BENEFICIOS Y RIESGOS DE DATA LAKE

Integrantes:

- Diego Porlles
- Paolo Lizárraga
- Anthony Belizario
- Daniel Loja

Abstract

A data lake is a centralized storage repository that contains big data from various sources in a raw, granular format. You can save structured, semi-structured, or unstructured data, which means that the data can be kept in a more flexible format for future use. When saving data, a data lake associates it with metadata tags and identifiers so that it can be extracted more quickly.

Definición:

Un data lake es un repositorio de almacenamiento centralizado que contiene big data de varias fuentes en un formato granular y sin procesar. Puede guardar datos estructurados, semiestructurados o no estructurados, lo que significa que los datos pueden conservarse en un formato más flexible para usarlos en un futuro. Al guardar datos, un data lake los asocia con identificadores y etiquetas de metadatos para poder extraerlos más rápidamente.

Características de Data Lake

Incluye funcionalidades de orquestación y programación de trabajos (por ejemplo, a través de YARN). La ejecución de la carga de trabajo es un requisito previo para Hadoop empresarial.

YARN proporciona administración de recursos y una plataforma central para entregar herramientas consistentes de operaciones, seguridad y control de datos en los clústeres de Hadoop, asegurando que los flujos de trabajo analíticos tengan acceso a los datos y la potencia informática que requieren.

Ventajas de Data Lake

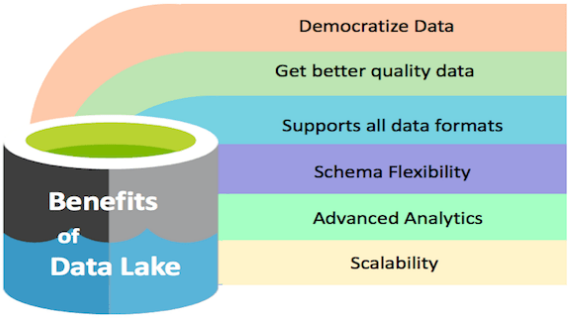
Un data lake funciona a partir de un principio llamado schema-on-read o esquema contra escritura. Esto significa que no existe un esquema predefinido en el que deban encajarse los datos antes de almacenarlos. Tan solo cuando los datos se leen durante el tratamiento se analizan y adaptan en un esquema según convenga. Esta característica ahorra mucho tiempo que normalmente se dedica a la definición del esquema.

Esto también permite almacenar datos tal y como estén, en cualquier formato.

Data Lake frente a almacenes de Datos

Un data lake y un almacén de datos son semejantes en su finalidad y objetivo fundamentales, motivo por el que se confunden fácilmente:

- Ambos son repositorios de almacenamiento que consolidan los distintos depósitos de datos de una organización.
- El objetivo de ambos es crear un depósito de datos único que alimente distintas aplicaciones.



Entonces, ¿Por qué un Data Lake?

Porque ayuda a resolver el problema persistente de accesibilidad e integración de datos. Mediante el uso de infraestructuras de big data, las empresas están comenzando a reunir volúmenes de datos crecientes para análisis o simplemente para almacenarlos para uso futuro indeterminado.

¿Cuáles son los siguientes pasos? ¿Cómo es analizada la información?

En un data lake no todo es blanco o negro, está claro que habrá una capa de raw data, pero también puede haber otras capas donde los datos comiencen a ser depurados y preparados para su explotación.

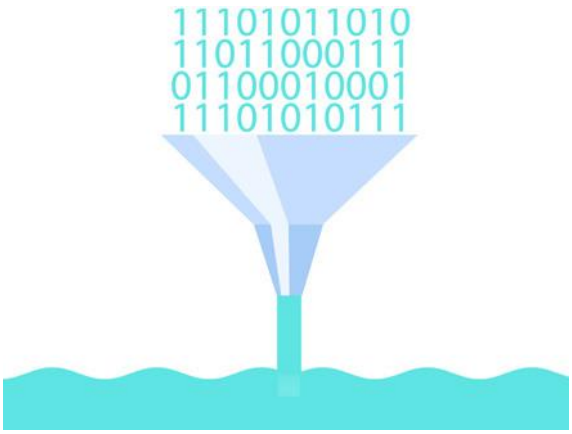
En este sentido, la tecnología normalmente usada para construir este tipo de repositorios es HDFS (Hadoop Distributed File System), un sistema de ficheros distribuido en el que podemos crear diferentes capas a través de las cuales los datos vayan pasando determinados controles de calidad y adaptación a las distintas necesidades de negocio. Sin embargo, se pueden emplear otras tecnologías o una combinación de ellas en función de las distintas capas que se quieran desarrollar.

Las capas más comunes que se pueden encontrar en un data lake son las siguientes:

Data Ingestion: Es una capa temporal de carga en la que los datos pasan checks básicos antes de ser almacenados en la capa de raw data. Si bien no es necesaria, se puede implementar para llevar a cabo:

- Controles básicos de calidad, como posibles filtros según el origen de los datos, descartando fuentes desconocidas.
- Procesos de encriptación de los datos en caso de requerirse por motivos de seguridad.
- Registros sencillos de metadata y trazabilidad mediante tags,

almacenando el origen de los datos, fecha y hora de carga, el formato y otras características técnicas, su privacidad y nivel de seguridad, algoritmo de encriptación, etc.



Data Storage (Raw Data): Es una capa sin esquema establecido donde todos los datos, estructurados o no estructurados, son almacenados sin sufrir adaptaciones. Es una capa en la que se precisa de analistas expertos en data discovery mediante herramientas big data (Hive, Spark, Map Reduce…).

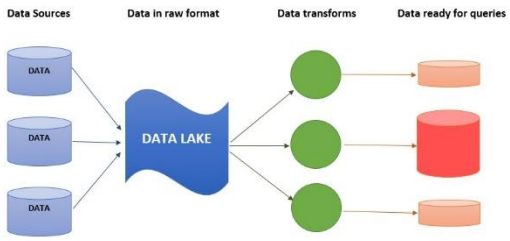
Data Processing (Zona de Confianza): Una vez que los analistas de datos han realizado data discovery en el raw data, se puede ver la necesidad de procesar y adaptar determinados sets de datos para alojarlos en una capa de uso recurrente. En esta capa pueden tener lugar procesos avanzados de data quality, integridad y otras adaptaciones para disponer de una capa de confianza de exploración de datos a la que tengan acceso otros usuarios.

Data Access (Zona de Consumo): Esta es una capa más avanzada donde, finalmente, los datos se ponen a disposición de analistas de negocio. Estos analistas podrán generar informes y análisis para responder a preguntas de negocio y afianzar la toma de decisiones.

Siendo estas capas opcionales, cada organización puede definir su propia estrategia a la hora de implantar un data lake. Sin embargo, al margen del número de capas que se opte por implantar, siempre existe una estrategia común básica: almacenar todos los datos de la organización con independencia de que tengan uso en la actualidad o no, dejándolos disponibles para posibles futuras necesidades no detectadas en el momento de desarrollar el data lake.

De esta manera, a medida que surjan nuevas necesidades de negocio los datos irán pasando de la capa raw data a capas más avanzadas, dejando nuevos sets de datos a disposición de los usuarios de negocio.

Es por ello, que un data lake es un repositorio de datos vivo y en constante evolución, adaptado para escalar de forma flexible en función del negocio y su propia evolución.



¿Cuáles son los beneficios de una data lake?

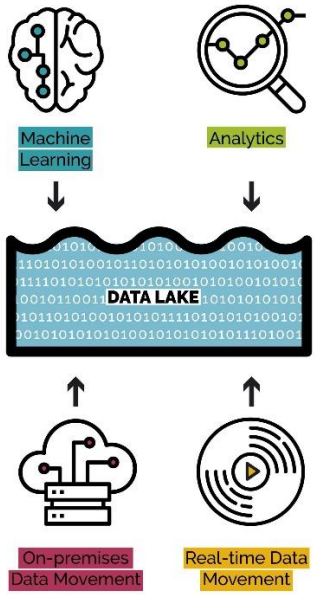
El principal beneficio de un data lake es la centralización de fuentes de contenido dispares. Una vez reunidas (de sus "silos de información"), estas fuentes pueden ser combinadas y procesadas utilizando big data, búsquedas y análisis que de otro modo hubieran sido imposibles. Las fuentes de contenido dispares a menudo contienen información confidencial que requerirá la implementación de las medidas de seguridad apropiadas en el data lake.

Las medidas de seguridad en el data lake pueden ser asignadas de manera que se otorga acceso a cierta información a los usuarios del data lake que no tienen acceso a la fuente de contenido original. Estos usuarios tienen derecho a la información, pero no pueden acceder a ella en su fuente por alguna razón.

Es posible que algunos usuarios no necesiten trabajar con los datos en el origen de contenido original, sino consumir los datos resultantes de los procesos incorporados a dichos orígenes. Puede haber un límite de licencias para el origen de contenido original que impide que algunos usuarios obtengan sus propias credenciales. En algunos casos, la fuente de contenido original se ha bloqueado, está obsoleta o se desactivará en breve, sin embargo, su contenido sigue siendo valioso para los usuarios del data lake.

Una vez que el contenido está en el data lake, puede normalizarse y enriquecerse. Esto puede incluir extracción de metadatos, conversión de formatos, aumento, extracción de entidades, reticulación, agregación, des-normalización o indexación.

Los datos se preparan "según sea necesario", lo que reduce los costos de preparación sobre el procesamiento inicial (tal como sería requerido por los data warehouses. Una estructura de big data permite escalar este procesamiento para incluir los conjuntos de datos más grandes posibles.



Los usuarios, de diferentes departamentos, potencialmente dispersos por todo el mundo, pueden tener acceso flexible a un data lake y a su contenido desde cualquier lugar. Esto aumenta la reutilización del contenido y ayuda a la organización a recopilar más fácilmente los datos necesarios para impulsar las decisiones empresariales.

La información es poder, y un data lake pone la información de toda la empresa

en manos de muchos más empleados para hacer a la organización un todo más inteligente, más ágil y más innovadora.

Maneras de Mejorar la calidad en Data Lake

1. Uso de Machine Learning y NLP.

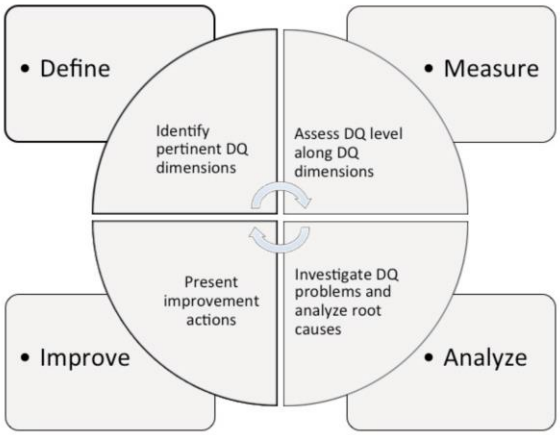
Machine Learning puede cambiar el juego porque puede capturar el conocimiento tácito de las personas que mejor conocen los datos, y luego convertirlos en algoritmos, que se pueden usar para automatizar el procesamiento de datos a gran escala. Esta es exactamente la forma en que Talend está aprovechando el aprendizaje automático de Spark, para aprender de los administradores de datos durante la comparación de datos y la deduplicación de las muestras de datos, y luego aplicarlo a gran escala de datos para miles de millones de registros.

2. Establecer los estándares de calidad de datos ágiles.

Para que las empresas aprovechen al máximo sus proyectos de transformación digital y creen un lago de datos ágil, deben diseñar procesos de calidad de datos desde el principio. Las organizaciones deben centrarse en estandarizar lo siguiente para mantener la calidad de big data.

3. Emplear marcos de gestión de calidad de datos.

Otra categoría de marcos se centra en la madurez de los procesos de gestión de la calidad de los datos. Su objetivo es evaluar el nivel de madurez de la gestión de DQ para comprender las mejores prácticas en organizaciones maduras e identificar áreas de mejora. Los ejemplos populares de dichos marcos incluyen la Gestión de la calidad de los datos total (TDQM), la Integración del modelo de madurez de la capacidad (CMMI), los Objetivos de control para la información y la tecnología relacionada (CobiT), la Biblioteca de infraestructura de tecnología de la información (ITIL) y Six Sigma.



PROCESO DE CREACIÓN DE UN DATA LAKE

Adquisición de Datos:

Obtención de datos y metadatos, así como su preparación para una eventual inclusión en el data lake. Este proceso, identificar las fuentes y los conjuntos de datos que son de mayor valor para un determinado proceso, puede ser muy exigente. Consiste en determinar, que datos, con qué granularidad (nivel de detalle), cuál es la frecuencia con la qué se pueden obtener o si se pueden leer de una vez, etc. Para realizarse bien, hay que tener un conocimiento adecuado del uso que se quiere dar a los datos de cara a anticipar las necesidades de los usuarios.

Data Curation/Grooming Data:

Es el conjunto de procesos/pasos por los que los datos crudos son transformados en datos consumibles por las aplicaciones analíticas. Para ello toman en consideración los metadatos obtenidos en el paso anterior y aplican transformaciones a los datos para que puedan ser utilizados en un punto anterior. Son cosas tan “sencillas” como transformar un fichero csv en una matriz o, más complejas, como capturar el contenido de una hoja de cálculo compleja u obtener el texto de un PDF y categorizarlo o, todavía más complejas, como integrar conjuntos de datos. Otros procesos asociados podrían ser la normalización de datos o la generación de datos derivados aplicando técnicas de inteligencia artificial. Un aspecto importante a considerar es la necesidad de guardar el proceso seguido especificándolo todo lo posible tanto para los pasos como para los datos (versiones, formatos, etc.).

Provisión de Datos:

Son el conjunto de procesos que permiten acceder a los datos contenidos en el data lake de acuerdo con las políticas que tiene establecidas. Para evitar el acceso a datos inapropiados, el data lake debería proveer un modo de visualización de conjuntos de datos que permitiese determinar (por sus excepciones, contenido, etc.) su adecuación a un determinado fin. Por lo general, junto con los datos debería poder visualizarse la metainformación de los mismos, incluyendo el contexto de los datos, de cara a que el usuario pueda entenderlo y utilizarlos adecuadamente en sus análisis. Finalmente, una vez el usuario ha seleccionado el conjunto de datos de su interés, antes de que pueda ser extraído sacado del data lake debería informársele de las políticas y licencias que atañen a su uso.

Preservación de los Datos:

Son el conjunto de procesos y políticas que determinan qué datos deben conservarse, hasta cuándo y cuáles no. Otros objetivos de estos procesos es determinar cómo debe evolucionar la infraestructura para garantizar la disponibilidad de suficiente espacio y el rendimiento adecuado para acceder a los datos.

CONCLUSIONES

Data Lake no es un Datawarehouse. Ambos están optimizados para diferentes propósitos, y el objetivo es utilizar cada uno para lo que fueron diseñados y no malgastar recursos en algo que no tendrá uso en la organización.

Recuerda, tener centralizada la información es el camino, pero el objetivo es hacer que la información esté disponible y a eso se puede llegar por muchos caminos.

Bibliografía:

https://www.itainnova.es/blog/big-data-y-sistemas-cognitivos/que-es-data-lake-definicion-creacion-y-ejemplos/

https://www.powerdata.es/data-lake

http://websinergia.com.mx/blog/2016/05/30/que-es-un-data-lake/

https://www.evaluandocrm.com/la-inestabilidad-del-big-data-latam-los-riesgos-toda-empresa-conocer/

https://www.applying.pe/data-lake-mejorar-la-calidad-data-lake/#:~:text=El%20desaf%C3%ADo%20con%20Data%20Lake&text=Finalmente%2C%20uno%20de%20los%20mayores,reglamentarios%20que%20otros%20datos%20no.