Informe de práctica de Tecnologías SIEM (2025)

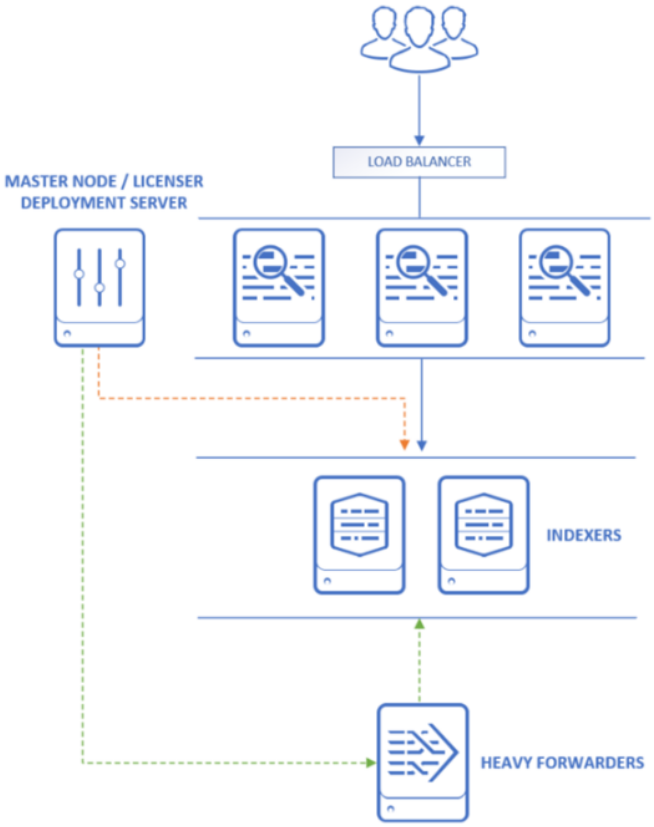
Autor: Pedro Oller Serrano

20/02/2025

Enunciado de la práctica

Entras en el Departamento de Seguridad de una importante compañía del sector servicios del país como experto en SIEM. Durante el primer día, tu jefe, el responsable de seguridad, te realiza una breve introducción sobre el sistema de monitorización de eventos de seguridad que tienen desplegado en la empresa con el fin de comenzar a darte visibilidad sobre dicha solución.

Lo primero que te indica es que se trata de una solución *on premise* basada en *Splunk*. La arquitectura desplegada es como la siguiente:



1. Dado que él directamente no gestiona este proyecto, y de cara a unos requisitos de auditoría interna que le han fijado, te solicita que le **especifiques**, sobre la ***actual arquitectura***, ***dónde se dispone de alta disponibilidad*** y ***dónde no***, en cuyo caso habrás de ***plantear una solución plausible para tal fin.***

2. Adicionalmente, y con motivo de realizar una ***disminución de costes***, te solicita que lleves a cabo un **estudio** sobre la viabilidad de ***modificar la arquitectura actual empleando otras alternativas posibles y que se las plantees, indicando sus características.***

3. De cara a realizar ese posible cambio de arquitectura, es necesario disponer de ciertos datos para buscar soluciones dentro del mercado que se ajusten a dicho escenario. Es por ello que, sabiendo que ***un evento son 560 bytes*** y que se necesita **almacenarlos** un ***mínimo de dos años*** por cumplimiento normativo, con un ***consumo***

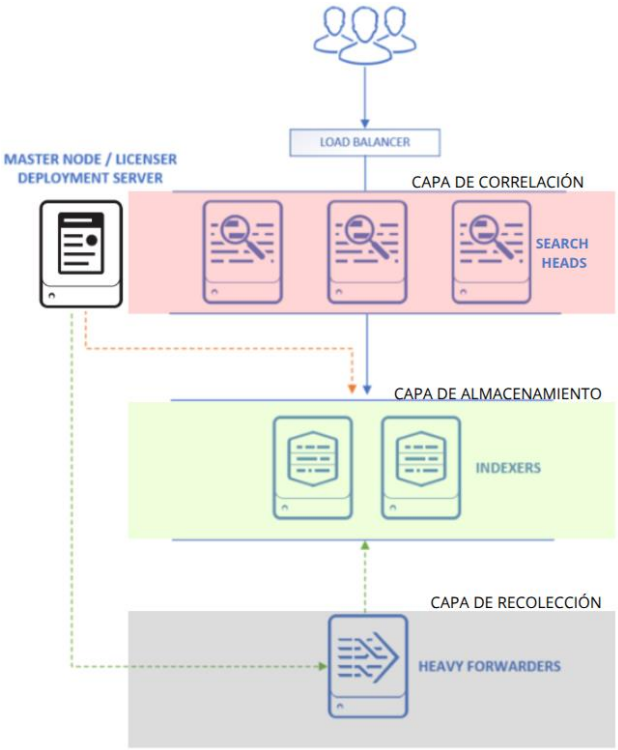
medio de ingesta de 3000 EPS (events per second), se requiere conocer el ***tamaño en disco*** que se precisará para almacenar esa cantidad de datos, asumiendo una ***ratio de compresión de 10:1***.

4. Por otro lado, es necesario ***saber cómo otros fabricantes afrontan este tipo de implementaciones*** y conocer sus puntos débiles y fuertes para tener una clara visión de sus funcionalidades (limitar la respuesta a ***dos fabricantes*** únicamente).

Resolución de la práctica

Especificaciones sobre la arquitectura e implementación de HA (High Availability)

En la siguiente figura está representado el esquema de la arquitectura implementada, imagino que el "Master node/ Licenser Deployment Server" hace referencia al servidor de licencias, por ello me he tomado la libertad de modificar el símbolo por el que emplea Splunk, para mayor claridad del ejercicio:

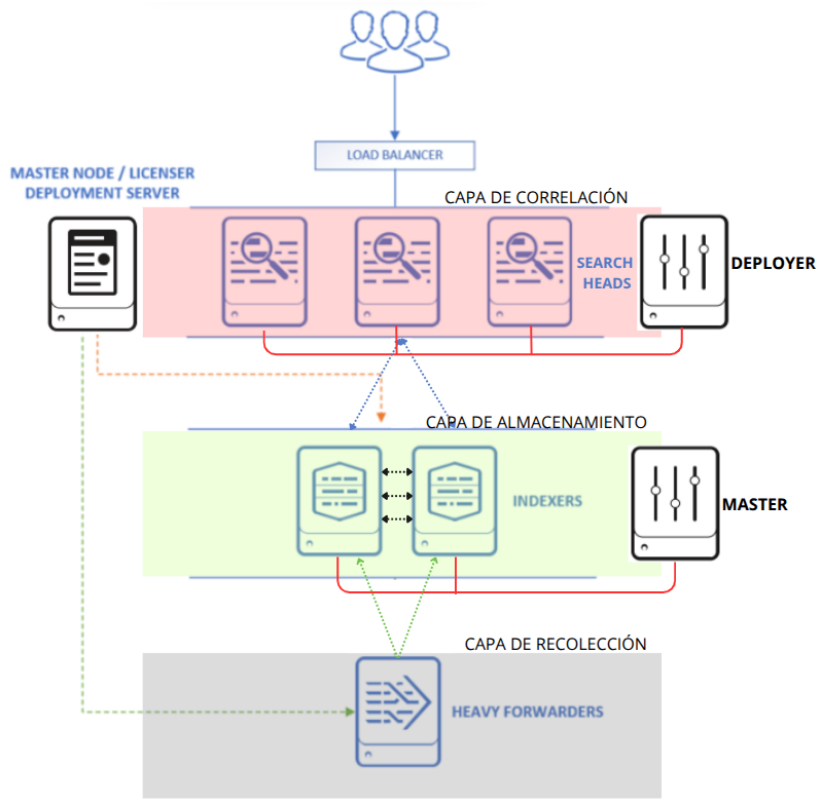


Primeramente, se llevará a cabo una descripción de la arquitectura por capas, para posteriormente, aplicar HA, tanto a la capa de almacenamiento como a la capa de correlación:

- Capa de recolección: En esta capa nos encontramos un *Heavy Forwarders,* que es un tipo de *Forwarder* con funcionalidades avanzadas de procesamiento, cuya función es al de recibir, procesar y enviar los datos a los dos *indexers* de la capa de almacenamiento. Algunas de sus características principales son:
  - Aplica *parsing, indexing parcial* y filtrado antes de reenviar los datos.
  - Transforma y descarta datos no deseados, disminuyendo la carga en los *indexers*.
  - Recoge datos de fuentes como API, bases de datos y archivos.
  - Assegura la transmisión mediante SSL
  - Proporciona redundancia si está en HA

- Capa de almacenamiento: en esta capa tenemos dos *indexers* sin *cluster*,es decir, no están en HA. Por lo que pueden estar configurados para una distribución de los datos enviados por el *Heavy Forwarders* o para configurar una separación de datos o para un simple respaldo por si uno cae que el otro continúe, pero sin replicación automática ya que no dispone de *clúster* por lo que se perderán los datos del *indexer* que haya caído.
- Capa de Correlación: en esta capa podemos observar un total de tres *Search Head* sin *clúster*, por lo que no hay replicación ni HA. Podrían estar configurados para una distribución de la carga o como respaldo manual, aunque sin *clúster no* hay *failover* por lo que si uno falla habría que configurar los demás y se perderían las búsquedas programadas, paneles y configuraciones.
- Maestro de licencias (LM): Su función es la de activar las funciones con licencia y realizar un seguimiento del volumen de introducción de datos diario.

Como se observa en la descripción de la arquitectura actual, no existe HA en ninguna capa. Por lo que, se debe crear otra arquitectura con HA para la capa de almacenamiento y para la capa de correlación, un esquema aproximado podría ser:



Siguiendo el esquema de la arquitectura, en la *capa de correlación* será necesario el despliegue de un *Deployer* para la configuración de los 3 *Search Heads en clúster* que se encarga de sincronizar los cambios entre los tres *Search Heads* . Con esta configuración, conseguimos balancear la carga, compartir paneles, alertas y búsquedas entre ellos. Además, si uno falla, los otros dos siguen funcionando.

Para la *capa de almacenamiento, s*erá necesario el despliegue de un *clúster Master* para la configuración de los 2 *Indexers en clúster*. El *Master* se encarga de gestionar la replicación y asegurar que cada dato tenga al menos una copia disponible. *C*on esta configuración conseguimos replicar los datos automáticamente y si uno falla el otro sigue disponible.

Esta arquitectura es robusta y con una alta tolerancia a fallos, garantizando un acceso constante a los datos. En contraposición, la replicación, duplica el almacenamiento.

Reforma de la arquitectura actual

Primeramente, comenzaremos definiendo las especificaciones y calculando el tamaño del disco para poder reformar la arquitectura reduciendo los costes en base a estos datos.

Especificaciones:

Espacio requerido por evento.	560 bytes
Tiempo mín. de almacenamiento.	730 días
Eventos por segundo.	3000 EPS
Ratio de compresión	10:1

Comenzamos con el cálculo del número de eventos en los dos años de tiempo mínimo de almacenamiento:

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

Aplicamos la ratio de compresión, es decir, por cada 10 eventos almacenamos 1:

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

Lo multiplicamos por los bytes que ocupa cada evento y lo representamos en Tb:

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

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

$$9.63\ Tb$$

Luego el tamaño necesario mínimo de almacenamiento es de 9.63 Tb (alrededor de 13.526 Gb/día), redondeando al alza, se necesitan unos 10 Tb de almacenamiento.

Con estas especificaciones, pasamos a contestar las preguntas del cuestionario de *Splunk* para definir los requisitos para los niveles de indexación y búsqueda:

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

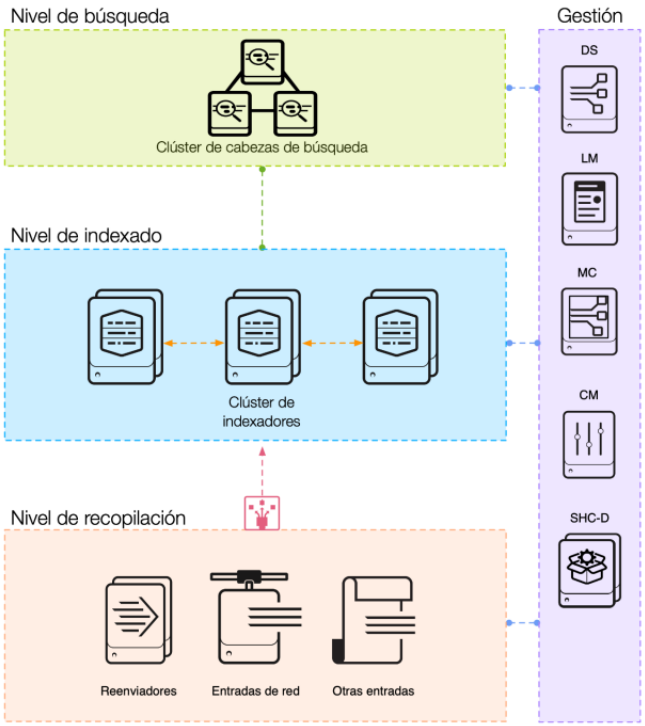
10	¿Tiene directrices de seguridad altamente restrictivas que impiden la ubicación conjunta de fuentes de datos de registro específicas en servidores/indexadores compartidos?	Es posible que no se permita que los datos de registro altamente confidenciales se ubiquen conjuntamente con conjuntos de datos 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

Respuesta de sí o no a las diez preguntas:

Pregunta	Respuesta	Justificación
1	Sí	Alrededor de 14 Gb/día
2	Sí	Se requiere de alta disponibilidad de la recopilación/indexado de datos ya que estamos en una empresa importante de servicios.
3	Sí	Al estar en el departamento de seguridad los puntos ciegos en los resultados de las búsquedas no son permisibles
4	No	No se especifica y además trabajamos sobre una empresa
5	No	No se especifica que se requiera para una supervisión continua, además en el propio esquema inicial ni siquiera estaba implementada la redundancia en los *Search Head*.
6	Sí	No se especifica, pero en el caso de una empresa importante del sector, imagino que se necesitará dar cobertura a más de 50 usuarios.
7	No	Se supone el caso de una empresa con una única sucursal.
8	No	El enunciado no especifica que se quiera implemantar *Splunk* para la seguridad empresarial.
9	No	El enunciado no lo requiere y además por el esquema de la arquitectura inicial puede suponer que no es necesario.
10	No	El enunciado no menciona nada de las directrices de seguridad

Se obtiene como afirmativas las cuestiones: 1, 2, 3 y 6. Siguiendo las instrucciones de *Splunk* deberíamos implementar la topología de la cuestión afirmativa con el número más alto, es decir la 6, la cual es, "*D/C/M*" para la topología de indexador y la *"3"* para la topología de búsqueda. Como vemos nos recomienda 3 topologías, para poder elegir entre ellas, se debe consultar las demás cuestiones afirmativas: 1, 2 y 3. Luego, como la cuestión 3 es la segunda más alta y recomienda la topología de indexador, "C", nos quedamos con la arquitectura "*C3*".

Para una topología *"C3"* se debe implementar un *clúster distribuido + SHC en* un único sitio. El esquema de la arquitectura sería algo como:



Es un diseño avanzado para garantizar alta disponibilidad, rendimiento y resiliencia. Ventajas respecto del presentado en el enunciado:

- **Alta Disponibilidad:** Los fallos no interrumpen el servicio.
- **Protección de Datos:** Replicación para evitar pérdida de eventos.
- **Balanceo de Carga:** Distribuye la ingesta y las búsquedas.
- **Escalabilidad:** Se pueden agregar nodos sin interrupciones.

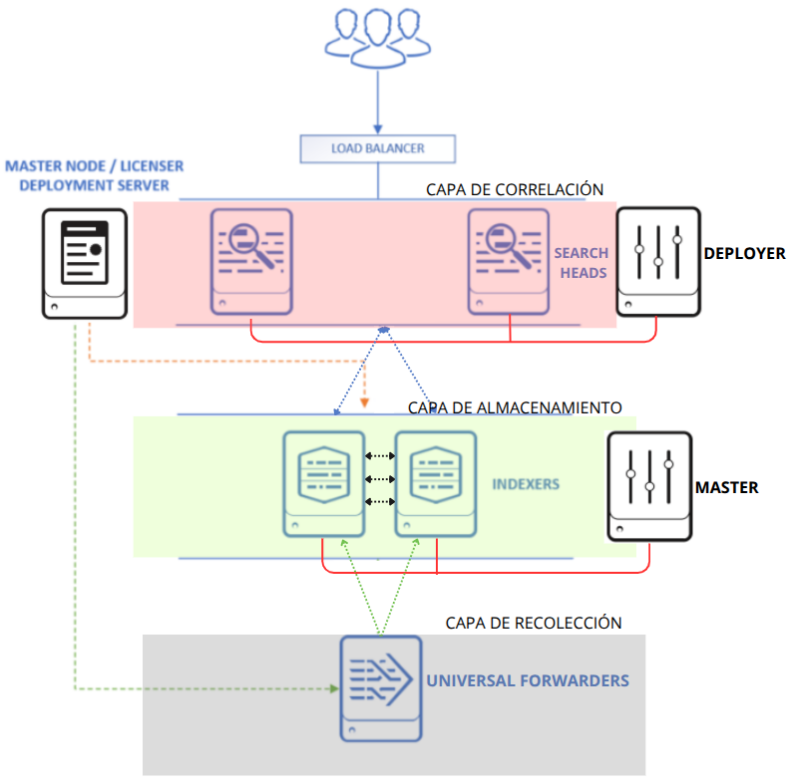
Como en el enunciado el jefe nos pide reducir costes, aunque no tengo claro si se refiere a después de aplicar HA en las capas de almacenamiento y correlación o antes, supongo que después porque es una gran empresa. Por lo que, para reducir costes sin perder demasiadas prestaciones y mantener una HA básica, podemos modificar la arquitectura como sigue:

- Capa de correlación: mantener únicamente 2 *Search Head* y así habrá HA pero sin excesiva redundancia. Además, se puede mantener un *Deployer* liviano.

- Capa de almacenamiento: se mantienen los dos *indexers* en *clúster* con una replicación 1:1 (se duplica el almacenamiento). Se puede configurar *SmartStore Splunk* para mover datos fríos a almacenamiento en la nube o a un NAS local, manteniendo los datos *hot y warm* en disco rápido y los *cold* en un almacenamiento más económico. Luego el espacio necesario con replicación ascendería a 20 Tb. Repartido en alrededor de 5 Tb de disco rápido (más caro) y lo demás en almacenamiento algo más barato.

- Capa de recolección: cambiar el *Heavy Forwader* por *Universal Forwarder,* más ligero y económico.

El esquema de la arquitectura sería el siguiente:



Una segunda opción reduciendo aún más prestaciones, suponiendo que la cuestión 6 es negativa, es decir, no se necesitará dar cobertura a más de 50 usuarios. Las cuestiones 1, 2 y 3 siguen siendo afirmativas, para mantener la actividad de la empresa. Por lo que, se necesitaría una topología *"C1"*. Esta, ofrece un equilibrio entre coste y resiliencia. Es ideal si se necesita asegurar los datos. Ventajas:

- **Protección de Datos:** Si un *Indexer* falla, el otro conserva los datos gracias a la replicación.
- **Coste Moderado:** Se mantiene un solo *Search Head*, sin licencias adicionales.
- **Escalabilidad:** Se pueden agregar más *Indexers* según aumente la ingesta de datos.

Desventajas de la topología *"C1"* :

- **Sin Alta Disponibilidad en Búsquedas:** Si el *Search Head* falla, no hay continuidad.

Arquitectura de la topología *"C1"*:

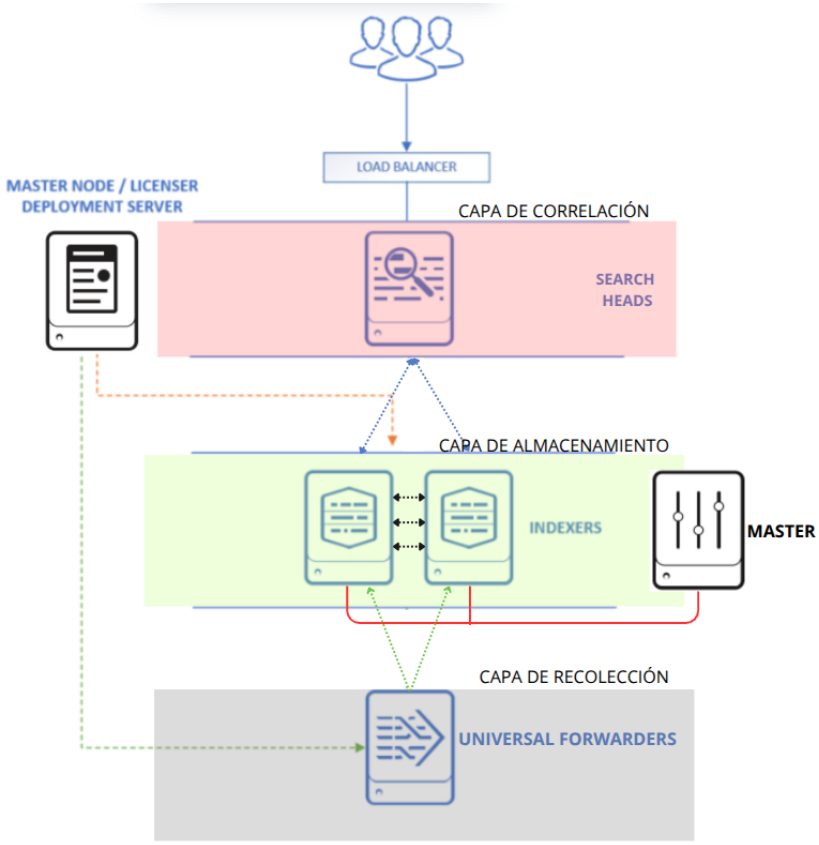
- Capa de correlación: mantener únicamente 1 *Search Head,* sin clúster por lo que, si falla, las búsquedas se detienen.

- Capa de almacenamiento: se mantienen los dos *indexers* en *clúster* con una replicación 1:1 (se duplica el almacenamiento). Se puede configurar *SmartStore Splunk* para mover datos fríos a almacenamiento en la nube o a un NAS local, manteniendo los datos *hot y warm* en disco rápido y los *cold* en un almacenamiento

más económico. Luego el espacio necesario con replicación ascendería a 20 Tb. Repartido en alrededor de 5 Tb de disco rápido (más caro) y lo demás en almacenamiento algo más barato.

- Capa de recolección: cambiar el *Heavy Forwader* por *Universal Forwarder,* más ligero y económico.

El esquema de la topología *"C1"* sería:



Análisis Comparativo de Implementaciones

Siguiendo el informe Gartner de 2020, los principales competidores con *Splunk* serían *IBM QRadar y Securonix*.

Características y comparativa de *QRadar y Securonix:*

Características	IBM QRadar	Securonix
Arquitectura	On-premise, virtualizado y nube híbrida.	Nativo en la nube (Saas)
Ingesta de datos	Basada en appliances o instancias virtuales.	Ingesta escalable en la nube con almacenamiento externo.
Análisis de seguridad	Reglas de correlación tradicionales y analítica avanzada.	Análisis basado en Machine Learning y UBA
Escalabilidad	Requiere Data Nodes adicionales para crecimiento.	Escalable horizontalmente sin infraestructura propia.
Almacenamiento	Local con archivado externo (Data Nodes)	En la nube, con almacenamiento flexible
Gestión de alertas	Priorización automatizada de incidentes.	Alertas inteligentes basadas en riesgos y anomalías.
Uso	Interfaz tradicional, no es demasiado amigable	Interfaz moderna y más intuitiva.
Coste	Alto, con licenciamiento por EPS.	Modelo de pago por uso.

- Puntos débiles y fuertes de *QRadar:*

Puntos fuertes*:*

- Detección precisa con reglas predefinidas y personalizables.
- Integración nativa con IBM Watson para análisis avanzado.
- Despliegue flexible (*on-premise*, nube o híbrido).

Puntos débiles:

- Coste elevado y consumo intensivo de recursos.
- Escalabilidad limitada sin inversión en nuevos nodos.

- Puntos débiles y fuertes de *Securonix:*

Puntos fuertes*:*

- Nativa en la nube, sin necesidad de infraestructura física.
- Análisis avanzado con *Machine Learning* y *User*. Además, emplea *Entity Behavior Analytics* (UEBA).
- Modelo de coste flexible, basado en volumen de datos y uso.

Puntos débiles:

- Dependencia de la nube para almacenamiento y procesamiento.
- Requiere conectividad constante y puede tener latencia si la red es limitada.

En conclusión, si se prefiere una solución más tradicional, robusta y *on-premise,* con analítica avanzada, control total y, además se cuenta con infraestructura propia, yo recomendaría el uso de *IBM QRadar*. Sin embargo, si se busca una solución *cloud-native,* que sea escalable, con analítica avanzada de comportamientos y que se

requiera de una reducción de los costes y licencias, entonces recomendaría más bien *Securonix*.