ARQUITECTURA DE SISTEMAS DE OBTENCIÓN Y EXPLOTACIÓN DE INFORMACIÓN EN FUENTES ABIERTAS

Autor: Tomás Aguado Gómez
Director/es: Norberto Fernández García

I. INTRODUCCIÓN Y CONTEXTO

La globalización del mundo de la seguridad y su progresiva sofisticación han ocasionado que grupos terroristas, cibercriminales, incluso servicios de inteligencia hostiles hayan convertido el mundo virtual en un teatro de operaciones más en el que desarrollar actividades. Este nuevo "campo de batalla" centrado en Internet hace que la obtención de información de fuentes a abiertas sea uno de los campos claves para luchar contra este tipo de actividades.

Recuperar y analizar información de fuentes abiertas se ha convertido en la actualidad en un problema centrado en la obtención, almacenamiento, clasificación y análisis de grandes volúmenes de datos obtenidos de fuentes web cada vez más sofisticadas y especializadas que, en ocasiones, dificultan exprofeso la extracción de información.

La capacidad de generar una imagen unificada y coherente de lo que está "pasando en Internet" a partir del enjambre de fuentes de datos existentes es un problema tan complejo como urgente de resolver. En los últimos años además a esta complejidad inherente se ha añadido el hecho de que ciertos actores han utilizado este medio para difundir noticias falsas o adulteradas con el objeto de influir sobre las conductas de diferentes comunidades.

Es necesario pues contar con un sistema no sólo de extracción sino también de análisis y fusión de información con las fuentes de datos de *backoffice* que pueda ayudar a los analistas en esta disciplina. Si bien existen herramientas de pago con cierta trayectoria en este ámbito, su adopción en este campo para algunas organizaciones, especialmente las relacionadas con seguridad y defensa, supone ciertas desventajas y servidumbres que no pueden asumir. Asimismo, también existen proyectos de código abierto orientados a satisfacer estas necesidades, aunque no siguen un enfoque holístico y sólo resuelven parte del problema planteado.

Se hace pues necesario plantear una arquitectura abierta basada en componentes estándar que permita adaptarse a las necesidades y particularidades de la obtención y análisis OSINT.

II. DESARROLLO Y RESULTADOS

Tras realizar un intensivo análisis del estado del arte que permita establecer ciertos bloques constructivos básicos que podrían utilizarse para implementar la arquitectura, se desarrolla la arquitectura en base a cuatro pilares fundamentales:

1. Almacenar. Los módulos necesarios para el almacenamiento y procesamiento distribuido de grandes volúmenes de información basados en plataformas como Hadoop, pero sin desdeñar otros más tradicionales como las bases de datos relacionales o las bases de datos de grafos/documentales.
2. Descubrir y desarrollar. Una vez almacenados los datos, se analizaron las operaciones que se podrían realizar por parte de usuarios avanzados, así como las herramientas necesarias. Estos usuarios avanzados tienen como objetivo final generar productos de datos para los usuarios finales y la interacción principal con los datos será directamente a través de lenguajes de programación como R o Python.
3. Consumir. Este módulo trata de establecer las funcionalidades necesarias para poder explotar la información desde el punto de vista de los analistas o usuarios finales. Este es quizá el entorno en el que mejor se muevan las herramientas comerciales.
4. Gobernar. Esta infraestructura pasa a convertirse en el nodo central de las tareas de análisis, no sólo de la información OSINT sino también del resto de datos de la organización. Es tan importante y compleja desde el punto de vista tecnológico como organizacional que deben establecerse las estructuras y procedimientos de gobierno necesarios. Desde el punto de vista técnico se propone utilizar de bases sistemas de control de metadatos y linaje del dato integrados en el sistema como Apache Atlas.

No sólo se abordan estos bloques principales sino también ciertas cuestiones transversales de capital importancia a la hora de definir la arquitectura.

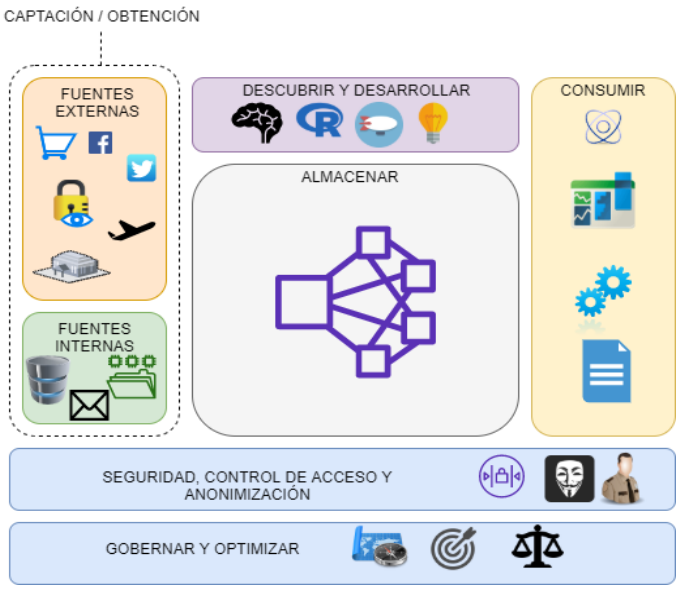

Figura 1 Vista modular de la arquitectura propuesta.

1. Seguridad. Introducir todos los datos en un único repositorio de datos sobre el que accede un ecosistema enormemente complejo de aplicaciones supone un gran desafío desde el punto de vista de la seguridad. Por esta razón ya se plantea desde la concepción de la arquitectura un sistema de control de acceso integrado en la propia plataforma que permita satisfacer esta necesidad.
2. Análisis avanzado. En lo que se refiere a información OSINT, no sólo crece el volumen de datos sino también su falta de estructura. La aparición de algoritmos de tratamiento avanzado unida a la mejora y democratización de los que ya existían permite obtener valor de tipos de datos que antes se desechaban.
3. Roles. La explotación de este tipo de arquitecturas plantea también la necesidad de contar con ciertos roles específicos relacionados con el análisis de datos y con su gobernanza que permitan aprovechar de forma efectiva las capacidades de este tipo de arquitecturas.

III. CONCLUSIONES

Tras el desarrollo de la arquitectura ha podido constatarse la complejidad del problema que se trataba de resolver. No sólo se ha planteado una arquitectura genérica que permitiría resolver el problema de la obtención y el análisis de información OSINT, sino que también se han planteado opciones concretas de implementación en todos los módulos que permitirían instanciar una arquitectura de este tipo.

Esto se considera imprescindible de cara a comprobar la viabilidad de la arquitectura. Asimismo, se plantean los flujos de información y los protocolos de información que deberán existir entre todos los módulos de la arquitectura garantizando que las herramientas concretas propuestas los soportan.

Plantear la arquitectura a través de las perspectivas anteriormente descritas ha permitido poder abordar la complejidad del problema planteado de forma modular. Este enfoque hace posible además poder prescindir además de ciertos módulos en caso de que no sean necesarios o ya se encuentren implementados dentro de cada organización.

Durante este desarrollo se detectan posibles carencias, mejoras y diferentes opciones que no se han explorado en profundidad, pero se plantean como mejoras posibles de la arquitectura. Estas mejoras se plantean no sólo desde el punto de vista técnico sino también desde el organizacional.