SIEM Technologies Practice Report (2025)

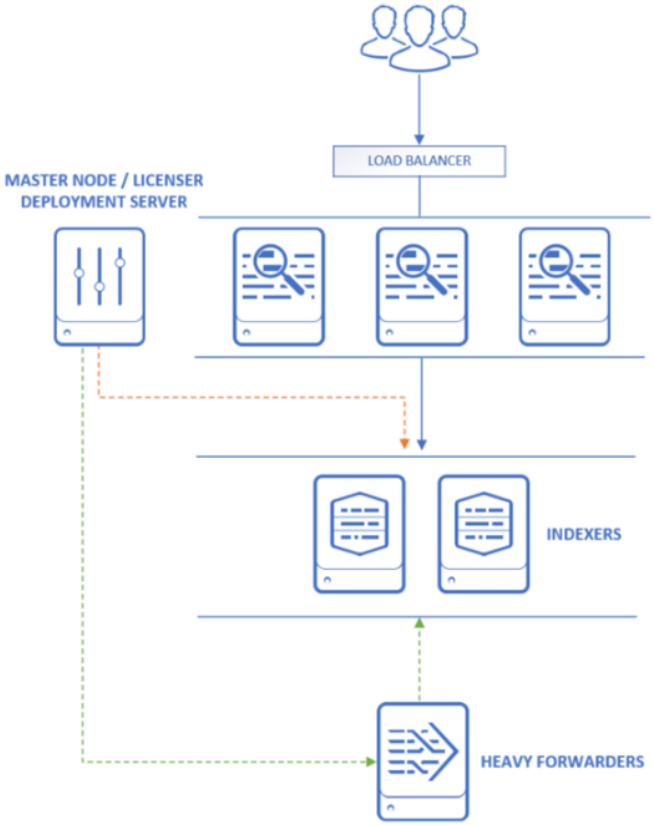
Author: Pedro Oller Serrano

20/02/2025

Statement of practice

You join the Security Department of a major company in the country's service sector as a SIEM expert. On the first day, your boss, the security manager, gives you a brief introduction about the security event monitoring system that they have deployed in the company in order to start giving you visibility into this solution.

The first thing it tells you is that it is an *on-premise* solution based on *Splunk*. The deployed architecture looks like this:



1. Given that he does not directly manage this project, and in view of some internal audit requirements that have been set for him, he asks you to **specify**, on the ***current architecture***, ***where high availability is available*** and ***where not***, in which case you will have ***to propose a plausible solution for this purpose.***

2. In addition, and in order to reduce ***costs***, it asks you to carry out a **study** on the feasibility of ***modifying the current architecture using other possible alternatives and to consider them, indicating their characteristics.***

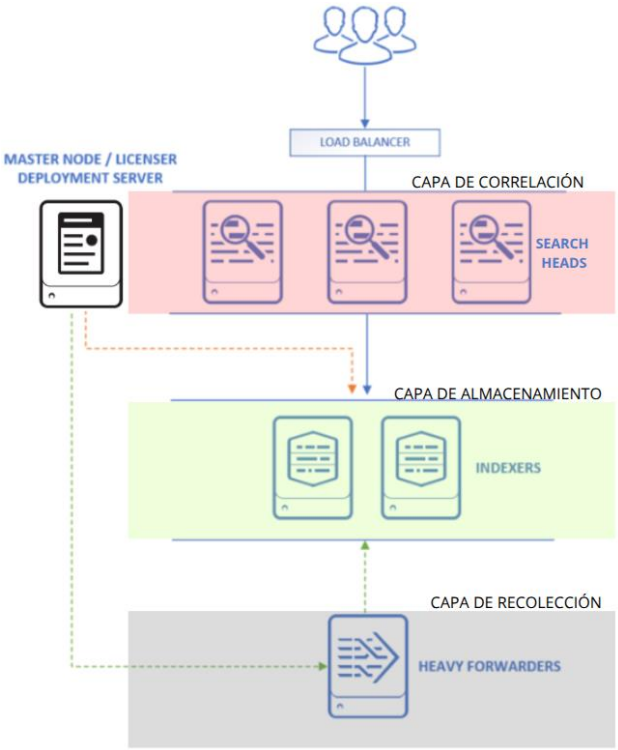
3. In order to make this possible change in architecture, it is necessary to have certain data to look for solutions within the market that fit this scenario. That is why, knowing that ***an event is 560 bytes*** and that it is necessary **to store them** for a ***minimum of two years*** due to regulatory compliance, with an **average ingestion consumption of 3000 EPS** (events per second), it is necessary to know the ***disk size*** that will be required to store that amount of data, assuming a ***compression ratio of 10:1***.

4. On the other hand, it is necessary ***to know how other manufacturers deal with this type of implementation*** and to know their weaknesses and strengths in order to have a clear vision of their functionalities (limit the response to ***only two manufacturers***).

Resolution of the practice

Specifications on the architecture and implementation of HA (High Availability)

In the following figure is represented the schematic of the implemented architecture, I imagine that the "Master node/ Licenser Deployment Server" refers to the license server, so I have taken the liberty of modifying the symbol by which Splunk uses, for greater clarity of the exercise:



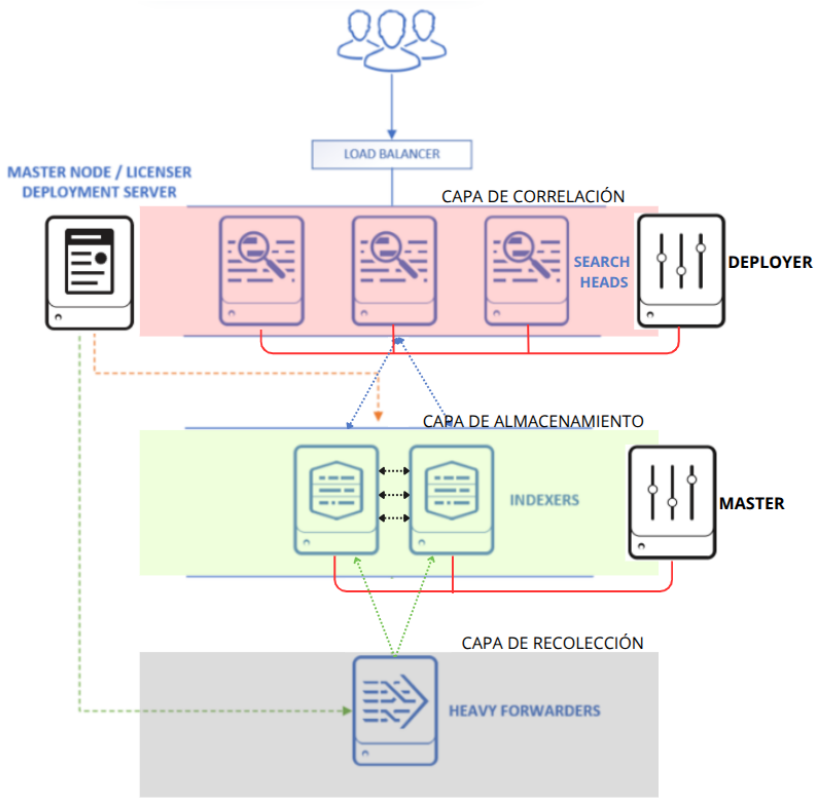
First, a description of the architecture by layers will be carried out, and then HA will be applied, both to the storage layer and to the correlation layer:

- Collection layer: In this layer we find a *Heavy Forwarders,* which is a type of *Forwarder* with advanced processing functionalities, whose function is to receive, process and send the data to the two *indexers* of the storage layer. Some of its main features are:
  - Apply *parsing, partial indexing,* and filtering before forwarding the data.
  - Transform and discard unwanted data, reducing the load on indexers .
  - It collects data from sources such as APIs, databases, and files.
  - Secures transmission via SSL
  - Provides redundancy if you are in HA
- Storage layer: in this layer we have two *indexers* without *a cluster*, that is, they are not in HA. So they can be configured for a distribution of the data sent by the

Heavy Forwarders or to configure a data separation or for a simple backup in case one goes down that the other continues, but without automatic replication since it does not have a *cluster* so the data of the *indexer* that has fallen will be lost.

- Correlation Layer: in this layer we can see a total of three *Search Heads* without *a cluster*, so there is no replication or HA. They could be configured for load distribution or as a manual backup, although without a *cluster* there is *no failover* so if one fails the others would have to be configured and the scheduled searches would be lost. Panels and settings.
- License Master (LM): Its role is to activate licensed functions and track the daily data entry volume.

As noted in the description of the current architecture, there is no HA at any layer. Therefore, another architecture with HA must be created for the storage layer and for the correlation layer, an approximate scheme could be:



Following the scheme of the architecture, in the *correlation layer* it will be necessary to deploy a *Deployer* for the configuration of the 3 *Search Heads in cluster* that is responsible for synchronizing the changes between the three *Search Heads*. With this setup, we managed to load balance, share dashboards, alerts, and searches with each other. Also, if one fails, the other two continue to work.

For the *storage layer, the deployment of a Master cluster will be required* for the configuration of the 2 *Indexers in cluster*. The *Master* is responsible for managing

replication and ensuring that each piece of data has at least one copy available. *With this configuration we manage* to replicate the data automatically and if one fails the other is still available.

This architecture is robust and with a high fault tolerance, ensuring constant access to data. Replication, in contrast, doubles storage.

Refurbishment of the current architecture

First, we will start by defining the specifications and calculating the size of the disk in order to reform the architecture reducing costs based on this data.

Specs:

Space required per event.	560 bytes
Min. storage time.	730 days
Events per second.	3000 EPS
Compression ratio	10:1

We start with calculating the number of events in the two years of minimum storage time:

$$3000\ \frac{eventos}{segundo} 3600\ seg\ \cdot\ 24\ \frac{h}{dia}\ 365\ \frac{dias}{año}\ \cdot\ 2\ año\ =\ 189.216.000.000\ eventos$$

We apply the compression ratio, i.e. for every 10 events we store 1:

$$\frac{189.216.000.000}{10} = 18.921.600.000\ eventos$$

We multiply it by the bytes that each event occupies and represent it in Tb:

$$18.921.600.000 \cdot 560 \frac{bytes}{evento}\ eventos =$$

$$10.596.096.000.000\ bytes\ \cdot \frac{1\ Kb}{1024\ bytes} \cdot \frac{1\ Mb}{1024\ Kb} \cdot \frac{1\ Gb}{1024\ Mb} \cdot \frac{1\ Tb}{1024\ Gb} =$$

$$9.63\ Tb$$

Then the minimum necessary storage size is 9.63 Tb (about 13.526 Gb/day), rounding up, about 10 Tb of storage is needed.

With these specifications, we now answer the questions in the *Splunk* questionnaire to define the requirements for indexing and search levels:

Nº	Pregunta	Consideraciones	Repercusión sobre la topología	Categoría de topología de nivel de indexador ♦	Categoría de topología de nivel de búsqueda ♦
1	¿Es su introducción de datos prevista inferior a ~300 GB/día?	Considere un crecimiento a corto plazo en la introducción diaria (~6-12 meses)	Candidato para una implementación de un único servidor, dependiendo de las preguntas relacionadas	S	1
			con la disponibilidad		
2	¿Requiere una alta disponibilidad de la recopilación/ indexado de los datos?	Si no tiene intención de utilizar Splunk para la supervisión de casos de uso que requieran una introducción de datos continua, una interrupción temporal del flujo de datos entrantes podría ser aceptable, siempre que no se pierdan datos de registro.	Requiere una implementación distribuida para dar cobertura a la introducción continua	D	1
3	Suponiendo que una cabeza de búsqueda realice una búsqueda: ¿Tienen sus datos que poder buscarse completamente en todo momento (por ej. no puede permitirse ningún impacto en la integridad de los resultados de las búsquedas?	Si su caso de uso está calculando mediciones de rendimiento y supervisión de uso general empleando funciones agregadas, por ejemplo, una interrupción aislada del indexador podría no afectar materialmente el cálculo de datos estadísticos sobre un número elevado de incidencias. Si su caso de uso es la auditoría de seguridad y la detección de amenazas, los puntos ciegos en los resultados de las búsquedas son muy probablemente poco deseables.	Requiere indexadores agrupados en clústeres con un factor de replicación de al menos dos (2). Nota: aunque un factor de replicación de 2 proporciona una protección mínima contra el fallo de nodo de indexador único, el factor de replicación recomendado (y predeterminado) es de 3.	C	1
4	¿Tiene centros de datos	Los requisitos de recuperación de	El funcionamiento	M	2

	múltiples y requiere la recuperación automática de su entorno de Splunk en caso de una interrupción del centro de datos?	desastres pueden dictar el funcionamiento continuo de dos instalaciones (activas/activas) o prescribir objetivos RTO/RPO para la recuperación de desastres manual	continuo requerirá la agrupación en clústeres de los indexadores en varios emplazamientos y al menos dos cabezas de búsqueda activas para garantizar la protección contra los fallos tanto en el nivel de introducción/ indexación de los datos como en el nivel de las búsqueda.		
5	Suponiendo una introducción de datos continua y sin pérdidas, ¿requiere alta disponibilidad para el nivel de búsqueda de cara al usuario?	Si se está utilizando Splunk para la supervisión continua casi en tiempo real, las interrupciones en el nivel de búsqueda no son tolerables probablemente. Esto puede ser cierto o no para otros casos de uso.	Requiere cabezas de búsqueda redundantes, y potencialmente la agrupación en clústeres de las cabezas de búsqueda	D/C/M	3
6	¿Necesita dar cobertura a un gran número de usuarios simultáneos y/o una carga de trabajo de búsqueda significativamente programada?	Los requisitos para más de ~50 usuarios/ búsquedas simultáneos normalmente requieren la ampliación horizontal del nivel de búsqueda	Puede ser necesaria una topología que utilice una agrupación en clúster de cabezas de búsqueda en el nivel de búsqueda	D/C/M	3
7	En un entorno de múltiples centros de datos, ¿necesita que se sincronicen los artefactos de los	Esto decidirá si los usuarios disfrutan de una experiencia vigente y coherente en el caso de una	Requiere una agrupación en clúster "extendida" de las cabezas de búsqueda entre sitios con una	M	4

	usuarios (búsquedas, paneles y otros objetos de conocimiento) entre sitios?	interrupción del sitio.	configuración apropiada. **Importante:** Aunque una SHC extendida puede mejorar la disponibilidad para los usuarios durante un fallo de sitio completo, no puede garantizarse que todos los artefactos se repliquen entre ambos sitios en todo momento. Esto puede afectar aplicaciones específicas que dependen de artefactos coherentes y vigentes, como la Aplicación Splunk para la seguridad empresarial. La agrupación en clústeres de cabezas de búsqueda por sí sola no puede proporcionar una solución DR completa. Otros beneficios para SHC sí se aplican.		
8	¿Tiene intención de implementar la Aplicación Splunk para la seguridad empresarial (ES)?	Asegúrese de leer y comprender las limitaciones específicas a las que está sujeta la Aplicación Splunk para la seguridad empresarial según se documenta con cada topología.	ES requiere un entorno de cabezas de búsqueda exclusivo (ya sea autónomo o agrupado en clúster).	D/C/M	+10
9	¿Tiene un entorno distribuido geográfica-mente que esté sujeto a normativas de custodia de datos?	Las normativas de algunos países no permiten que los datos generados dentro del país abandonen los sistemas de ese país.	Dichas normativas prohíben la implementación de un nivel de indexado central de Splunk y requieren que se desarrolle una arquitectura personalizada por parte de una colaboración entre Splunk/socio y el cliente que tenga en cuenta los detalles de dicha implementación en profundidad. En otras palabras, no hay una SVA para cumplir este requisito.	Personalizada	Personalizada

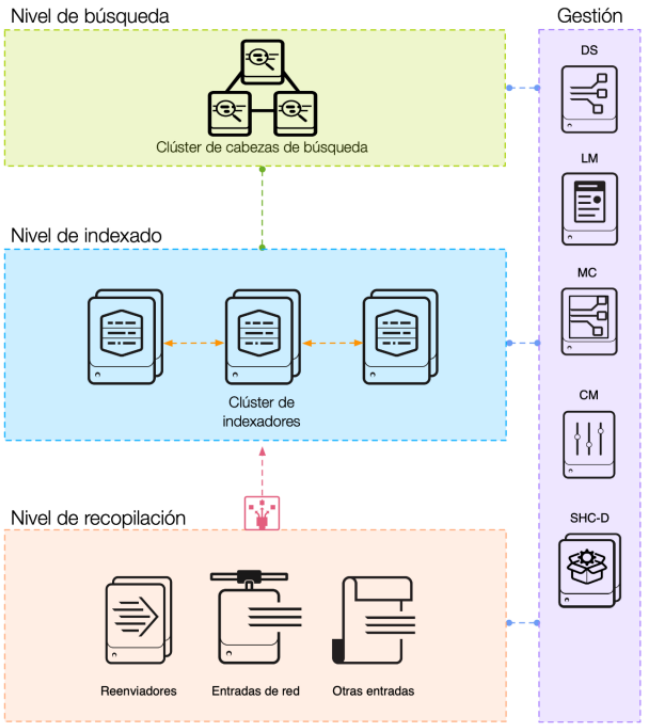
10	¿Tiene directrices de seguridad altamente restrictivas que impiden la ubicación conjunta de fuentes de datos de registro específicas en servidores/indexadores compartidos?	Es posible que no se permita que los datos de registro altamente confidenciales se ubiquen conjuntamente con conjuntos de datos de riesgo inferior en el mismo sistema físico o dentro de la misma zona de red en base a directrices corporativas.	Se necesitan entornos de indexado independientes y múltiples, potencialmente con un nivel de búsqueda híbrido compartido. Esto va más allá del ámbito de las SVA y requiere un desarrollo arquitectónico personalizado.	Personalizada	Personalizada

Yes or no answer to the ten questions:

Question	Answer	Justification
1	Yes	About 14 Gb/day
2	Yes	High availability of data collection/indexing is required since we are in an important service company.
3	Yes	Being in the security department, the blind spots in Search results are not permissible
4	No	It is not specified and we also work on a company
5	No	It is not specified that it is required for continuous monitoring, and in the initial scheme itself redundancy was not even implemented in the *Search Heads*.
6	Yes	It is not specified, but in the case of an important company in the sector, I imagine that it will need to cover more than 50 users.
7	No	This is the case of a company with a single branch.
8	No	The statement does not specify that you want to implement *Splunk* for enterprise security.
9	No	The statement does not require it and also because of the scheme of the initial architecture it can assume that it is not necessary.
10	No	The statement does not mention anything about the safety guidelines

Questions 1, 2, 3 and 6 are obtained as affirmative. Following Splunk's instructions we should implement the topology of the affirmative question with the highest number, i.e. 6, which is, "*D/C/M*" for the indexer topology and *"3"* for the search topology. As we can see, he recommends 3 topologies, in order to choose between them, the other affirmative questions must be consulted: 1, 2 and 3. Then, since question 3 is the second highest and recommends the indexer topology, "C", we are left with the "*C3*" architecture.

For a *"C3"* topology, a distributed cluster + SHC *must be deployed in* a single site. The architecture scheme would be something like:



It is an advanced design to ensure high availability, performance, and resiliency. Advantages over the one presented in the statement:

- **High Availability:** Failures do not interrupt the service.
- **Data Protection:** Replication to prevent loss of events.
- **Load Balancing:** Distributes ingestion and searches.
- **Scalability:** Nodes can be added without interruption.

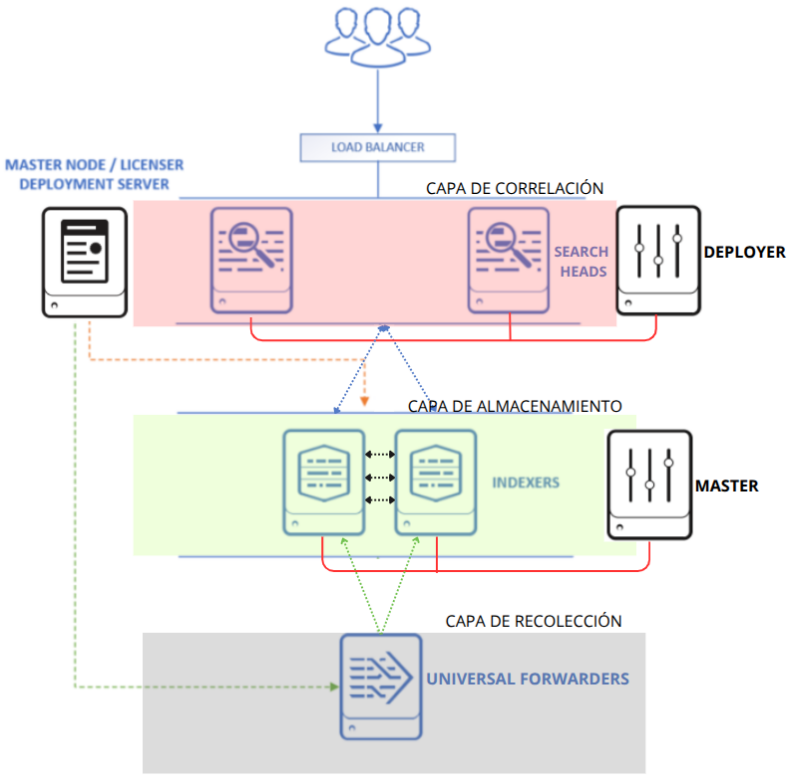
As in the statement, the boss asks us to reduce costs, although I am not sure if he is referring to after applying HA in the storage and correlation layers or before, I suppose after because it is a large company. Therefore, to reduce costs without losing too many features and maintain a basic HA, we can modify the architecture as follows:

- Correlation layer: keep only 2 *Search Heads* and there will be HA but without excessive redundancy. In addition, a lightweight *Deployer* can be maintained.

- Storage layer: keep both *indexers* in *cluster* with 1:1 replication (storage is doubled). Configurable *SmartStore Splunk* to move cold data to cloud storage or an on-premises NAS, keeping the data *Hot and warm* on fast disk and the *cold* in cheaper storage. Then the space needed with replication would amount to 20 Tb. Divided into about 5 Tb of fast disk (more expensive) and the rest in somewhat cheaper storage.

- Harvesting layer: change the *Heavy Forwader* for *Universal Forwarder,* lighter and cheaper.

The scheme of the architecture would be as follows:



A second option reduces even more features, assuming that question 6 is negative, that is, it will not be necessary to cover more than 50 users. Questions 1, 2 and 3 remain affirmative, in order to maintain the company's activity. Therefore, a "C1" topology would be needed . This offers a balance between cost and resilience. It is ideal if you need to secure the data. Advantages:

- **Data Protection:** If one *Indexer* fails, the other retains the data thanks to replication.
- **Moderate Cost:** A single *Search Head is maintained*, with no additional licenses.
- **Scalability: More** Indexers *can be added* as data ingestion increases.

Disadvantages of the *"C1" topology*:

- **No High Availability in Search:** If the *Search Head* fails, there is no continuity.

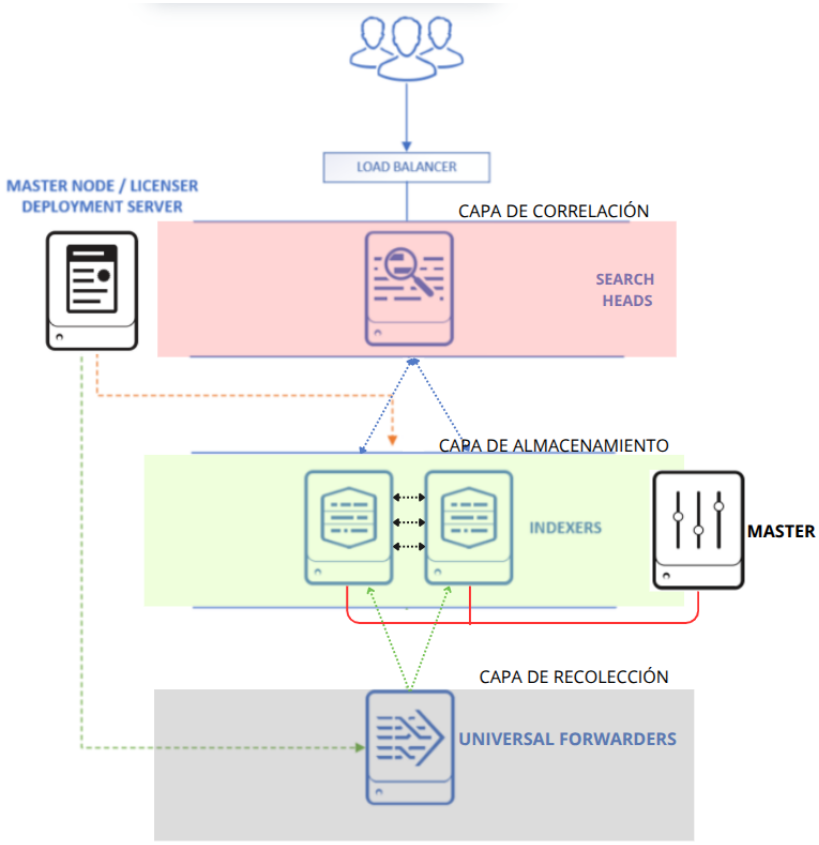
Architecture of the *"C1" topology*:

- Correlation layer: keep only 1 *Search Head,* without a cluster so that, if it fails, searches stop.

- Storage layer: keep both *indexers* in *cluster* with 1:1 replication (storage is doubled). Configurable *SmartStore Splunk* to move cold data to cloud storage or an on-premises NAS, keeping the data *Hot and warm* on fast disk and the *cold* in cheaper storage. Then the space needed with replication would amount to 20 Tb. Divided into about 5 Tb of fast disk (more expensive) and the rest in somewhat cheaper storage.

- Harvesting layer: change the *Heavy Forwader* for *Universal Forwarder,* lighter and cheaper.

The schematic of the topology *"C1"* would be:



Comparative Analysis of Implementations

Following the 2020 Gartner report, the main competitors with *Splunk* would be *IBM QRadar and Securonix*.

Features and comparison of *QRadar and Securonix:*

Characteristics	IBM QRadar	Securonix
Architecture	On-premise, virtualized and hybrid cloud.	Cloud native (SaaS)
Data ingestion	Appliance-based or virtual instance-based.	Scalable cloud ingestion with external storage.
Security Analysis	Traditional correlation rules and advanced analytics.	Analysis based on Machine Learning and UBA
Scalability	Requires additional Data Nodes for growth.	Horizontally scalable without its own infrastructure.
Storage	Local with external archiving (Data Nodes)	In the cloud, with flexible storage
Alert management	Automated prioritization of incidents.	Intelligent alerts based on risks and anomalies.
Use	Traditional interface, not too friendly	Modern and more intuitive interface.
Cost	High, licensed by EPS.	Pay-as-you-go model.

- QRadar *'s strengths and weaknesses:*

Strengths*:*

- Accurate detection with predefined and customizable rules.
- Native integration with IBM Watson for advanced analytics.
- Flexible deployment (*on-premise*, cloud, or hybrid).

Weaknesses:

- High cost and intensive consumption of resources.
- Limited scalability with no investment in new nodes.

- Securonix's strengths and weaknesses *:*

Strengths*:*

- Cloud-native, no need for physical infrastructure.
- Advanced analytics with *Machine Learning* and *User.* In addition, it employs *Entity Behavior Analytics* (UEBA).
- Flexible cost model, based on data volume and usage.

Weaknesses:

- Reliance on the cloud for storage and processing.
- It requires constant connectivity and may have latency if the network is limited.

In conclusion, if you prefer a more traditional, robust and *on-premise solution,* with advanced analytics, total control and, in addition, you have your own infrastructure, I would recommend the use of *IBM QRadar.* However, if you are looking for a *cloud-native* solution that is scalable, with advanced behavioral analytics and that requires a reduction in costs and licenses, then I would rather recommend *Securonix*.