Revista mktDescubre - ESPOCH FADE
N° 11 Junio 2018, pp. 50 - 66
Latindex Folio Catálogo 20003
ISSN - 1390 - 7352 IEPI Título N° 3232 - 12

MAPEO SISTEMATICO DE LITERATURA DE UN DATA LAKE

Margarita Alejandra Aucancela Guamán
maucancela@espoch.edu.ec

Myriam Johanna Naranjo Vaca
myriam.naranjo@hotmail.com

José Francisco Betún Yuquilema
jesebetun@gmail.com
Escuela Superior Politécnica de Chimborazo
Facultad de Administración de Empresas

RESUMEN

El crecimiento exponencial de los datos en las organizaciones ha generado el desarrollo de nuevas tecnologías como un Data Lake o Laguna de Datos. En este trabajo se han planteado preguntas de investigación que permitieron determinar su definición, utilidad, importancia, arquitectura, funciones y aportes que genera la utilización de esta tecnología, para ello se llevó a cabo un Mapeo Sistemático de Literatura (MSL). Como resultados se definió que un Data Lake es un repositorio de datos de bajo costo que permite almacenar datos estructurados, no estructurados y semi estructurados. La tecnología que permite la implementación de un Data Lake es Hadoop, lo que obliga a los analistas de los datos a investigar sobre su implantación.

PALABRAS CLAVES: arquitectura, data lake, funciones, gestión de datos, ventajas, desventajas.

ABSTRACT

The exponential growth of data in organizations has generated the development of new technologies such as a Data Lake or Data Lagoon. In this work have been raised research questions that allowed to determine its definition, utility, importance, architecture, functions and contributions that generates the use of this technology, for it was carried out a Systematic Mapping of Literature (MSL). As a result, it was defined that a Data Lake is a low-cost data repository that allows the storage of structured, unstructured and semi-structured data. The technology that allows the implementation of a Data Lake is Hadoop, which forces data analysts to investigate its implementation.

KEYWORDS: Architecture, advantages, analytics, data lake, data management, disadvantages, functions

Fecha recepción: marzo 2018
Fecha aceptación: junio 2018

1. INTRODUCCIÓN

El volumen de crecimiento de los datos se duplica cada 18 meses (Management Solutions, 2015), esto ha generado la necesidad de desarrollar tecnologías que permitan su gestión y explotación. Sin una correcta administración, es imposible convertir los datos en información fiable, tanto para un uso operativo básico como para la toma de decisiones estratégicas. El sinfín de señales que denotan que una empresa posee problemas o dificultades recurrentes en la administración de datos se caracteriza por (POWER DATA, 2014):

- Disponer de datos de mala calidad o baja calidad: inexactos, inconexos, duplicados, incorrectos, caducos...
- Inventario agotado o sobre abastecido.
- Derroches de correo debido a direcciones incorrectas.
- Errores de facturación.
- Pérdidas de ingresos por oportunidades perdidas.
- Gestión ineficaz de los datos maestros que impidan una visión única.
- Creciente cantidad y diversidad de datos.
- Sanciones debido al incumplimiento de la regulación.
- Falta de inversiones en TI para tratar los datos masivos.
- No poder trasformar esos datos en inteligencia de negocio.

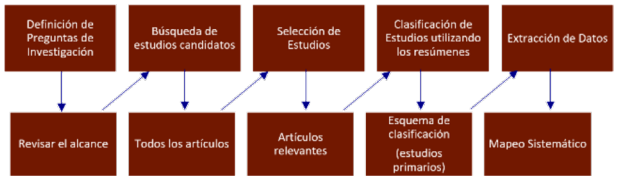
Una solución prometedora que hoy en día se plantea a las organizaciones es el uso de una laguna de datos o Data Lake. Pero ¿Qué es un Data Lake? ¿Para qué sirve un Data Lake? ¿Cuál es la arquitectura de un Data Lake? ¿Cómo implementar un Data Lake? ¿Por qué es importante la gestión de un Data Lake? ¿Cuáles son las funciones de un Data Lake? ¿En qué sectores es posible utilizar un Data Lake? ¿Cuáles son los aportes que ofrece esta tecnología a las organizaciones? ¿Cuáles son las ventajas de la gestión de un Data Lake? ¿Cuáles son las desventajas de la gestión de un Data Lake?

Este trabajo de investigación pretende responder las preguntas planteadas utilizando un mapeo sistemático de literatura, con el objetivo de ampliar el conocimiento sobre esta nueva tecnología que promete solucionar problemas de gestión de datos, ampliar la inversión y la innovación en este tipo de tecnologías, así como generar valor comercial.

2. MATERIALES Y MÉTODOS

El mapeo sistemático de literatura según (Petersen, 2015) es un método para construir clasificaciones y conducir análisis temáticos a efecto de obtener un mapa visual del conocimiento existente dentro de un tema amplio. El análisis de los resultados se realiza categorizando los hallazgos y calculando la frecuencia de publicaciones dentro de cada

categoría, para determinar la cobertura de las distintas áreas de un tema de investigación específico. La información generada se puede combinar para responder preguntas de investigación específicas y ahorrar tiempo y esfuerzo en la investigación. Para que esto sea posible, los mapeos sistemáticos de literatura deben ser de calidad en términos de completitud y rigurosidad (Kitcheham, 2011). Siguiendo las directrices de (Petersen, 2015) y (Kitcheham, 2011), en el presente estudio se dio seguimiento a las siguientes fases: Definición de las preguntas de investigación, Búsqueda de estudios candidatos, Selección de Estudios, Clasificación de Estudios utilizando los resúmenes y la Extracción de Datos (ver Figura 1).



imagenes\proceso de mapeo.png
Figura 1
Etapas del Proceso de Mapeo Sistemático de Literatura
Fuente: (Kitcheham, 2011) & (Petersen, 2015)
Elaborado por: Los autores

El mapeo sistemático de literatura cuenta con investigaciones anteriormente publicadas, que permiten dar respuesta a las preguntas planteadas y descartar aquellas investigaciones que no aportan al tema de investigación

Definición de las preguntas de Investigación

Según los lineamientos propuestos por (Petersen, 2015), las preguntas de investigación son un elemento clave del mapeo sistemático de literatura, en vista de que reflejan sus objetivos y proporcionan una visión general del tema de investigación (Souag, Mazo, Salinesi, & Comyn-Wattiau, 2015).

Fuente de Datos y Búsqueda de Estudios Candidatos

En esta segunda etapa se realizó una búsqueda inicial con palabras clave como: Data Lake, Data Management, Characteristics. La búsqueda se realizó en las bases científicas de Scopus, IEEE, ACM, Springer, Science Direct y Google Académico. Una vez seleccionados los artículos que para esta ocasión se denominan: estudios candidatos, se procede a revisar el título, el resumen y las palabras clave para obtener términos afines, comunes y generales, con los cuales se elaboraron cadenas de búsqueda para refinar la obtención de artículos fuente.

Selección de estudios

Una vez seleccionada la cadena de búsqueda CD4, siguiendo las directrices de (Kitcheham, 2011) se aplicó criterios de inclusión y exclusión para la selección

N°	CADENA DE BUSQUEDA
CD1	(Data Management) AND (Data lake OR mining OR Concepts and elements of Data mining) AND (Management OR application) AND (Using OR benefits).
CD2	(Data lake OR Data Management) AND (Data mining OR Concepts) AND (managing data lakes OR services) AND (Differences OR Description).
CD3	(Data lake) AND (Management OR Implementation OR Description) AND (structure).
CD4	(Data lake) AND (managing data lakes OR services) AND (Differences OR application) AND (Using OR Advantage OR Approach).
CD5	(Data lake OR Data Management) AND (Data mining OR Number) AND (Advantage OR Approach) AND (Using OR benefits OR Differences).

imagenes\tabla cadena de busqueda.png

de artículos, los cuales se llamarán estudios candidatos.

- Los criterios de inclusión consideran todos los artículos relacionados al tema de investigación
- Los de exclusión se aplican para aquellos artículos técnicos que no se relacionan con el tema de investigación;

Clasificación de artículos

Una vez obtenidos los estudios candidatos 51 en número, se sometieron a votación de los investigadores, obteniendo un total de 27 estudios los que se llamaran primarios

Extracción de datos

Una vez clasificados los artículos se procedió a revisarlos para responder las preguntas de investigación planteadas

3. RESULTADOS

En este apartado, se contestó las preguntas de investigación, a continuación, su respuesta.

:

P1. ¿Qué es una Data Lake?

Un Data lake o laguna de datos, es un repositorio digital construido para almacenar una gran cantidad de datos en formato nativo, es decir que los datos son ingresados en su sistema de procesamiento sin comprometer su estructura. Según (Alserafi, 2016), un Data Lake es un enorme repositorio de datos en bruto.

Las fuentes de información que alimentan a un Data Lake son: Medios de Pago, Redes Sociales: videos, imágenes, audios, Bases de datos relacionales, Sistemas de Información Gerencial: CRM, SCM y ERP, logs de servidores, correos electrónicos, registros de llamadas telefónicas, dispositivos móviles y datos en especial

del Internet de las cosas(IoT) (Hagstroem, Roggendorf, Saleh, & Sharma, 2017).

Las tecnologías empleadas para la implementación de un Data Lake son de la familia Hadoop (LaPlante & Sharma, 2016), estas se caracterizan por ser de bajo costo y por las mejoras que presentan en la captura, el refinamiento, el archivo y la exploración de datos brutos dentro de una empresa (Fang, 2015). Otra tecnología es la de Cloudera.

Un Data Lake contiene datos multi-estructurados, que en su mayoría tienen un valor no reconocido para la organización (Fang, 2015), sin embargo, estos permiten obtener conocimiento e información que aportan a las decisiones clave de la organización. (LaPlante & Sharma, 2016). La generación del conocimiento se apoya en el uso de aplicaciones de análisis dinámico que no necesariamente requieren de una estructura estática como un Datawarehouse. (Miloslavskaya & Tolstoy, 2016).

Por otro lado, el Big data ha motivado el surgimiento de esta tecnología, según (Capgemini, 2017) las tecnologías del Big Data, incluido el Data Lake evolucionan cada 6 meses.

P2. ¿Para qué sirve un Data Lake?

Hay una variedad de formas en que se puede utilizar un Data Lake:

- Ingestión de fuentes de datos multi estructuradas (Khine & Zhao, 2017): estructuradas, semiestructuradas y no estructuradas (también conocidas como big data), como lecturas de equipos, datos de telemetría, registros, transmisión de datos, etc.
- Un Data Lake es un lugar conveniente para colocar datos para el análisis experimental (Stodder, 2017) , incluso antes que su valor o propósito haya sido completamente definido. La agilidad es importante para todos los negocios en estos días, por lo que un Data Lake puede jugar un papel importante en situaciones de “prueba de valor” debido al enfoque “ELT” (Extract, Load and Transform), al soporte analítico avanzado en tiempo real y a la posibilidad de ejecutar algoritmos de aprendizaje automático.
- Un Data Lake puede ser utilizado como plataforma ETL u otras rutinas de preparación de datos y creación de perfiles del sistema de almacenamiento de datos para que las organizaciones no se vean obligadas a expandir sus datas warehouse y sistemas de integración de datos existentes, y paguen lo que cuesta esa expansión (Stodder, 2017)
- Soporte para arquitectura Lambda que incluye una capa de velocidad, capa de lote y capa de servicio.
- Mas que un Datawarehouse (DW), un Data Lake permite contener datos

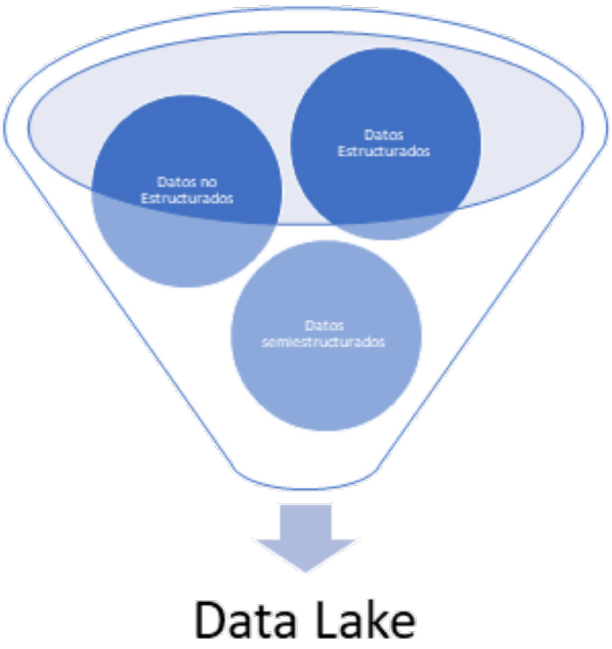
que no se almacenan fácilmente en un DW o que no se consultan con frecuencia. Se puede acceder al Data Lake a través de consultas federadas (capa de virtualización de datos) que hacen que la separación del DW sea transparente a los usuarios finales.

- Archivado y Almacenamiento histórico de datos de toda la organización para respaldar las actividades de análisis e informes que puedan arrojar resultados valiosos.
- Un Data Lake puede convertirse en un tipo de almacén de datos operacional desde el cual se pueden mover, transformar y limpiar datos para las herramientas de la data warehouse, data marts o business intelligence (BI) según lo necesiten los usuarios (Stodder, 2017).
- Soporte de aplicaciones. Además del análisis, un Data Lake puede comportarse como la fuente de datos de una aplicación de tipo front-end. (SQL Chick, 2016)
- Un Data Lake puede trabajar directamente con datos en clústeres Hadoop y almacenamiento basado en la nube utilizando técnicas de procesamiento en la base de datos para evitar el movimiento innecesario de datos (Stodder, 2017).
- Un Data Lake basado en los clústeres de Hadoop permite que las aplicaciones y los trabajos de análisis saquen el máximo provecho del procesamiento distribuido asociado a un almacén de datos lógico, las bases de datos columnares, la computación de memoria y la rápida transmisión e interacción de datos (Stodder, 2017).

P3. ¿Cuál es la arquitectura de un Data Lake?

Un Data Lake utiliza una arquitectura plana centrada en los datos, que almacena grandes volúmenes de datos en varios formatos, es decir datos sin procesar (Knowledgent Group Inc., 2014). Básicamente, en un Data lake los datos ingresan por procesamiento por lotes o procesamiento en tiempo real. Cada entidad de datos en la laguna está asociada con un identificador único y un conjunto de metadatos extendidos (Salemink, 2017), los cuales forman un catálogo de datos, el mismo que es utilizado por los consumidores (científicos de datos, analistas de negocios) para crear esquemas específicos de acuerdo con sus necesidades. Cabe recalcar que en este caso el Data Lake ejerce un rol de proveedor de datos que proporciona datos y análisis como un servicio (DAaS) (Knowledgent Group Inc., 2014).

La comprensión de la naturaleza de los datos se delega al consumidor de los datos en el momento de su recuperación (es decir, el tiempo de consulta). Cuando se recuperan los datos, el usuario transformará esos datos según las partes de la empresa para adquirir información comercial. (Khine & Zhao, 2017).

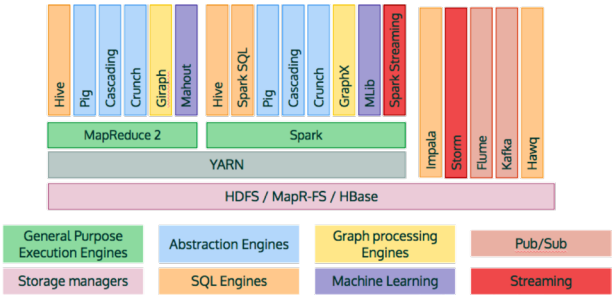


imagenes\arquitectura data lake.png
Figura 2
Vista simplificada de un Data Lake
Fuente: (Khine & Zhao, 2017)
Elaborado por: Los autores

P4. ¿Cómo implementar un Data Lake?

Hadoop (plataforma de datos orientada a objetos de alta disponibilidad), es una herramienta de Big data ampliamente utilizada para la carga de datos en un Data Lake. Hadoop es un framework que permite el proceso distribuido de grandes volúmenes de datos entre clusters de computación (Farrugia, Claxton, & Thompson, 2016). Está diseñado para escalar desde un solo servidor a miles de máquinas, cada una ofreciendo capacidad de cálculo y de almacenamiento (Capgemini, 2017).

Hadoop tiene dos componentes principales: HDFS (Hadoop Distributed File System) y el motor MapReduce. HDFS File System maneja el punto único de falla y escalabilidad al replicar múltiples copias de bloques de datos en diferentes nodos del clúster. Todos los datos almacenados en estos bloques de datos se procesarán en el enfoque MapReduce. Los datos se recuperarán como una lista de pares clave-valor, es decir, la fase del mapa. Las mismas claves de datos se barajarán, clasificarán y enumerarán en grupos para realizar las operaciones necesarias, es decir, reducir la fase. Todos los datos producidos por una empresa se incluirán en el grupo de datos Hadoop Cluster.



imagenes\stack tecnológico de Hadoop.png
Figura 5
Stack Tecnológico de Hadoop
Fuente: (Capgemini, 2017)

Para la carga en tiempo real, un Data Lake utiliza el marco de procesamiento de flujo de: Apache Spark o Apache Flink. Los datos requeridos se transformarán de acuerdo con la necesidad de los sistemas analíticos sobre la marcha en el tiempo de consulta. Además, puede incluir una base de datos semántica, un modelo conceptual y agregar una capa de contexto para definir el significado de los datos y su interrelación con otros datos. Se puede decir que la estrategia de Data Lake incluye almacenar todo tipo de datos (variedad de datos) de las bases de datos SQL y NoSQL, así como combinar los conceptos de OLTP con OLAP. (Khine & Zhao, 2017)

Para implementar un Data Lake se deben seguir los siguientes pasos (Cito Research, 2014):

Etapa 1: Manejo de datos a escala. La primera etapa implica instalar el canal de la comunicación para extraer y transformar los datos a escala. En esta etapa, la analítica puede ser bastante simple, pero se aprende mucho sobre cómo hacer que Hadoop funcione de la manera que desee.

Etapa 2: Transformación de la estructura y análisis muscular. La segunda etapa implica mejorar la capacidad de transformar y analizar datos. En esta etapa, las empresas buscan las herramientas que son más apropiadas para sus habilidades y comienzan a adquirir más datos y crear aplicaciones. Las capacidades del almacén de datos de la empresa y el lago de datos se utilizan juntas.

Etapa 3: Amplio impacto operacional. La tercera etapa implica obtener datos y análisis en manos de la mayor cantidad de gente posible. Es en esta etapa que el Data Lake y el almacén de datos de la empresa comienzan a funcionar al unísono, cada uno jugando su papel. Un ejemplo de la necesidad de esta combinación es el hecho de que casi todas las grandes compañías de datos que comenzaron con un Data Lake eventualmente agregaron un almacén de datos empresarial para operacionalizar sus datos. Del mismo modo, las empresas con almacenes de datos empresariales utilizan Hadoop.

Etapa 4: Capacidades de la empresa. En esta etapa más alta del Data Lake, las capacidades empresariales se agregan al Data Lake. Pocas compañías han alcanzado este nivel de madurez, pero muchas lo harán a medida que crezca el uso de big data, lo que requerirá gobernabilidad, cumplimiento, seguridad y auditoría.

P.5 ¿Por qué es importante la gestión de un Data Lake?

Los incrementos en potencia de procesamiento de computadoras, capacidad y uso de almacenamiento en la nube y conectividad de red están convirtiendo la corriente de datos en la mayoría de las empresas en un maremoto: un flujo interminable de información detallada sobre perfiles personales de clientes, datos de ventas, especificaciones de productos, pasos de proceso y así sucesivamente. Los datos son un activo corporativo valioso y su administración efectiva puede ser vital para el éxito de una organización (Keith, 2013). Un enfoque ágil para el desarrollo de un Data Lake puede ayudar a las empresas a lanzar programas analíticos rápidamente y establecer una cultura de datos amigable a largo plazo (Hagstroem, Roggendorf, Saleh, & Sharma, 2017). El análisis de los datos puede proporcionar información útil

para la detección de fraude en el sector financiero o la mejora de la experiencia del usuario (Powerdata, 2018) o detectar discrepancias en los patrones de voltaje (Lawson, 2016).

Para que un proyecto de Big Data tenga éxito, según (GarryKillian, 2016) se necesitan de dos cosas que son: saber qué datos accionables (combinados) necesita para los resultados deseados y obtener los datos correctos para analizarlos lograr esos resultados. Por otro lado, (Oram, s.f.) recomienda considerar los siguientes aspectos relacionados con la gestión de datos:

- Adquisición e ingestión: cómo resolver estos problemas con un grado de automatización.
- Metadatos: cómo hacer un seguimiento de cuándo llegaron los datos y cómo se formatearon, y cómo ponerlos a disposición en etapas posteriores del proceso.
- Preparación y limpieza de datos: lo que necesita saber antes de preparar y limpiar sus datos, y lo que debe limpiarse y cómo hacerlo.
- Organizar flujos de trabajo: lo que debe hacer para combinar sus tareas (ingestión, catalogación y preparación de datos) en un flujo de trabajo de extremo a extremo.
- Control de acceso: cómo abordar los controles de seguridad y acceso en todas las etapas del manejo de datos. Un Data Lake permite que los datos de toda una organización estén disponibles de manera inmediata y se utilicen solo aquellos requeridos para un determinado análisis. Esto facilita una toma de decisiones más acertada.

En resumen, la gestión de un Data Lake es muy importante para las estrategias de datos empresariales ya que responden mejor a las realidades de los datos actuales: volúmenes y variedades de datos mucho mayores, mayores expectativas de los usuarios y la rápida globalización de las economías.

P.6 ¿Cuáles son las funciones de un Data Lake?

Un Data Lake tiene las siguientes funciones:

Ingesta de datos

Las organizaciones tienen una serie de opciones cuando transfieren datos a un Hadoop Data Lake. La administración de la ingesta permite el control de los datos ingresados, de dónde provienen, cuándo llegan y dónde se almacena.

Gobernanza de los datos

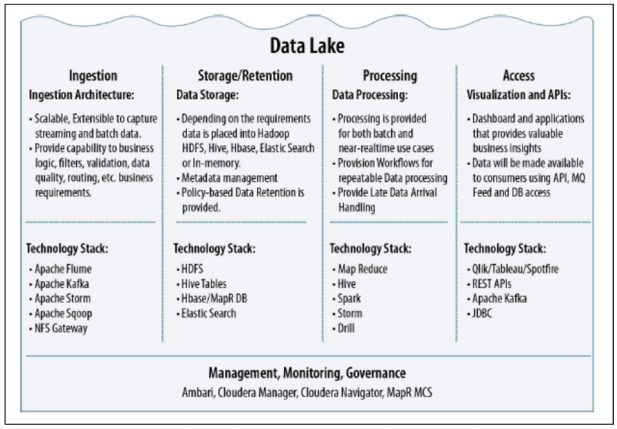
Una parte importante de la arquitectura de un Data Lake es colocar los datos en un área transitoria antes de moverlos al depósito de datos bruto.

Es en esta área donde todas las fuentes de datos posibles, externas o internas, o se mueven a Hadoop o se descartan. Al igual que con la visibilidad de los datos,

un proceso de administración de la ingesta impone reglas de gobernanza que se aplican a todos los datos que pueden ingresar al Data Lake.

Almacenamiento y retención de datos

Un Data Lake por definición proporciona un almacenamiento de datos mucho más rentables que un Datawarehouse, debido al modelo tradicional schema-on write, el cual es altamente ineficiente, incluso en la nube.



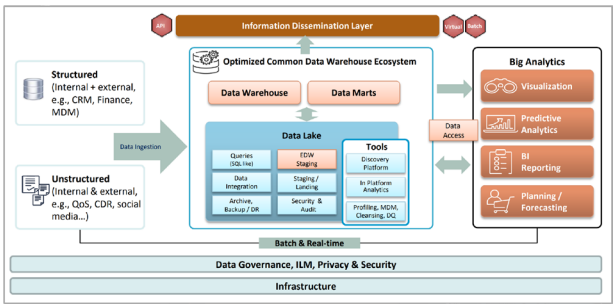
imagenes\funciones data lake.png
Figura 3
Funciones de un data lake
Fuente: (DATAFLOQ, 2016)

Procesamiento de datos

El procesamiento es la etapa en la que los datos se pueden transformar en formato estandarizado, esta etapa es necesaria porque en la etapa de ingesta de datos, el usuario no toma ninguna decisión sobre transformar o estandarizar los datos.

Acceso a los datos

Esta etapa es donde se consumen los datos del Data Lake. Existen varias formas de acceder a los datos: consultas, extracciones basadas en herramientas, o extracciones que deben suceder a través de una API. Algunas aplicaciones necesitan obtener los datos para realizar análisis u otras transformaciones en sentido descendente. (LaPlante & Sharma, 2016)



imagenes\proceso de almacenamiento de datos.png
Figura 4
Data Lake en acción
Fuente: (Capgemini, 2017)

P7. ¿En qué sectores es posible utilizar un Data Lake?

Un Data Lake genera valor en diferentes áreas. Los sectores que se benefician del uso de un Data Lake son:

Servicios de salud

Muchos grandes proveedores de servicios de salud mantienen millones de registros para millones de pacientes, incluidos informes semiestructurados como imágenes de radiología, notas no estructuradas de médicos y datos capturados en hojas de cálculo y otras aplicaciones informáticas comunes. Mediante la implementación de un Data Lake basado en una arquitectura Hadoop, un proveedor de servicios de salud puede habilitar el procesamiento distribuido de Big data, mediante el uso de estándares, permitiendo el almacenamiento de registro

en sus formatos nativos para su posterior análisis; esto evita el problema de forzar la categorización de cada tipo de datos, como sería el caso de un Datawarehouse tradicional.

Servicios financieros

Al pasar los datos a un Data Lake de Hadoop, los bancos pueden almacenar y analizar múltiples flujos de datos y ayudar a los gerentes regionales a controlar el riesgo de la cuenta en sucursales distribuidas. Son capaces de descubrir qué analistas de riesgo tomaron decisiones de cuenta que iban en contra de la información de riesgo por parte de terceros. El resultado neto es un mejor control del fraude. Con el tiempo, la acumulación de datos en un Data Lake permite al banco construir algoritmos que detecten patrones sutiles, pero de alto riesgo que los analistas de riesgo bancario pueden haber fallado en identificar previamente.

Otros aportes del análisis están enmarcados en el control de morosidad, la prevención de la fuga de cliente, estimación del nivel de renta, capacidad de ahorro, creación de nuevos servicios financieros, personalización de promociones y campañas de bonificación (Management Solutions, 2015).

Operaciones de vuelo

Las cancelaciones de vuelos, los retrasos en las salidas, la congestión en los tiempos de rodaje y las demoras de mantenimiento en el aire son problemas cada vez más frecuentes que afectan negativamente el rendimiento, el consumo de combustible, la tasa de emisiones y la satisfacción del cliente en los principales aeropuertos del mundo. Una solución prometedora para mejorar la eficiencia en las operaciones de vuelo es el uso de un Data Lake (Martínez-Prieto, y otros, 2017).

Por otro lado, el procesamiento de las opiniones de los pasajeros es un aspecto clave para determinar los sentimientos y medir el nivel de satisfacción de los pasajeros (Sankaranarayanan & Lalchandani, 2017).

En microempresas

Un Data Lake puede ayudar a las microempresas. Por ejemplo, los minoristas pueden almacenar todo el comportamiento de compra de un cliente en el Data Lake de Hadoop. Al capturar los datos de la sesión web (historiales de las sesiones de todos los usuarios en una página), las microempresas pueden hacer cosas como ofrecer ofertas oportunas basadas en la navegación web de un cliente y el historial de compras. (LaPlante & Sharma, 2016)

P8. ¿Cuáles son los aportes que ofrece esta tecnología a las organizaciones?

Los Data Lake son creados como un marco de gestión de datos integrado, elimina el costoso y engorroso proceso de preparación de datos de ETL que un Datawarehouse tradicional requiere. Los

datos se ingresan sin problemas en el data lake, donde se administran utilizando etiquetas de metadatos que ayudan a ubicar y conectar la información cuando los usuarios comerciales la necesiten.

Este enfoque libera a los analistas para la importante tarea de encontrar valor en los datos sin involucrar TI en cada paso del proceso, permitiéndoles conservar los recursos de TI. Hoy, todos los departamentos de TI están siendo obligados hacer más con menos. En tales entornos, los datos bien gobernados administrados, ayudan a las organizaciones a aprovechar de manera más efectiva todos sus datos para obtener información comercial y tomar buenas decisiones. (LaPlante & Sharma, 2016).

P9. ¿Cuáles son las ventajas de la gestión de un Data Lake?

Cada organización o industria puede beneficiarse básicamente de un Data Lake y usarlo según sus necesidades. Un caso ejemplar es su uso para finalizar los silos de datos dentro de su organización, centralizar los datos y obtener un mejor acceso a todas las fuentes de datos dispares dentro de su empresa.

Los casos de uso populares incluyen lograr vistas de 360 grados de los clientes y analizar las redes sociales, pero también permite a las organizaciones de salud optimizar los tratamientos y permite a los fabricantes obtener información de los datos de los sensores. Las ventajas de un Data Lake son numerosas:

Almacenamiento de bajo costo y extremadamente escalable

Los costos de almacenamiento de datos en un Data Lake son bajos y puede escalar fácilmente a volúmenes extremos (Simon, 2018).

Compatible con múltiples lenguajes de programación y marcos

Gracias a los datos brutos que se almacenan en el Data Lake, los desarrolladores pueden trabajar con múltiples lenguajes de programación como Python o Java y usar diferentes marcos como Hive o Pig.

Agnóstico de datos y acceso inmediato a todos los datos

Un Data Lake puede contener cualquier dato, desde datos de máquina estructurados hasta datos de redes sociales no estructurados en una ubicación central. Considerando que el 80% de los usuarios de un Data Lake son operativos, esta tecnología pone a su disposición la capacidad de generar informes, ver métricas clave de rendimiento o dividir el mismo conjunto de datos en una hoja de cálculo todos los días (Campbell, 2015)

Datos centralizados que no tienen que ser movidos

Con un Data Lake, todos los datos se encuentran en una ubicación central. Los silos ya no son necesarios, lo que facilita el acceso y la mezcla de las diferentes

fuentes de datos. Además, ya no es necesario mover los datos de un almacén a otro. resuelven problemas comerciales al permitir la democratización, reutilización, exploración y análisis de datos (Maroto, 2018).

Más información debido a los datos brutos

Con un Data Lake, las organizaciones pueden almacenar los datos en formato sin procesar, lo que significa que no se pierde información en el camino. En el futuro, a medida que surjan oportunidades adicionales para aprovechar los datos, las empresas pueden volver a los datos originales en busca de respuestas. (DATAFLOQ, 2016)

P10. ¿Cuáles son las desventajas de la gestión de un Data Lake?

Incluso una solución flexible de próxima generación como Data Lake está sujeta a su propio conjunto de desafíos. Aunque hay un gran volumen de datos disponible para los usuarios en Data Lake, los problemas pueden surgir cuando estos datos no se gestionan cuidadosamente:

- Falta de control de datos. Sin la estructura y los controles para administrar y mantener la calidad, consistencia y cumplimiento de los datos, un Data Lake puede convertirse rápidamente en un pantano.
- Mala accesibilidad. Aunque los datos pueden estar disponibles, su valor es limitado si los usuarios son incapaces de encontrar o entender los datos.
- Baja calidad de datos y linaje. Los usuarios necesitan conocer el contexto de los datos y saber de dónde provienen.
- Falta de seguridad de datos. Los datos cargados en una laguna de datos sin ningún tipo de supervisión pueden llevar a riesgos de incumplimiento.

Para maximizar el valor de un Data Lake y evitar estas dificultades, las organizaciones deben garantizar que sus implementaciones aborden factores de éxito críticos. (Knowledgent Group Inc., 2014)

4. DISCUSIÓN

En el presente trabajo se plantearon algunas preguntas de investigación que buscan determinar los aspectos teórico-conceptuales de un Data Lake. Para el caso de la respuesta a la pregunta planteada sobre su definición se tiene que Un Data lake o laguna de datos, es un repositorio digital construido para almacenar una gran cantidad de datos en formato nativo, es decir que los datos son ingresados en su sistema de procesamiento sin comprometer su estructura, esto se corrobora con lo que dice (Alserafi, 2016) al definir a un Data Lake como un repositorio masivo de datos en bruto.

Por otro lado, cuando se habla de su utilidad, considerando en primer lugar al Big Data y sus tecnologías como emergente, esto

se reafirma con el criterio de (Capgemini, 2017) y (Management Solutions, 2015).

Cuando se habla de su utilidad un Data Lake permite la ingesta de datos de diferentes fuentes: redes sociales, internet de las cosas, base de datos relacional, datawarehouse tradicional, dispositivos móviles, registros de llamadas telefónicas, logs de servidor, etc, convirtiendo a un Data Lake en un repositorio que almacena toda la información de una organización, esto lo ratifica (Khine & Zhao, 2017). Un Data Lake puede ser utilizado para realizar un análisis dinámico de datos, que genere un nuevo conocimiento que aporte a la toma de decisiones en la organización, también puede ser utilizado como fuente de datos para un Datawarehouse tradicional, como soporte para otras aplicaciones, para la explotación de datos en línea, entre otros, esto lo ratifican (SQL Chick, 2016) (Capgemini, 2017) (Alserafi, 2016) (Management Solutions, 2015).

A diferencia de un Datawarehouse tradicional el cual utiliza un ETL/ELT para alimentar su repositorio, un Data Lake emplea una arquitectura Lambda para la ingesta de datos (LaPlante & Sharma, 2016). Los sectores que se han beneficiado de la implementación de un Data Lake son los proveedores de servicios de salud, las organizaciones financieras y las microempresas esto lo ratifican (LaPlante & Sharma, 2016) (Management Solutions, 2015).

Las respuestas en relación con las preguntas planteadas sobre ventajas y desventajas ratifican lo que mencionan (DATAFLOQ, 2016) y (Knowledgent Group Inc., 2014) al considerar que un Data Lake permite un almacenamiento de bajo costo y extremadamente escalable, es compatible con múltiples lenguajes de programación y marcos, es de acceso inmediato a todos los datos y almacena todo tipo de datos, permitiendo contar con datos centralizados que no tienen que ser movidos. Sin embargo, esta tecnología aun requiere un mayor control cuando existe un incremento en el volumen de los datos, mejoras en su accesibilidad, calidad y linaje, además de seguridad.

La mayoría de información sobre un data lake se obtiene de artículos técnicos que las organizaciones generan, más que de información de artículos científicos, esto demuestra el interés de las organizaciones por explotar esta tecnología, este criterio lo ratifica (Fang, 2015)

Un Data lake al ser una nueva tecnología demanda de innovación para permitir el análisis de los datos, de tal manera que realmente permita la generación de información útil para la toma de decisiones. Algunos autores debido a esta demanda no le ven un futuro prometedor a esta tecnología, sin embargo, otros lo ven como una oportunidad de explotación y desarrollo, plantean soluciones como el

uso de inteligencia artificial, el contexto como servicio y módulos que permitan el análisis de la información. .

4. DISCUSIÓN

Un Data lake o laguna de datos, es un repositorio digital construido para almacenar una gran cantidad de datos en formato nativo. Los datos que alimentan un Data Lake provienen de diferentes fuentes como son: Redes Sociales: videos, imágenes, audios, Bases de datos relacionales, Sistemas de Información Gerencial: CRM, SCM y ERP, logs de servidores, correo electrónico, Dispositivos móviles y datos en especial del Internet de las cosas.

Además, un Data Lake es utilizado para realizar un análisis experimental de datos o el análisis en tiempo real, permite extraer conocimiento que contribuya a la toma de decisiones. Otra de las utilidades es que puede ser utilizado como repositorio de soporte para otras aplicaciones. Un Data Lake permite realizar consultas federadas y tiene un procesamiento distribuido. Los sectores que se han beneficiado de esta tecnología son los Proveedores de Servicios de Salud, las entidades que brindan servicios bancarios, las aerolíneas y las microempresas. Las funciones que tiene un Data Lake son: Ingesta, Almacenamiento, Procesamiento, Acceso. Aún hay mucho por descubrir sobre esta tecnología emergente, investigaciones futuras pueden realizarse en función de los resultados que tiene su implementación, en cómo mejorar la calidad de los datos que almacenan, en cómo se ampliar o mejorar el procesamiento de los datos, en las técnicas de análisis, en cómo gestionar las consultas federadas, además de ampliar la información sobre la arquitectura lamba y otras arquitecturas que mejorarán la ingesta de datos.

5. FUENTES BIBLIOGRÁFICAS

1. Knowledgent Group Inc. (2014). How to Design a Successful Data Lake. KNOWLEDGENT WHITE PAPER, 1-13.
2. Alserafi, A. (2016). Towards Information Profiling: Data Lake Content Metadata Management. 16th International Conference on Data Mining Workshops (ICDMW), 178-185.
3. Campbell, C. (26 de 01 de 2015). Blue Granite. Obtenido de Top Five Differences Between Data Lakes And Data Warehouses: https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses
4. Capgemini. (2017). Obtenido de BigData y Ecosistema Hadoop: Capgemini - UV: https://www.uv.es/capgeminiuv/documents/BigData_UV_20170328_v3.pdf
5. Cito Research. (2014). Obtenido de Putting the Data Lake to Work A Guide

to Best Practices: https://citoresearch.com/
6. DATAFLOQ. (23 de Noviembre de 2016). The Future of Big Data: How Data Lakes Open New Possibilities for Your Organization. Obtenido de DATAFLOQ: https://datafloq.com/read/Data-Lakes-Open-Possibilities-Your-Organization/1695
7. Fang, H. (2015). Managing Data Lakes in Big Data Era What's a data lake and why has it became popular in data management ecosystem. Cyber Technology in Automation, Control and Intelligent Systems, 820 - 824.
8. Farrugia, A., Claxton, R., & Thompson, S. (2016). Towards social network analytics for understanding and managing enterprise data lakes. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). San Francisco. Obtenido de https://ieeexplore.ieee.org/document/7752393/
9. GarryKillian. (2016). I-SCOOP. Obtenido de Data lakes and big data analytics: the what, why and how of data lakes: https://www.i-scoop.eu/big-data-action-value-context/data-lakes/
10. Hagstroem, M., Roggendorf, M., Saleh, T., & Sharma, J. (Agosto de 2017). A smarter way to jump into data lakes. Obtenido de McKinsey & Company : https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/a-smarter-way-to-jump-into-data-lakes
11. Keith, G. (2013). Data and the Enterprise. En G. Keith, Principles of Data Management (págs. 1-10). Edinburgo: The British Computer Society. Obtenido de https://www.bcs.org/upload/pdf/data-management-chapter1.pdf
12. Khine, P. P., & Zhao, S. W. (2017). Data Lake: A New Ideology in Big Data Era. ResearchGate.
13. Kitcheham, B. (2011). Using mapping studies as the basis for further research - A participant-observer case study. Information and Software Technology, 53(6), 638-651.
14. LaPlante, A., & Sharma, B. (2016). Architecting Data Lakes. Estados Unidos: O'Reilly Media, Inc.
15. Lawson, L. (5 de July de 2016). 4 Best Practices for Data Lakes. Obtenido de Enterprise Apps Today: http://www.enterpriseappstoday.com/data-management/best-practices-data-lakes.html
16. Management Solutions. (2015). Obtenido de Data Science y la transformación del sector financiero: https://www.managementsolutions.com/sites/default/files/publicaciones/esp/Data-Science.pdf
17. Maroto, C. (2018). A Data Lake Architecture with Hadoop and Open Source Search Engines. Obtenido de Search Technologies: https://www.searchtechnologies.com/blog/search-data-lake-with-big-data

18. Martínez-Prieto, M. A., Bregon, A., García-Miranda, I., Álvarez-Esteban, P. C., Díaz, F., & Scarlatti, D. (2017). Integrating Flight-related Information into a (Big) Data Lake. Conferencia de Sistemas de Aviónica Digital (DASC), IEEE / AIAA 2017 (págs. 1 - 10). St. Petersburg, FL, US: IEEE.
19. Miloslavskaya, N., & Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016, 300-305.
20. Oram, A. (s.f.). Obtenido de Managing the Data Lake: https://www.oreilly.com/data/free/managing-the-data-lake.csp
21. Petersen, K. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 1-18.
22. POWER DATA. (06 de Agosto de 2014). POWER DATA. Obtenido de POWER DATA: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/392573/10-se-ales-de-que-su-empresa-tiene-problemas-de-gesti-n-de-datos
23. Powerdata. (2018). Obtenido de El valor de la gestión de los datos: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/topic/data-lake
24. Salemink, I. (15 de Noviembre de 2017). Dutch Enterprise Data Lake, fishing in clear water. Obtenido de unstats.un.org: https://unstats.un.org/unsd/bigdata/conferences/2017/presentations/day3/session1/p-sessionB/1%20-%20Data%20Lake%20-%20Irene%20Salemink%20-%20Statistics%20Netherlands%20-%2016-9_FINAL.pdf
25. Sankaranarayanan, H. B., & Lalchandani, J. (2017). Passenger reviews reference architecture using big data lakes. 7ma Conferencia Internacional sobre Cloud Computing, Data Science & Engineering - Confluence, 2017 , 204 - 209.
26. Simon, P. (2018). Data lake and data warehouse – know the difference. Obtenido de SAS: https://www.sas.com/en_us/insights/articles/data-management/data-lake-and-data-warehouse-know-the-difference.html#/
27. Souag, A., Mazo, R., Salinesi, C., & Comyn-Wattiau, I. (2015). Reusable knowledge in security requirements engineering: a systematic mapping study. Springer-Verlag(21), 251-283.
28. SQL Chick. (2 de Octubre de 2016). Data Lake Use Cases and Planning Considerations. Obtenido de SQL Chick: https://www.sqlchick.com/entries/2016/7/31/data-lake-use-cases-and-planning
29. Stodder, D. (23 de 05 de 2017). TDWI. Obtenido de Managing-the-data-lake-monster: https://tdwi.org/articles/2017/05/23/managing-the-data-lake-monster.aspx