer an integration with external storage systems like HDFS or cloud storage services like Amazon S3, however some integration efforts have been carried out by the community.

Archiving old data and keeping it stored is crucial for many reasons:

- Data value is perishable, keeping unnecessary data on the live system will slow it down.
- Deleting data limits your ability to derive downstream value for the future. Value can be derived from historical data.
- Different teams want to have access to the same data but for different purposes. Keeping old data allows multiplying the value of that data via different teams deriving new insights from it. The data can be used for things that the original team collecting it never envisioned.
- Data needs to be retained for compliance, but also for forensic analysis.

For most existing SIEM solutions today, once data is archived and is out of the live system, unless it is restored, the SIEM will not use it. Moreover, how archived data is handled or where it is stored or transferred is up to the user and is usually done manually. As there are a diverse number of options for where to store archived data: some SIEM users will opt for attached storage, other will use an in-house Distributed file systems such HDFS, a commercial cloud storage solution like Amazon S3 or Amazon Glacier, etc. Or even use “scp” operations to another device.

Regardless of the archiving solution chosen, the actual archiving process consists of running scripts that are often custom built for a specific IT environment, so a script used by one customer may not be useful for another customer’s needs, and a change in archive option requires rewriting the archiving script.

Finally, archiving retired data from a SIEM can be costly and can pose security and reliability problems if archived data is not handled correctly.

These issues served to inspire us to develop a SIEM extension that can handle data archiving in a reliable, flexible, and secure manner leveraging public Clouds (e.g., Amazon S3, Amazon Glacier, Windows Azure, Blob Store, etc.). The goal is to offer a secure and elastic solution for SIEM data archival regardless of the data retention needs with the ability to customize policies to fit retention requirements.

Previous work done by some members of the consortium on cloud-of-clouds storage [77] can be leveraged to build the SIEM storage extension capable of long term data retention without risking the confidentiality of such data and at a reasonable cost.

5 Role of SIEMs in the future: barriers and enablers

5.1 Introduction

The changing nature of security threats, proliferation of mobile devices, globalization, the explosion of social media and the quick change in regulation are speeding the evolution of Security Information and Event Management. The purpose of this chapter is to analyse those external factors to the project that could potentially affect the future of SIEM systems and the related technology, like the ones developed within DiSIEM. This exercise also intends to consider exploitation aspects in the early stages of the project to guide the project work.

This section is intended to analyse the factors that could affect the evolution of the SIEMs in the mid-term and long-term according to political (including regulation), societal and economic trends.

After analysis of each factor, it is possible to tag them as either enabler or a barrier for the progress of SIEMs. Consequently, future SIEMs should take advantage of those factors identified as enabler, and in the opposite way, coming SIEM evolutions should minimize the effect of the barriers or even avoid them.

5.2 Technological factors

After more than a decade of evolution of Security Information and Event Management, this technology is now becoming a “security big data analytics” platform. Big data will be the base of any effective SIEM product in the near future. According to Gartner report dated on August 2016 [1], it is assumed that by the end of 2017, at least 60% of major SIEM vendors will incorporate advanced analytics and UEBA functionality into their products.

The big amount of data managed by present SIEMs is a challenge not only for data computation but also for data storage and usability. The more information they receive, the more complex it is its later processing, storage and visualization. In the analysis performed to different SIEM solutions in Section 3, it has been identified that although some of them have already taken into consideration measures to increase the processing capabilities in real-time, there are still limited long-term data storage and reporting capabilities. Both factors, storage and visualization, are actually some of the extensions that will be covered by the DiSIEM project (see Section 4.4 for visualization and 4.5.3 for storage), trying to go in the right direction of the Security business needs.

5.2.1 SIEM technology improvements

After the analysis of the current state of the art on SIEMS performed in this deliverable, the following needs are identified:

- **Less intrusive architecture.** SIEMs must be integrated in a company’s infrastructure as easily as possible. Nowadays, the infrastructure of a company may evolve quickly, adapting to new software, more advanced

hardware, etc. Then it is of big importance to have a modular product that can be adapted quickly to customer needs. At the same time, the SIEM must respect the systems legacy existing in any organization infrastructure and be able of dealing with the diverse data sources in the managed infrastructure.

- **Enhanced visualization capabilities.** The big amount of data and information sources managed by current and future SIEMs makes essential to have not only a friendly user interface, adaptable to different roles with different informational needs (security managers, security analysts, etc.), easily configurable, usable, and as simple and intuitive as possible, but also flexible enough to handle real-time data under constant change
- **SIEM connectors.** An adaptive SIEM product must provide the customer with easily adaptable connectors that enable extensibility. That means a company can make use of specific APIs available to easily integrate other components or products in their environment directly by their R&D department (where existing), with light support of the SIEM provider. These connectors and APIs must be properly documented with guidelines and specific examples adapted to customer needs.
- **Scalability.** Scalability of the SIEM solution is necessary at different levels. Organizations are quickly changing and growing the number of connecting devices (e.g: BYOD paradigm, IoT environments), therefore SIEMs must be able to monitor as many devices as possible, in order to ensure full protection of the IT infrastructure. Similarly, not only the hardware infrastructure can grow in an organization but also the number of security events collected at the edge of the SIEM infrastructure, since current companies or organizations are evolving to full-connected environments (digital transformation) and that leads to more sensors, and more devices connected to the same network. Then the challenge is to have a more refined correlation, affecting multiple layers and with more complex rules, together with higher processing and storage capabilities.
- **Secure long-term data storage.** Close related to the previous point, there is a need to support large volumes of data archival in a secure and flexible way. Most of existing solutions are expensive and require external hardware infrastructure customized for each specific organization.
- **Diversity-enhanced risk assessment.** Diversity in data sources feeding a SIEM is necessary to improve its risk analysis capacity. This includes from events generated by widely extended intrusion detection systems (IDSs) or vulnerabilities scanners to more advanced detectors related to User and Entity Behaviour Analysis (UEBA) and support for threat intelligence.

5.2.2 Technology trends

The quick evolution of technology may affect the way SIEMs evolve in the future. Here we perform a brief analysis of technological trends, as reported by Atos [78], highlighting connections with potential changes in future SIEMs. The technological trends identified are the following:

- **Cloud storage.** This technology permits to store bigger volume of data, accessible from anywhere and without hardware limitations because it can be based on a system that ensures scalability in terms that up to now are impossible to achieve in a physical server. This is something that can be

clearly seen as an enabler in SIEM technology since big data analytics of network events can be performed in a more efficient way, without worries about the amount of logs, information, etc. that are stored. In fact, it is one of the technological trends that DiSIEM project will take advantage of.

- **Cloud service integration.** This is treated separately to cloud storage because it is more focused on executing software in a remote server, not only keeping data “statically” in a cloud infrastructure. In the same way as the data storage advantages in cloud, this technology makes possible to ensure scalability and high availability of software applications since they are not restricted to the hardware of a local server, and can be launched from anywhere. This is of course something SIEMs can take advantage of.
- **Mobile technologies.** Mobile devices are growing. It brings new threats that should be analysed by SIEM systems. Now there exist Mobile Device Management (MDM) solutions to secure, monitor and support mobile devices. This kind of devices is e.g. smartphones, tablets, laptops, mobile printers and OTA (On-The-Air) technology.

In this respect, it is a trend that employees use Company owned devices as well as personal devices for office work. A need would be to secure corporate data. Working at home, e.g. with personal computer, what now is commonly called BYOD (Bring Your Own Device), is a trend in Cyber-security. This leads to several potential problems:

- BYOD devices are not managed by the IT team so they are not under the policy control of the company.
- Some of them (e.g., smartphones, tablets) do not have any security solution pre-installed. Their use in a corporate network can be a threat.
- The data in these devices are not encrypted. If those devices are lost or stolen, it can be manipulated.
- Applications installed in those devices cannot be tracked and may be potential treat if they are connected to corporate network.

Looking at this technology trend focused on SIEMs, it can be seen as an enabler because it will be possible to take more data from mobile devices adding information to the data analysis of SIEMs. However, as a drawback (or barrier) these devices are connected to external networks that may be weaker in terms of security and more difficult to monitor by the SIEM.

- **Big data analytics.** As introduced before, SIEMs are evolving to data analytics systems. Data in a connected environment grows exponentially and it makes necessary to have powerful analysis tools capable of real time analysis of events, support to decision making, etc. The growth in data analytics methods is clearly an enabler for SIEM systems.
- **Machine learning technologies.** The new high performance computers, with powerful hardware, and modern programming languages, together with the data analytics explained above, is making possible to create data models, fed by the experience of cause-effect analysis. This analysis permits machines to learn and decide. SIEMs can take advantage of these technologies to make event detection and decision making more intelligent.

- **High performance computing.** Modern high level computational capacity (supercomputers) makes feasible big amount of data analysis in a very short time. This is evidently an enabler for future SIEMs able to better analyse larger amounts of events in less time.
- **Self-adaptive security.** Advanced Security Operations Centres (SOC) and Security Information and Event Management are integral to new security strategies. Organizations have deployed real-time controls in a proactive security approach that is helping to prevent cyberattacks. Trusted probes are being deployed within secure environments to look for evidence of specific suspicious behaviours. These preventive security technologies help agencies anticipate attacks. Behind the scenes, streaming analytics and High Performance Computation (HPC) analyse millions of events in real time. Meanwhile distributed analytics in combination with HPC supports vital forensic analysis after an incident is detected. They help organizations quickly identify where the attack started, who is behind it and how it was carried out.
- **DevOps.** DevOps is a term used to refer to a set of practices that emphasizes the collaboration and communication of both software developers and other IT professionals while automating the process of software delivery and infrastructure changes. Then building, testing and software deployment happen rapidly, frequently and more reliably [79]. Since SIEM is software, which is provided with some capabilities adapted to the customer needs, it will be benefited from this new philosophy of Software development. This will make easier and faster to adapt the software to new customer requests, due to modifications in the infrastructure, new threats identified, etc.
- **Internet of Everything.** The internet of everything (IoE) is a ubiquitous communication network that effectively captures, manages and leverages data from billions of real-life objects and physical activities. It extends the concept of Internet of Things (IoT) by also including people, processes, locations and more. Each object transfer data over a network without human-to-human or human-to-computer interaction. The impact of this technology in SIEMs is that they provide large amount of data and events for analysis. Moreover, security, privacy and trust must be considered in developing IoT solutions.
- **5G Networks.** 5G represents the next generation of communication networks and services, an approach for fulfilling the requirements of future applications and scenarios. This technology will increase the data transfer speed, and then could affect to the amount of data analysed by a SIEM in a network per time unit. This can impose a difficulty for SIEMs in events detection.
- **Social media analytics.** The emerging discipline of social network analysis brings complex patterns of connection, influence and reciprocity that are the main factors in customer behaviour. Meanwhile, social media analysis, with its social media monitoring and sentiment analysis, helps companies derive intelligence from customer conversations on digital media. Social networks like Twitter provide a wealth of information that may be explored by cybersecurity companies as well as by hackers, as attack victims use on-line social media to discuss their experience and knowledge about attacks,

vulnerabilities and exploits. The DiSIEM project will try to take advantage of such analysis to improve existing SIEMs capabilities.

5.3 Societal factors

Society is becoming strongly dependent on information and communications technology (ICT), which is leading to a rapid social, economic, and governmental development. However, this development has brought new threats to critical digital infrastructures. It is then imperative for organizations (countries, companies, etc.) to make suitable investment in cybersecurity so they can fully realize its benefits. It is demonstrated that the number of cybersecurity threats is correlated with the economic and social development of a nation; in general, more developed nations enjoy better cybersecurity [80].

In this section, we briefly introduce how the changes in societal habits related with technology may affect the future of SIEMs.

Firstly, we must understand the human interaction with a SIEM. A SIEM is not only a tool that is deployed over an infrastructure and it works on its own. The human interaction is the key for this system to succeed, from the designer of the tool to the SIEM operator. Then it is important to analyse the human factors that could affect the evolution of such systems.

The following societal factors have been identified and are analysed from a SIEM perspective:

- **Generation Z.** Modern generations have growth with mobiles, internet, etc. They understand the world as a big network in which everything is connected to the internet. Everybody speak through devices connected to the Internet, with all the information immediately available. It can be assumed that people of the future will be more aware of cybersecurity and will bring the companies clearer conscience of the risks of the threats in the network [81]. The citizens from this generation will require more security actions from organizations, generating new opportunities of growth for SIEMS and antiviruses.
- **Growth of social networks.** There is a huge growth of social networks usage among the young generations in the last few years. According to wearesocial.com 2016 annual report, the penetration of active social media users is up to 31% of global population, with an increase of 10% in the last year. This is specifically significant for mobile social users, which has a penetration of 27%, and have growth 17% last year [82]. Then, social networks activity is a source of data that cannot be disregarded, and it can be of very high importance in security events analysis. The DiSIEM project will explore this opportunity.
- **Cyber attacks.** In the new connected societies, the development of the internet has led to a new type of attacks, the cyber-attacks. A cyber-attack is any type of malicious act that targets ICT systems, infrastructures networks, etc. Then attacks to critical infrastructures can be considered the new weapons. As a reference, HACKMAGEDDON reports 1061 cyber-attacks in 2016, 73% being cybercrime, 13% hacktivism, 9% cyber espionage and 5%

cyber warfare [83]. That makes SIEMs to be essential in any infrastructure in which data is of relevance or whose attack may cause operation disruption, even damage to population, not only from a single company interests perspective but also from users, citizens, and (more generally) people perspective.

- **Deep web.** The deep web is the part of the World Wide Web whose contents are not indexed by standard search engines [84]. The deep web can be considered as a barrier by SIEM systems, since it makes difficult to retrieve data from the network.

5.4 Economic factors

Among the economic factors that could affect the future of SIEMs the following can be highlighted:

- **Short term/temporary work.** In 2014 the main type of employment relationship in the EU was full-time permanent contracts, with 59 % of the share of employment, although this is decreasing, while the share of non-standard forms of work is increasing. If this trend continues, it may well become the case that standard contracts will only apply to a minority of workers within the next decade [85]. Due to the new types of work, tending to shorter term jobs, people do not stay in the same company for long time now, especially in the first period of their career. In the past, employees had solid links with the organization for which they work; however, nowadays people tend to change employment more often, reducing the commitment with companies. The consequence is that companies need to minimize the employee's ramp up to learn a new tool, or a new way of working. Therefore, this factor makes essential that future SIEMs have improved and more friendly interfaces at the level of decision taking, configuration rules, link to new sources and sensors. Moreover, the documentation and support to users is of high importance.
- **Freelance.** Self-employment is increasing against the usual company paid employment [86]. Freelance do not work for a company as an employee but as a service provider. This type of work may be a threat for companies because the devices used by freelancers do not belong to the IT department and cannot be easily monitored. On top of that, they do not have strong bonds with the company that hires their services. However, cybersecurity freelance can be a good choice for SIEM providers because they may possess a wider knowledge about potential threats that may affect an organization, since they accumulate a lot of experience from different companies.
- **Cyber security jobs are continuously growing.** The estimated growth in cyber security jobs is of 35% by 2020 [87]. This reflects the importance of cybersecurity for the companies, and that can be an opportunity for SIEMs to grow in the market.
- **Bigger companies, globalization.** The global market makes easier for the big technological companies survive and grow more [88]. However, the level of criticality of that information may be higher. The future SIEMs should be dimensioned for such big companies and global networks.
- **SMEs companies.** The SMEs will become bigger targets of Cyber Attacks in the future [89]. Then they should be the new target for SIEM market growth,

since currently commercial SIEMs are more focused on big infrastructures. SMEs will need less expensive solutions adapted for their specific needs.

- **Reduction of time to market.** Rapid change in the market makes necessary that all products, in our case SIEMs, or services adapt quickly to changes in the market. The future SIEMs should be more adaptive, more modular and more flexible to be in line with the demand of new business.

5.5 Political factors

Citizens depend on a stable, secure, and resilient environment. This environment includes nowadays the “cyberspace”. Protection of individual properties and business or personal sensitive information in the cyberspace is getting critical and the political organizations must take part in this. They must design the security framework, principles and rules to reduce the risks in the population. This risk may economically affect private companies and public institutions. These regulations may affect the evolution of SIEMs in the future, since, in some instance, they analyse sensitive information to detect security events in the network.

The following sections describe some policies and regulations that may affect the SIEMs data collection.

5.5.1 EU regulation in Data protection

In January 2012, the European Commission proposed a comprehensive reform of data protection rules in the EU. On 4 May 2016, the official texts of the Regulation and the Directive REGULATION (EU) 2016/679 have been published in the EU Official Journal [90]. While the Regulation entered into force on 24 May 2016, it shall apply from 25 May 2018. The EU Member States have to transpose the directive it into their national law by 6 May 2018.

The objective of this new set of rules is to give back citizens the control over their personal data, and to simplify the regulatory environment for business. The data protection reform is a key enabler of the Digital Single Market which the Commission has prioritized. The reform will allow European citizens and businesses to fully benefit from the digital economy [91].

There already existed an official definition of “Personal data” in the Article 30 of Directive 95/46/EC and Article 15 of Directive 2002/58/EC. According to Directive 95/46/EC “Data protection directive”, it says that “Personal data shall mean any information relating to an identified or identifiable natural person ...”.

A number of provisions of the Directive contain a substantial degree of flexibility in order to find an appropriate balance between protections of the data subject’s rights on the one side and on the other side the legitimate interests of data controllers [92].

In order to understand how this regulation may affect the data collected by SIEMs, we can see for example how EC understands the propriety of the IP

address in a network (commonly analysed by security software). In the internet, every computer is identified by a single numerical IP address of the form A.B.C.D. where A, B, C and D are numbers in the range of 0 to 255. The working Party has considered IP addresses as data relating to an identifiable person, and issued the following statement: “Internet access providers and managers of local area networks can, using reasonable means, identify Internet users to whom they have attributed IP addresses as they normally systematically “log” in a file the date, time, duration and dynamic IP address given to the Internet user. The same can be said about Internet Service Providers that keep a logbook on the HTTP server. In these cases there is no doubt about the fact that one can talk about personal data in the sense of Article 2 a) of the Directive: “[...] Especially in those cases where the processing of IP addresses is carried out with the purpose of identifying the users of the computer (for instance, by Copyright holders in order to prosecute computer users for violation of intellectual property rights), the controller anticipates that the “means likely reasonably to be used” to identify the persons will be available e.g. through the courts appealed to (otherwise the collection of the information makes no sense), and therefore the information should be considered as personal data. “ [93]

In conclusion, in some cases the dynamic IP addresses are considered personal data, and then should be protected by the EU directives mentioned above. When the person is not identifiable by the IP then it is not considered as personal data, and then could be used for analysis. Consequently, the way SIEMs process and store data must be in line with the directives on data protection.

Moreover, the regulation in data protection affects the SIEMs in the way they can store the data, where is the database located and if it is saved with adequate level of data protection.

5.5.2 Investment R&I on EU Cybersecurity

Recently the EU Commission announced an increment in the investment on cybersecurity (cPPP) in order to put more efforts to reduce cyber-threats in the European Union. This can be summarized in the text of the press release launched in Brussels on fifth of July of 2016: “The Commission today launches a new public-private partnership on cybersecurity that is expected to trigger €1.8 billion of investment by 2020. This is part of a series of new initiatives to better equip Europe against cyber-attacks and to strengthen the competitiveness of its cybersecurity sector.” [94]

The following phrases reveal the clear need of investment on cybersecurity in Europe:

- Andrus Ansip, Vice-President for the Digital Single Market, said: "Without trust and security, there can be no Digital Single Market. Europe has to be ready to tackle cyber-threats that are increasingly sophisticated and do not recognise borders. Today, we are proposing concrete measures to strengthen Europe's resilience against such attacks and secure the capacity needed for building and expanding our digital economy." [94]

- Günther H. Oettinger, Commissioner for the Digital Economy and Society, said: "Europe needs high quality, affordable and interoperable cybersecurity products and services. There is a major opportunity for our cybersecurity industry to compete in a fast-growing global market. We call on Member States and all cybersecurity bodies to strengthen cooperation and pool their knowledge, information and expertise to increase Europe's cyber resilience. The milestone partnership on cybersecurity signed today with the industry is a major step." [94]

This is an initiative of the Commission to establish contractual public private partnership on Cybersecurity (cPPP) between the European Union and the European Cybersecurity Organisation. The adoption and evolution of SIEMs can then be empowered by this investment in cybersecurity. Summary of enablers and barriers

After the study of the technological, social and political aspects that could affect the evolution of SIEMs in the future, the following table gives a summary of which of them can be considered as barriers and which of them are enablers of this technology. In some cases, the limit between barrier and enabler is not so clear because the same factor could be seen as positive and negative depending on the point of view:

	Enablers	Barriers
Technology factors	<ul style="list-style-type: none"> • Cloud storage capacity • High performance computing • Internet of Everything (as data sources) • Big data analytics • Machine learning • Self-adaptive security • Social media analytics 	<ul style="list-style-type: none"> • Cloud storage security mechanisms • BYOD trend • Internet of Everything (increasing number of devices) • 5G Networks
Societal factors	<ul style="list-style-type: none"> • Generation Z • Growth of social networks • Open development environments and communities. • Growth of cyber terrorism 	<ul style="list-style-type: none"> • Deep web
Economic factors	<ul style="list-style-type: none"> • Growing cyber security jobs • Globalization • SMEs 	<ul style="list-style-type: none"> • Freelance commitment to the Company • Short term work • Reduction of time to market
Political factors	<ul style="list-style-type: none"> • Increment of Investment in R&I of European Commission 	<ul style="list-style-type: none"> • EU personal data protection

Table 10: Summary of SIEM enablers and barriers

6 Summary and Conclusions

This document presents a deep analysis of some of the leader SIEM solutions available in the market, namely HP ArcSight, IBM Q Radar and Intel McAfee, together with some more visionary options such as AlienVault's SIEMs or Atos XL-SIEM, and promising tools to be taken into consideration in a SIEM context, such as Elastic Stack and Splunk. The result of this analysis is summarized in the following table where the main strengths and weaknesses identified by each SIEM solutions are enumerated:

	Main Strengths	Main Weaknesses
HP ArcSight	<ul style="list-style-type: none"> • Capacity to support a large scale SOC • Number of available connectors • Resilience 	<ul style="list-style-type: none"> • Complex deployment • Platform lacks flexibility • Visualization capability
IBM Q Radar	<ul style="list-style-type: none"> • Modular architecture • Network traffic and log events correlation • Support for network flows and packets • Vulnerability and asset data for threat intelligence. • Can be deployed using Physical appliances, virtual appliances, IaaS, cloud services • IBM App Exchange to develop applications and extensions to IBM QRadar 	<ul style="list-style-type: none"> • Retiring data • Need third-party technologies for some basic checks • Integration issues for vulnerability management add-on for QRadar • Complex sales process • Data resilience granted for High Availability Deployment • Limited set of charts • Extensive IBM involvement needed for deployment and operation • High price • Simple Behavioural Analytics with UBA App • Limited Alerting Channels
Intel McAfee ESM	<ul style="list-style-type: none"> • Good integration with other McAfee security applications • Wide security solutions portfolio • Customizable visualization framework 	<ul style="list-style-type: none"> • Required additional McAfee security application to support advanced functionality. • High price. • Poor stability, performance and lack of support satisfaction reported by users
Alienvault OSSIM	<ul style="list-style-type: none"> • Open source • Good community support and integration with threat intelligence • Distributed as ISO 	<ul style="list-style-type: none"> • Poor processing capabilities • Data source integration • Simple risk model
Alienvault USM	<ul style="list-style-type: none"> • Low cost • Good community support and integration with threat intelligence • User friendly web interface 	<ul style="list-style-type: none"> • Data source integration • Simple risk model • Limited scalability and flexibility
Atos XL-SIEM	<ul style="list-style-type: none"> • Extended processing capabilities and resilience 	<ul style="list-style-type: none"> • No integration with external threat intelligence (IoCs)

	thanks to Apache Storm cluster and Esper CEP. <ul style="list-style-type: none"> Flexibility in security directives defined by user User friendly web interface 	<ul style="list-style-type: none"> Data storage capabilities Simple risk model
Elastic Stack	<ul style="list-style-type: none"> Open source Good documentation 	<ul style="list-style-type: none"> Needs DIY No correlation capabilities
Splunk	<ul style="list-style-type: none"> Real-time monitoring Log management capabilities for standard log types (e.g. Apache) Advanced security analytics Security monitoring use cases Incident response and management User and application monitoring Deployment and support simplicity 	<ul style="list-style-type: none"> Basic predefined correlation Data volume/day license models Planning and prioritization required not to exceed license Splunk UBA different infrastructure and license

Table 11: Strengths and weaknesses of analysed SIEMs.

To compare the different SIEMs analysed in this deliverable, the following table shows the opinion of partners about how the main features considered are addressed by each solution (☺ good - ☹ fair- ☹ bad), based not only on theory but on their own experience using some of these solutions.

	HP ArcSight	IBM Q Radar	Intel McAfee ESM	Alienvault USM	Atos XL-SIEM	Elastic Stack	Splunk
Data storage	☺	☺	☺	☹	☹	☺	☺
Processing Capabilities	☺	☺	☹	☹	☺	☹	☹
Flexibility directives	☹	☺	☹	☹	☺	☹	☹
Behavioural analysis	☹	☹	☹	☹	☹	☹	☹
Risk Analysis	☹	☺	☹	☹	☹	☹	☺
APIs	☹	☹	☹	☹	☹	☺	☺
Resilience	☺	☹	☹	☹	☺	☹	☺
Visualization	☹	☺	☺	☹	☺	☺	☺
Reaction Capabilities	☹	☹	☹	☹	☹	☹	☺
Deployment	☹	☹	☹	☹	☹	☺	☺
Price	☹	☹	☹	☺	☹	☺	☹

Table 12: Evaluation of analysed SIEMs

Once having made clear the main features provided by these solutions, this report shapes a starting point for the potential innovations to be carried out in work packages WP3, WP4, WP5 and WP6 to cover some needs identified. In particular, DiSIEM will focus on improving the following features:

- **Behavioural analysis.** It is clear from the previous table that all solutions analysed lack of this capability or it is poorly covered. One of the extensions to be developed in DiSIEM will be an application-based anomaly detector for complementing other sensors and detect fraud in application servers.
- **Risk Analysis and deployment.** These features are not either good covered by most of the solutions analysed. Techniques and tools for analysing, evaluating and guiding the optimal deployment of diverse security mechanisms in the managed infrastructure, including multi-level risk-based metrics, will be developed and it will be provided a framework for deploying diverse and redundant sensors.
Besides, in order to improve the threat intelligence capacity of SIEMs, it will be used OSINT for feeding a SIEM with information about security threats and about the identification of trends over time. DigitalMR's listening247 platform will be used as a base for this OSINT-based security threat predictor extension.
- **Visualization and reaction capabilities.** Although most of the solutions analysed provide user-friendly graphical interfaces, they have limited capabilities to deal with huge number of events collected. In DiSIEM it will be developed web-based visualisation and analysis extensions, which help the user to have a high-level insight of the situation and a more efficient decision making and reaction.
- **Data storage and price.** Although most of the solutions analysed include good data storage capabilities, they are limited by the hardware availability and usually it is required additional products (and event licenses based on data volume) with the consequent increase in the price. Secure and elastic solution based on cloud-of-clouds storage for long-term SIEM data archival in diverse public clouds (e.g., Amazon S3, Amazon Glacier, Windows Azure, Blob Store, etc.) will be developed in DiSIEM with the ability to customize policies to fit data retention needs.

Some other features considered in this deliverable such as resilience, processing capabilities or flexibility in security directives, will be not covered in the extensions to be developed in DiSIEM (at least initially) since they are out of the scope of the project.

It is also important to remark that although the proposed enhancements will be developed trying to be SIEM-independent so the outcomes can be reused by multiple SIEM solutions, they will be only integrated and validated through pilot deployments in the SIEMs provided by the partners involved in DiSIEM. In particular: HP ArcSight deployed in EDP environment, Splunk and Elastic Stack in AMADEUS environment and XL-SIEM in Atos environment. The specific requirements and limitations to DiSIEM results to enable their validation in these environments will be documented in WP7.

Finally, it has been also studied the role of the SIEMs in the near and long-term future taking into account different political, economic, technological and social factors which can act as enablers or barriers. From this analysis we can conclude that conditions are good to foster investment in improving and extending this technology as a key component not only for Security Operation Centres but also to provide cyber security management for SMEs with reduced security knowledge and capacities.

7 List of Acronyms

Acronym	Description
ACE	Advanced Correlation Engine
ADM	Application Data Monitor
AMQP	Advanced Message Queuing Protocol
AQL	Ariel Query Language
AWS	Amazon Web Services
BSD	Berkeley Software Distribution
BYOD	Bring Your Own Device
CEP	Complex Event Processing
CEF	Common Event Format
cPPP	contractual Public Private Partnership
CRE	Custom Rules Engine
CSV	Comma-separated values
CVSS	Common Vulnerability Scoring System
DAS	Direct Attached Storage
DDS	Data Distribution Service
DEM	Database Event Monitor
DPI	Deep packet inspection
DSS	Decision Support System
DRPC	Distributed Remote Procedure Call
DSM	Device Support Module
DMZ	Demilitarized zone
DRBD	Distributed Replicated Block Device
ELM	Enterprise Log Manager
ELK	Elastic
EPL	Event Processing Language
EPS	Events Per Second
ePo	ePolicy Orchestrator
ERC	Event Receiver
ESM	Enterprise Security Manager
FTI	Full-text index
GWT	Google Web Toolkit
GPL	General Public License
GTI	Global Threat Intelligence
HA	High Availability
HDFS	Hadoop Distributed File System
HIDS	Host intrusion detection systems
HPC	High performance Computing
ICT	Information and Communications Technology
IDS	Intrusion Detection Systems
IOC	Indicators Of Compromise
IoT	Internet of Things
IPFIX	Internet Protocol Flow Information Export
IPS	Intrusion Prevention Systems

IRC	Internet Relay Chat
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
MDM	Mobile Device Management
MEF	McAfee Event Format
MTTR	Mean Time To Remediate
MQ	Magic Quadrant
NAS	Network-attached storage
NBAD	Network Behavioural Anomaly Detection
NIDS	Network Intrusion Detection System
NRT	Near Realtime
OpenDXL	Open Data Exchange Layer
OSINT	Open-source intelligence
OSSIM	Open Source SIEM
OTA	On-The-Airs
OTX	Open Threat Exchange
PCAP	Packet capture
RAID	Redundant Array of Independent Disks
REC	Event Receiver
SAN	Storage Area Network
SCADA	Supervisory control and data acquisition
SDEE	Security Device Event Exchange
SDK	Software Development Kit
SEF	Standard Event Format
SNMP	Simple Network Management Protocol
SPL	Search Process Language
SQL	Structured Query Language
SOAP	Simple Object Access Protocol
SOC	Security Operations Center
SSD	Solid-State Drive
STIX	Structured Threat Information eXpression
TLS	Transport Layer Security
UBA	User Behaviour Analysis
UEBA	User and Entity Behaviour Analysis
USM	Unified Security Management
VA	Vulnerability Assessment
WMI	Windows Management Instrumentation
WMICL	Windows Management Instrumentation Command Line
WSDL	Web Services Description Language

8 References

- [1] Gartner, “Magic Quadrant for Security Information and Event Management,” 10 August 2016.
- [2] O. Rochford, K. M. Kavanagh and T. Bussa, “Critical Capabilities for Security Information and Event Management,” 2016.
- [3] K. Scarfone, «Comparing the best SIEM systems on the market,» September 2015. [Online]. Available: <http://searchsecurity.techtarget.com/feature/Comparing-the-best-SIEM-systems-on-the-market>.
- [4] I. Nirvana, «SIEM Product Comparison – 2016,» 2016. [Online]. Available: <http://infosecnirvana.com/siem-product-comparison-201/>.
- [5] [Online]. Available: ftp://public.dhe.ibm.com/software/security/products/qradar/documents/71MR1/SIEM/CoreDocs/QRadar_71MR1_HighAvailabilityGuide.pdf.
- [6] [Online]. Available: <http://www-03.ibm.com/software/products/en/x-force-threat-intelligence>.
- [7] [Online]. Available: <http://www-03.ibm.com/software/products/en/qradar>.
- [8] [Online]. Available: <http://www-03.ibm.com/software/products/en/qradar-risk-manager>.
- [9] [Online]. Available: http://www.ibm.com/support/knowledgecenter/en/SS42VS_7.2.7/com.ibm.qradar.doc/c_hwg_app_oview.html.
- [10] [Online]. Available: <http://www-03.ibm.com/security/engage/app-exchange/>.
- [11] «Alienvault Official Site,» [Online]. Available: <https://www.alienvault.com>.
- [12] Alienvault, “Life Cycle of a log,” 2014. [Online]. Available: https://www.alienvault.com/doc-repo/usm/security-intelligence/AlienVault_Life_cycle_of_a_log.pdf.
- [13] Alienvault, «Unified Security Management (USM) Deployment guide,» 23 January 2017. [Online]. Available: <https://www.alienvault.com/documentation/resources/pdf/usm-deployment-guide.pdf>.
- [14] Alienvault, «AlienVault USM for AWS solution guide,» 2015. [Online]. Available: http://www.infosecurityeurope.com/_novadocuments/85377?v=635660222800170000.
- [15] D. Hermanowski, “Open Source Security Information Management System Supporting IT Security Audit,” 2015. [Online]. Available: http://www.wil.waw.pl/art_prac/2015/pub_cybconfCybersec15_DH-OSSIM-ieee_REVIEW_RC05_ver_PID3720933.pdf.
- [16] AlienVault, «Open Threat Exchange (OTX),» [Online]. Available: <https://www.alienvault.com/open-threat-exchange>.

- [17] “FIWARE,” [Online]. Available: <https://www.fiware.org/>.
- [18] «ACDC Project,» [Online]. Available: <https://www.acdc-project.eu/>.
- [19] «FI-XIFI Project,» [Online]. Available: <https://www.fi-xifi.eu/home.html>.
- [20] «RERUM Project,» [Online]. Available: <https://ict-rerum.eu/>.
- [21] «WISER Project,» [Online]. Available: <https://www.cyberwiser.eu/>.
- [22] EsperTech, “Esper: Event processing for Java,” [Online]. Available: <http://www.espertech.com/products/esper.php>.
- [23] “Apache Storm,” [Online]. Available: <http://storm.apache.org/>.
- [24] “Structured Threat Information eXpression (STIX™),” [Online]. Available: <https://stixproject.github.io/>.
- [25] “RabbitMQ,” [Online]. Available: <https://www.rabbitmq.com/>.
- [26] “Apache Zookeeper,” [Online]. Available: <https://zookeeper.apache.org/>.
- [27] “ZeroMQ,” [Online]. Available: <http://zeromq.org/>.
- [28] Espertech, “Espertech performance results,” [Online]. Available: <http://www.espertech.com/esper/release-5.1.0/esper-reference/html/performance.html#performance-results>.
- [29] A. Mathew, “Benchmarking of Complex Event Processing Engine Esper,” 2014.
- [30] «Overview of the Event Processing Language (EPL),» [Online]. Available: https://docs.oracle.com/cd/E13157_01/wlevs/docs30/epl_guide/overview.html.
- [31] “Data Distribution Service,” [Online]. Available: <http://portals.omg.org/dds/>.
- [32] “Distributed RPC,” [Online]. Available: <http://storm.apache.org/releases/1.0.0/Distributed-RPC.html>.
- [33] “Daemontools,” [Online]. Available: <http://cr.yip.to/daemontools.html>.
- [34] “Supervisord,” [Online]. Available: <http://supervisord.org/>.
- [35] “Storm Apache: Fault tolerance,” [Online]. Available: <http://storm.apache.org/releases/current/Fault-tolerance.html>.
- [36] [Online]. Available: elastic.co/guide/index.html.
- [37] [Online]. Available: <https://www.credera.com/blog/technology-insights/java/can-querydsl-part-1-enhance-simplify-existing-spring-data-jpa-repositories/>.
- [38] [Online]. Available: https://lucene.apache.org/core/2_9_4/queryparsersyntax.html.
- [39] [Online]. Available: <https://www.elastic.co/guide/en/x-pack/current/xpack-introduction.html>.
- [40] [Online]. Available: <https://github.com/Yelp/elastalert>.
- [41] [Online]. Available: <https://www.elastic.co/subscriptions>.
- [42] A. Jaquit, “Security metrics,” Pearson Education, 2007.
- [43] The Center for Internet Security, “The CIS Security Metrics,” from <https://benchmarks.cisecurity.org/downloads/show-single/index.cfm?file=metrics.110>, 2016.

- [44] E. Yasasin and G. Schryen, "Requirements for IT Security Metrics - an Argumentation Theory Based Approach," In ECIS, 2015.
- [45] P. De Lutiis, "Information Security Indicators (ISI); Key Performance Security Indicators (KPSI) to evaluate the maturity of security event detection, GS ISI 003 V1.1.2," from http://www.etsi.org/deliver/etsi_gs/isi/001_099/003/01.01.02_60/gs_isi_003v010102p.pdf, 2014.
- [46] J. M. Torres, J. M. Sarriegi, J. Hernantes and A. Lauge, "Steering Security through Measurement," In TrustBus 2009, 5695, p. 95., 2009.
- [47] SANS, "Benchmarking Security Information Event Management (SIEM)," from http://www.sans.org/reading_room/analysts_program/eventMgt_Feb09.pdf, 2016.
- [48] J. Homer, S. Zhang, X. Ou, D. Schmidt, Y. Du, S. R. Rajagopalan and A. Singhal, "Aggregating vulnerability metrics in enterprise networks using attack graphs," *Journal of Computer Security*, 21(4), 561-597, 2013.
- [49] Alienvault, «Alienvault Security Intelligence,» [Online]. Available: <https://www.alienvault.com/solutions/security-intelligence>.
- [50] IBM, "IBM Security QRadar Risk Manager," from <http://www-03.ibm.com/software/products/en/qradar-risk-manager>, 2016.
- [51] IBM, "IBM Security QRadar Vulnerability Manager," from <http://www-03.ibm.com/software/products/en/qradar-risk-manager>, 2016.
- [52] IBM, "IBM Security QRadar SIEM," from <http://www-03.ibm.com/software/products/en/qradar-siem>, 2016.
- [53] IBM, "Asset risk levels and vulnerability categories, IBM QRadar Security Intelligence Platform 7.2.8," from http://www.ibm.com/support/knowledgecenter/en/SS42VS_7.2.8/com.ibm.qradar.doc/c_qvm_view_scan_rslthosts_.html, 2016.
- [54] First, "Common Vulnerability Scoring System, V3 Development Update,," from <https://www.first.org/cvss>, December 12, 2016.
- [55] HP, "HPE Security ArcSight ESM," from <http://www8.hp.com/us/en/software-solutions/siem-security-information-event-management/>, December 15, 2016.
- [56] F. Thiele, "ArcSight priority formula," from <http://h41382.www4.hp.com/gfs-shared/downloads-340.pdf>, 2014.
- [57] I. Kotenko, O. Polubelova, I. Saenko and E. Doynikova, "The ontology of metrics for security evaluation and decision support in SIEM systems," *Proceedings - 2013 International Conference on Availability, Reliability and Security, ARES 2013*, 638-645. <https://doi.org/10.1109/ARES.2013.84>, 2013.
- [58] V. Gupta, V. Lam, H. V. Ramasamy, W. H. Sanders and S. Singh, "Dependability and performance evaluation of intrusion-tolerant server architectures," In *Proc. of the LADC 2003*, LNCS 2847 (pp. 81-101), 2003.
- [59] S. Singh, M. Cukier and W. H. Sanders, "Probabilistic validation of an intrusion-tolerant replication system," In *Proc. of the IEEE/IFIP DSN 2003*. (pp. 615-624), 2003.

- [60] L. Wang, Z. Li, S. Ren and K. Kwiat, "Optimal Voting Strategy against Random and Targeted Attacks," In International Journal of Secure Software Engineering (IJSSE), 4(4), 25-46, 2013.
- [61] B. Littlewood and D. Wright, "The use of multilegged arguments to increase confidence in safety claims for software-based systems: A study based on a BBN analysis of an idealized example," In IEEE Transactions on Software Engineering, 33(5), 347-365, 2007.
- [62] R. D. Kulkarni, "Using Ensemble Methods for Improving Classification of the KDD CUP '99 Data Set," IOSR Journal of Computer Engineering, 16(5), 57-61. <http://doi.org/10.9790/0661-16535761>, 2014.
- [63] J. W. Ulvila and J. E. Gaffney, "Evaluation of Intrusion Detection Systems," Journal of Research of the National Institute of Standards and Technology, 108(6), 453-473. Journal Article. <http://doi.org/10.6028/jres.108.040>, 2003.
- [64] B. Littlewood and L. Strigini, "Redundancy and diversity in security," In ESORICS 2004 (pp. 423-438), 2004.
- [65] M. Garcia, A. Bessani, I. Gashi, N. Neves and R. Obelheiro, "Analysis of operating system diversity for intrusion tolerance," Software: Practice and Experience, 44(6), 735-770, 2014.
- [66] J. Reynolds, J. Just, E. Lawson, L. Clough, R. Maglich and K. Levitt, "The design and implementation of an intrusion tolerant system," In Proc. of the IEEE/IFIP DSN 2002. (pp. 285-290), 2002.
- [67] J. Oberheide, E. Cooke and F. Jahanian, "CloudAV: N-Version Antivirus in the Network Cloud," In USENIX Security Symposium (pp. 91-106), 2008.
- [68] N. G. Boggs and S. Stolfo, "Aldr: A new metric for measuring effective layering of defenses," In Fifth Layered Assurance Workshop (LAW 2011), 2011.
- [69] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath and P. Shakarian, "Darknet and Deepnet Mining for Proactive Cybersecurity Threat Intelligence, 1-6," from <http://arxiv.org/abs/1607.08583>, 2016.
- [70] D. Kergl, "Enhancing Network Security by Software Vulnerability Detection Using Social Media Analysis Extended Abstract," IEEE International Conference on Data Mining Workshop (ICDMW), 1532-1533. <https://doi.org/10.1109/ICDMW.2015.228>, 2015.
- [71] F. Jenhani, M. S. Gouider and L. B. Said, "A Hybrid Approach for Drug Abuse Events Extraction from Twitter," Procedia Computer Science, 96(September), 1032-1040. <https://doi.org/10.1016/j.procs.2016.08.121>, 2016.
- [72] L. Marinos, "ENISA threat taxonomy: A tool for structuring threat information. Initial report," from <https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends/enisa-threat-landscape/etl2015/enisa-threat-taxonomy-a-tool-for-structuring-threat-information>, 2016.
- [73] J. J. Cebula and L. R. Young, "A Taxonomy of Operational Cyber Security Risks," Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst, (December), 1-47, 2010.

- [74] C. D. Manning, “The Stanford CoreNLP Natural Language Processing Toolkit,” Baltimore, Maryland USA, Association for Computational Linguistics., 2014.
- [75] Esri, “The Geospatial Approach to Cyber security: An Executive Overview [White paper],” from <https://www.esri.com/~media/Files/Pdfs/library/whitepapers/pdfs/geospatial-approach-cybersecurity.pdf>, 2014.
- [76] A. Slingsby, J. Dykes and J. Wood, “Exploring uncertainty in geodemographics with interactive graphics,” In IEEE Transactions on Visualization and Computer Graphics, 17(12), 2545-2554, 2011.
- [77] «<https://supercloud-project.eu/>,» [Online].
- [78] Atos, “Ascent Look Out 2016+,” *Ascent*, 2016.
- [79] D. Samovskiy, «The Rise of DevOps,» 2010.
- [80] P. N. K. S. T. S. (. T. C. David Burt, «the Cybersecurity Risk Paradox,» Microsoft, 2014.
- [81] H.-y. S. T. S. R. C. N. J. R. R. L. & S. A. Mengtian Jiang, «Generational differences in online safety perceptions, knowledge, and practices,» Taylor & Francis Online, 2016.
- [82] S. Kemp, «DIGITAL in 2016,» We are social.
- [83] Hackmaggeddon, «2016 Cyber Attacks Statistics,» 2016.
- [84] N. Hamilton, «The Mechanics of a Deep Net Metasearch Engine».
- [85] European Parliament, «Precarious Employment in Europe. Part 1: Patterns, Trends and policy Strategy,» European Parliament, 2016.
- [86] E. F. o. I. P. EFIP, «Future Working: The Rise Of European’s Independent Professionals,» 2016.
- [87] Forbes, «One Million Cybersecurity Job Openings In 2016,» Forbes, 2016.
- [88] economist.com, «Why giants thrive,» 2016.
- [89] eqnetworks, «SMEs Will Become Even Bigger Targets of Cyber Attacks in 2016: 3 Options They Can Pursue,» eqnetworks.com, 2016.
- [90] E. Parliament, *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*, European Commission, 2016.
- [91] E. Commission, «EU Data Protection Reform, What benefits for businesses in Europe?,» European Commission, 2016.
- [92] Data Protection Working Party, *Privacy on the Internet - An integrated EU Approach to On-line Data Protection -*, European Commission, 2000.
- [93] Data Protection Working Party, *Opinion 4/2007 on the concept of personal data*, European Commission, 2007.
- [94] European Commission, «Commission signs agreement with industry on cybersecurity and steps up efforts to tackle cyber-threats,» European Commission, 2016.
- [95] K. M. Kavanagh, O. Rochford and T. Bussa, “Magic Quadrant for Security Information and Event Management. Gartner Group Report,,” 2016.
- [96] «McAfee Advanced Correlation Engine,» [Online]. Available: <http://www.mcafee.com/us/resources/data-sheets/ds-advanced->

correlation-engine.pdf.

- [97] H.-y. S. T. S. R. C. N. J. R. R. L. ȳ. S. A. Mengtian Jiang, «Generational differences in online safety perceptions, knowledge, and practices,» Taylor & Francis Online, 2016.
- [98] economist.com, «Why giants thrive,» 2016.

9 Appendix I: Elasticsearch Products for Log Ingestion

Logstash

Logstash is an open source data collection engine with real-time pipelining capabilities.

Although Logstash was originally designed for log collection and parsing, its capabilities extended beyond this use case to include data enrichment and transformation features: two powerful features that are heavily used.

Data enrichment consists of adding additional information from field values. This can be done using external resources (external dictionaries with translate, geoip lookup, etc.) or resources that are shipped with elasticsearch (urldecode, useragent parsing, key values, fingerprinting, etc.)

Logstash offers a wide range of plugins for data ingestion, data parsing, enrichment and data output.

Plugins can be categorized into three types; Input plugins (data source), Filter plugins (data parsing and enrichment) and Output plugins (data destination). Table below summarizes a sample of plugins.

Input Plugins	Filter Plugins	Output Plugins
File (logs)	Aggregate	Csv
http	Anonymize	Elasticsearch
Jdbc	Csv	Email
Kafka	Fingerprint	File
Log4j	Geoip	Rabbitmq
Rabbitmq	Grok (parsing)	Redis
Redis	Json	S3
S3	Kv	Tcp
Stdin	Ruby	
Syslog	Useragent	
Tcp	Urlcode	
Twitter	Xml	

Table 13 : Input, Filter and Output plugins in Elastic Stack

The plugins shipped with Logstash are enough to cover most use cases and for the rare occasions where the Logstash plugins fall short, Logstash offers the capability of adding custom plugins.

Beats

The Beats are open source data shippers that are installed as agents on servers to send different types of operational data to Elasticsearch. Beats can send data directly to Elasticsearch or send it to Elasticsearch via Logstash, which you can be used to parse and transform the data.

For Beats modules are developed by the Elastic Company: Packetbeat, Filebeat, Metricbeat, and Winlogbeat.

Filebeat

Filebeat is a log data shipper. Installed as an agent on servers, Filebeat monitors the log directories or specific log files, tails the files, and forwards events either to Elasticsearch or Logstash for indexing.

When Filebeat is started, it starts one or more prospectors that look in the paths specified for log files. For each log file that the prospector locates, Filebeat starts a harvester that reads a single log file.

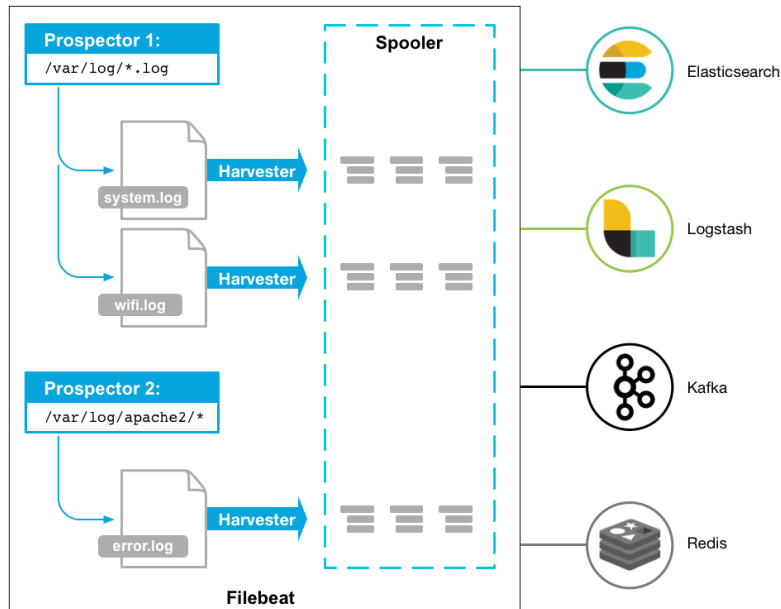


Figure 9-1 : Elasticsearch Filebeat diagram

Winlogbeat

Winlogbeat ships Windows event logs to Elasticsearch or Logstash. It can be installed as a Windows service on Windows XP or later.

Winlogbeat reads from one or more event logs using Windows APIs since the file format is proprietary, filters the events based on user-configured criteria, then sends the event data to the configured outputs (Elasticsearch or Logstash).

Winlogbeat can capture event data from any event logs running on your windows system. For example, you can capture events such as: application events, hardware events, security events, system events.

Metricbeat

Metricbeat is a lightweight shipper that you can be installed on servers to periodically collect metrics from the operating system and from services running on the servers. Metricbeat takes the metrics and statistics that it collects and ships them to the output specified. Possible outputs are: Elasticsearch, Logstash, Kafka, Redis, File and Console.

Metricbeat helps you monitor your servers by collecting metrics from the system and the following services: Apache, HAProxy, MongoDB, MySQL, Nginx, PostgreSQL, Redis, System Zookeeper.

Metricbeat consists of modules and metricsets. A Metricbeat module defines the basic logic for collecting data from a specific service, such as Redis, MySQL, and

so on. The module specifies details about the service, including how to connect, how often to collect metrics, and which metrics to collect.

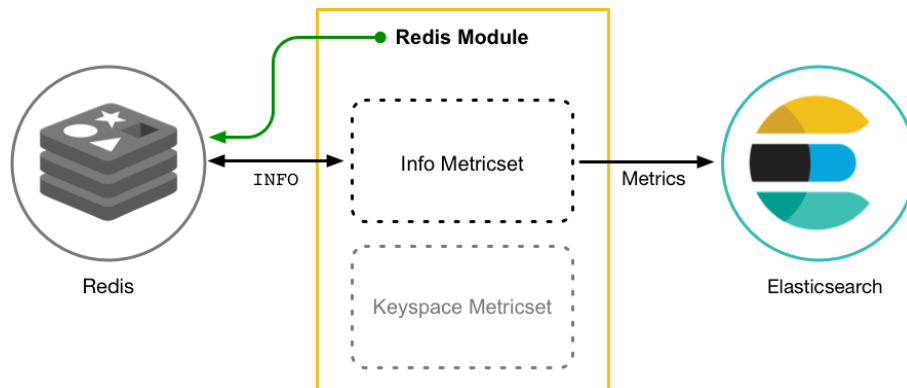


Figure 9-2 : Elasticsearch Metricbeat diagram

Packetbeat

Packetbeat is a real-time network packet analyzer that you can use with Elasticsearch to provide an application monitoring and performance analytics system. Packetbeat completes the Beats platform by providing visibility between the servers of your network.

Packetbeat works by capturing the network traffic between your application servers, decoding the application layer protocols (HTTP, MySQL, Redis, and so on), correlating the requests with the responses, and recording the interesting fields for each transaction.

Packetbeat sniffs the traffic between your servers, parses the application-level protocols on the fly, and correlates the messages into transactions.

Currently, Packetbeat supports many protocols, as: ICMP, DNS, HTTP, Cassandra, MySQL, PostgreSQL, Redis, etc.

Packetbeat can run on the same servers as your application processes or on its own servers. When running on dedicated servers, Packetbeat can get the traffic from the switch's mirror ports or from tapping devices.

After decoding the Layer 7 messages, Packetbeat correlates the requests with the responses in what we call transactions. For each transaction, Packetbeat inserts a JSON document into Elasticsearch.

PacketBeat supports adding new custom Protocols.

Customized Beat

If you have a specific use case to solve, you can create your own Beat. The libbeat library, written entirely in Golang, offers the API that all Beats use to ship data to Elasticsearch, configure the input options, implement logging, and more.

Custom Data Ingestion

StreamSets Data Collector: StreamSets Data Collector is a lightweight, powerful engine that streams data in real time. Use Data Collector to route and process

data in your data streams. Large number of data origins and destinations out of the box, including of course Elasticsearch as a data destination.

Hadoop Connector: Elasticsearch for Apache Hadoop (ES-Hadoop) is the two-way connector that solves a top wishlist item for any Hadoop user out there: real-time search. While the Hadoop ecosystem offers a multitude of analytics capabilities, it falls short with fast search. ES-Hadoop bridges that gap, letting you leverage the best of both worlds: Hadoop's big data analytics and the real-time search of Elasticsearch.

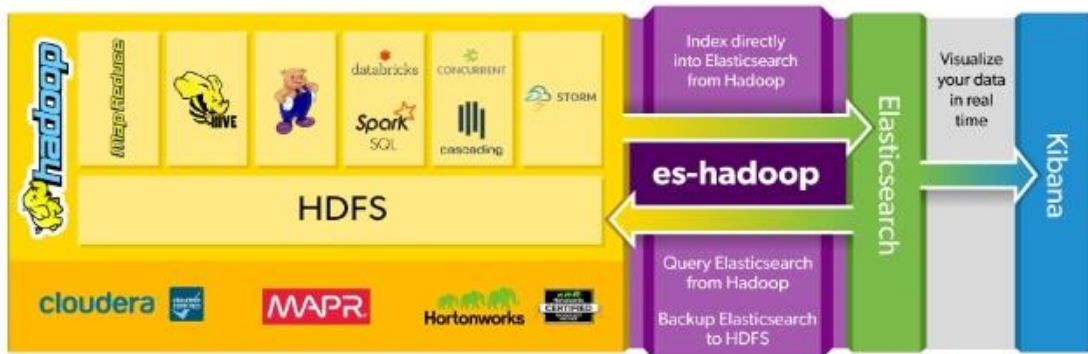


Figure 9-3 : Hadoop Elasticsearch ecosystem

ES-Hadoop connector allows to seamlessly Move Data between Elasticsearch and Hadoop: with a native integration and a rich query API, ES-Hadoop lets you index data directly into Elasticsearch from Hadoop, query Elasticsearch from Hadoop, and use HDFS as a long-term archive for Elasticsearch.

Combining Elasticsearch and Hadoop components allows performing sub-second search queries and analytics on data: While Hadoop lets you batch, join, and analyze to your heart's content, its queries are not the quickest. ES-Hadoop works with any flavor of Hadoop Distributions.

Index and Bulk API: Apart from the Elastic products used for data ingestion, the Elasticsearch's REST API offers the possibility to index data both in small chunks and in a bulk fashion. Clients in the following programming languages are maintained by Elastic: Java, Java REST client, Javascript, Groovy, .NET, PHP, Perl, Python, Ruby. Besides the officially supported Elasticsearch clients, there are several clients that have been contributed by the community for various languages, like Go, Haskell, Kotlin, Lua, Scala and others.

10 Appendix II: Elasticsearch deployment

Deployment of ELK

The minimal ELK set-up can be seen in Figure 10-1:

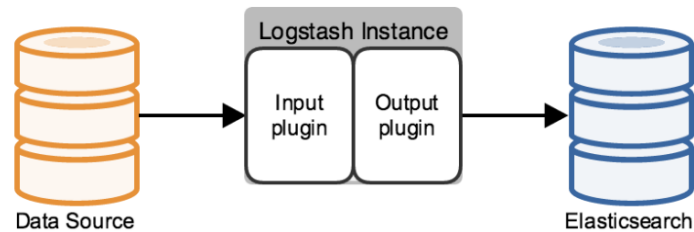


Figure 10-1: ELK minimal setup

But if a user would like to take full advantage of ELK capabilities, he/she should choose the following arrangement that resembles deployment where the user is making use 4 tiers (see Figure above):

- The input tier (i.e. that consumed data from source and consists of Logstash instances having adequate input plugins).
- The message broker (i.e. entity that serves as a buffer which holds ingested data but also works as a failover protection).
- The filter tier (i.e. of course applies data parsing).
- The indexing tier (i.e. the one that moves the processed data to Elasticsearch).

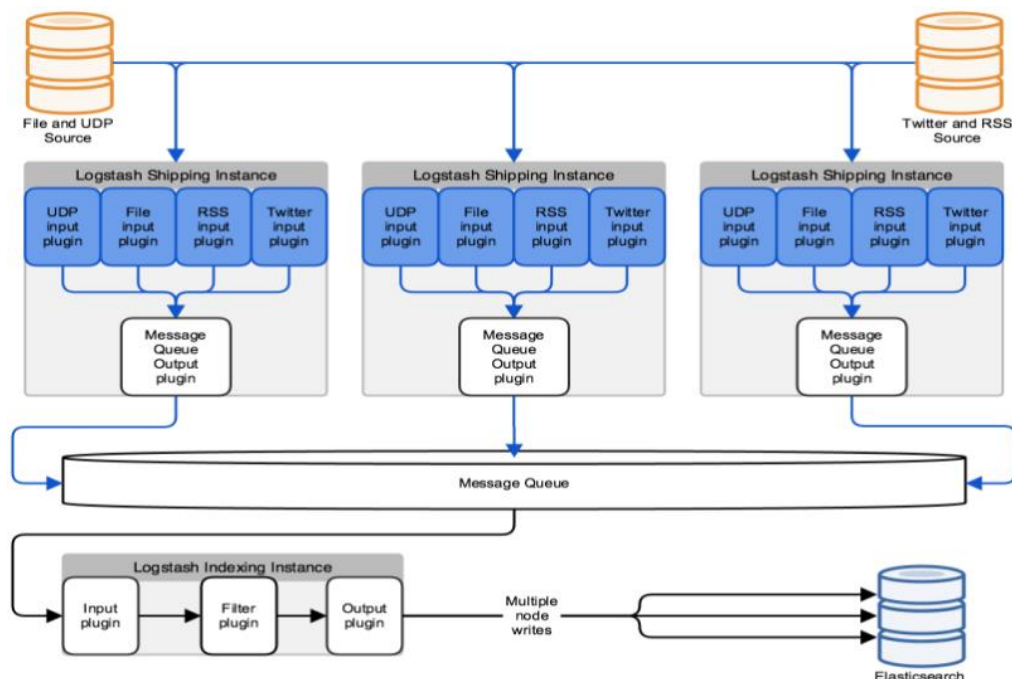


Figure 10-2 : ELK mature setup deployment

Elastic Cloud

Elastic Cloud is a hosted Elasticsearch solution that gives all capacity of the Elasticsearch and Kibana, reducing time to deployment without the worry of hosting and maintaining the infrastructure. We can deploy and manage a secure cluster running on Amazon with the latest versions of Elasticsearch and Kibana.

Advantages of Elastic Cloud:

- Scaling and upgrading are easy, being able to change the structure according the necessity.
- Provide free instance of Kibana, and backup of the cluster every 30 minutes.
- Network and hardware infrastructure monitored 24/7 by the Elastic team.
- Elastic Cloud supports most of the security features that are part of the Elastic Stack. These features are designed to:
 - Prevent unauthorized access with password protection, and role-based access control.
 - Preserve the integrity of your data with message authentication and SSL/TLS encryption.