
Universidad
Tecmilenio®

Universidad Tecmilenio

Infraestructura para Big Data

Data lake y data storage

Semana 12

Introducción

Un *data lake* (lago de datos) es un repositorio que contiene una gran cantidad de datos crudos, con lo cual se aprovecha todo el potencial de Big Data. Por lo tanto, las tomas de decisiones (que tienen como eje central el análisis de datos) permiten a las organizaciones obtener una ventaja competitiva para desenvolverse mejor en su mercado.

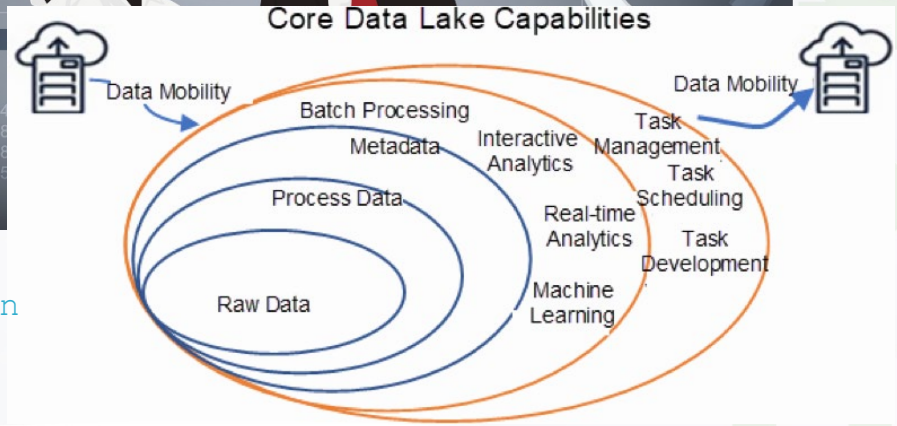
Aunque este concepto es bastante nuevo, Gorelik (s.f.) indica que ha tenido algunas etapas de madurez.

Explicación

Estas son algunas de las características más sobresalientes de un data lake (Software Testing Help, 2021):

- Puede almacenar diversos tipos de datos, por ejemplo, estructurados, semiestructurados y no estructurados.
- Permite varias capacidades de análisis, por ejemplo, aprendizaje automático, análisis predictivo, entre otros.
- Proporciona almacenamiento suficiente para todos los datos de una organización a bajo costo.
- Proporciona metadatos completos para administrar los elementos relacionados con los datos.

Capacidades básicas de un data lake:



AsparaDB. (2020). *Data Lake: Concepts, Characteristics, Architecture, and Case Studies*. Recuperado de https://www.alibabacloud.com/blog/data-lake-concepts-characteristics-architecture-and-case-studies_596910

Explicación

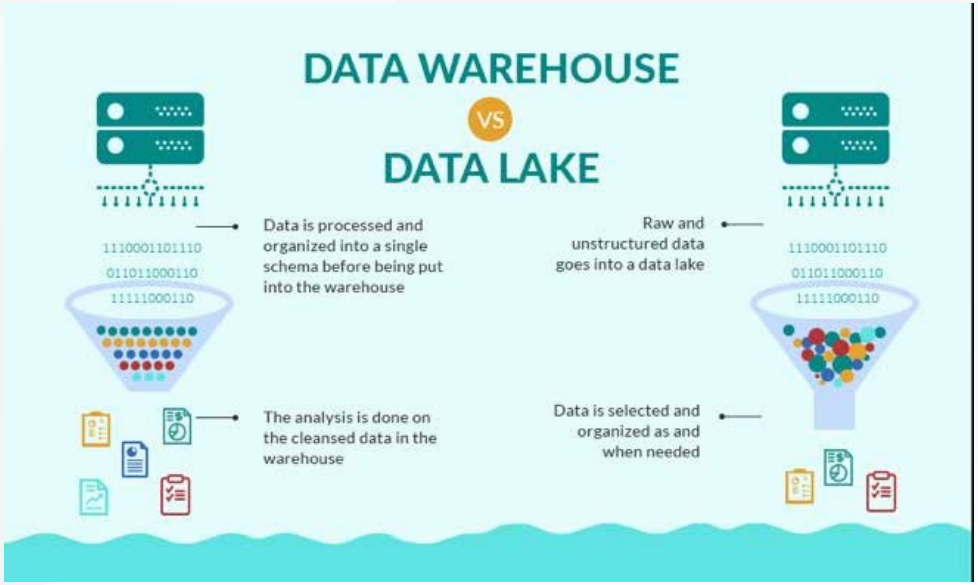
Explicación

De acuerdo con Power Data (s.f.), algunas de las diferencias entre el data warehouse y el

1. Conservación de datos.
2. Tipos de datos.
3. Usuarios.
4. Adaptabilidad.

Data warehouse

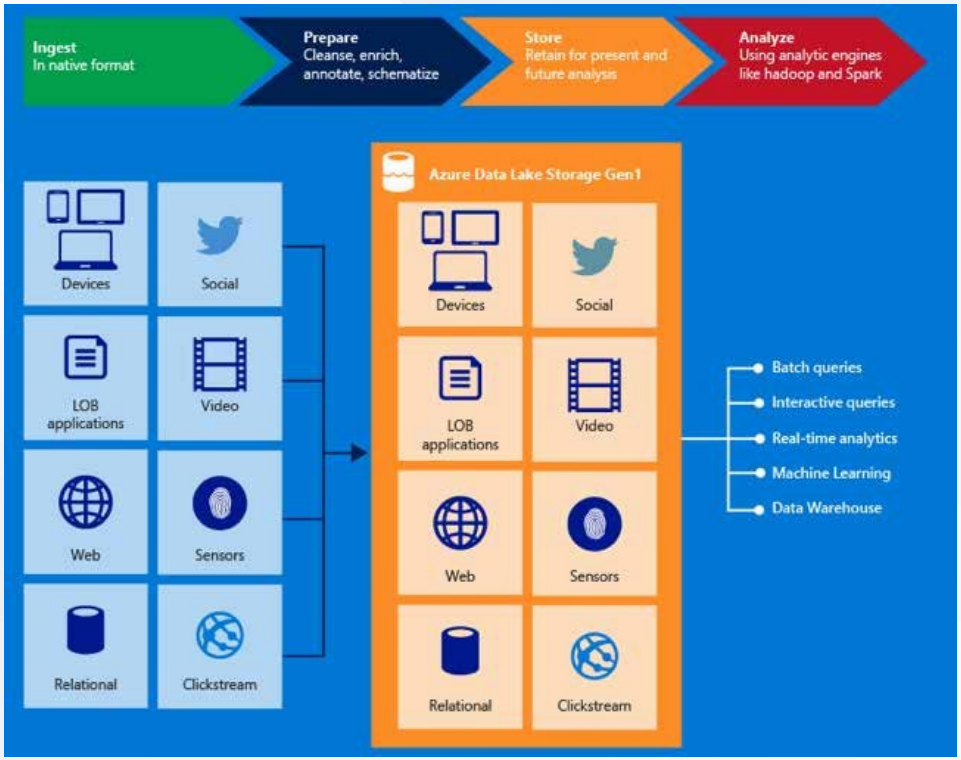
Es un sistema que permite incorporar y guardar información de diferentes fuentes en un solo almacén de datos (centralizado), permitiendo a una organización realizar potentes análisis para grandes volúmenes de información.



Dewan, S. (2018). *Data Warehouse Data Lake vs Data Warehouse: Which one should you go for?* Recuperado de https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/

Explicación

Azure Data Lake Storage Gen 1



Microsoft. (2021). *¿Qué es Azure Data Lake Storage Gen1?* Recuperado de https://docs.microsoft.com/es-es/azure/data-lake-store/data-lake-store-overview

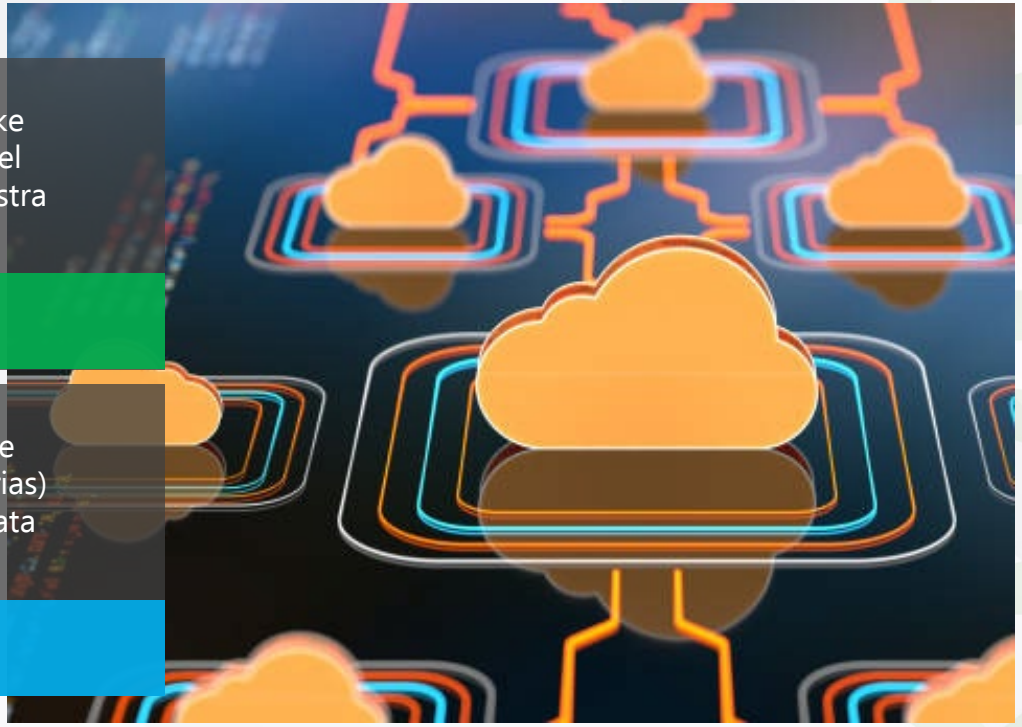
En **Microsoft Azure**, un data lake se configura para el almacenamiento de objetos en Hadoop, esto es, cuando levantamos nuestro clúster en HDInsight. Los datos de una organización o empresa se cargan primero por medio de HDInsight, creando tablas o cargando archivos con Hive. Posteriormente se aplican las herramientas de análisis y de procesamiento de datos.

Azure Data Lake Storage Gen1 es un sistema de archivos Hadoop Apache que es compatible con el sistema distribuido de Hadoop (HDFS), el cual funciona de manera natural en el ecosistema Hadoop y en todas sus distribuciones (Microsoft, 2021). Por lo tanto, es fácil analizar los datos almacenados en Data Lake Storage Gen1 usando
frameworks de analítica de Hadoop como MapReduce o Hive.

Explicación

Es importante destacar que, para crear un data lake storage, debemos seleccionarlo en los servicios del portal de Azure, asignarle un nombre y ligarlo a nuestra suscripción activa.

Posteriormente, se le deberá asignar una llave de identificación (se deben crear las que sean necesarias) para otorgar privilegios de lectura y escritura al data lake.



Para conocer un ejemplo, revisa el siguiente video:

Dirección de Producción de Contenidos. (2021, 27 de enero). *Big Data Tema 9* [Archivo de video]. Recuperado de https://youtu.be/vwwbnZMPWNE

Para conocer más sobre data lakes, revisa el siguiente video:

IBM Cloud. (2019, 19 de junio). *What is a Data Lake?* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=LxcH6z8TFpI&ab_channel=IBMCloud

Cierre

Es importante recordar que un data lake storage puede almacenar cualquier dato en su formato nativo (a diferencia del *data storage* convencional) sin requerir ninguna transformación previa. Por lo tanto, permite un análisis detallado a través de distintos *frameworks* y técnicas.

Por consiguiente, es una herramienta flexible para una empresa o industria con procesos que generan diversos tipos y tamaños de datos.

Bibliografía

- AsparaDB. (2020). *Data Lake: Concepts, Characteristics, Architecture, and Case Studies.* Recuperado de https://www.alibabacloud.com/blog/data-lake-concepts-characteristics-architecture-and- case-studies_596910
- Dewan, S. (2018). *Data Warehouse Data Lake vs Data Warehouse: Which one should you go for?* Recuperado de https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/
- Gorelik, A. (s.f.). *The Enterprise Big Data Lake.* Recuperado de https://www.oreilly.com/library/view/the-enterprise-big/9781491931547/ch01.html
- Microsoft. (2021). *¿Qué es Azure Data Lake Storage Gen1?* Recuperado de https://docs.microsoft.com/es-es/azure/data-lake-store/data-lake-store-overview
- Power Data. (s.f.). *Data lake: definición, conceptos clave y mejores prácticas.* Recuperado de https://www.powerdata.es/data-lake
- Software Testing Help. (2021). *What Is A Data Lake | Data Warehouse Vs Data Lake.* Recuperado de https://www.softwaretestinghelp.com/what-is-a-data-lake/

Universidad Tecmilenio

Infraestructura para Big Data

Introducción a la transformación y a la manipulación de datos

Semana 12

Introducción

En este tema se hablará sobre la manipulación y la transformación de datos. Por un lado, la primera se refiere a las acciones de trabajar con los datos, leerlos, guardarlos, copiarlos, aumentarlos, etc. Mientras que la segunda se refiere al "reajuste" o cambio en alguna de las propiedades inherentes del dato, como su tipo o tamaño.

Por lo tanto, verás cómo realizar estas acciones mediante Apache Spark.

Explicación

Características de la transformación de datos

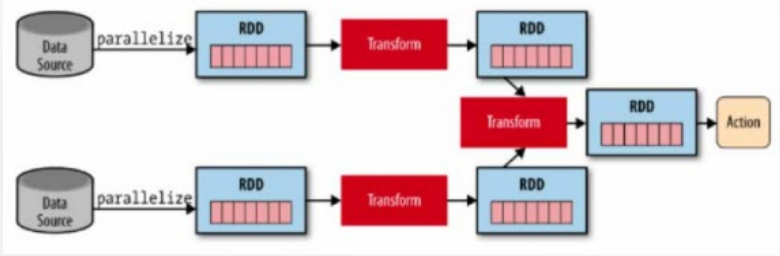
La transformación de datos es un proceso muy común dentro de los mundos de *machine learning*, ciencia de datos e inteligencia artificial. Es el proceso de convertir los datos de un formato hacia otro, es decir, del formato fuente (de origen) al formato que se requiere dentro del sistema destinatario. Durante este proceso, posiblemente sea necesario hacer otras acciones, por ejemplo, el refinamiento o combinación de datos. Este proceso recibe el nombre de ETL (*extract, load, transform*) (IBM Cloud Education, 2020):

Entre los beneficios que existen en la transformación de datos, se pueden señalar los siguientes (Talend, s.f.):

1. Alcanzar la mejor utilidad de los datos.
2. Administrar los datos de forma eficiente.
3. Desarrollar consultas de forma más rápida.
4. Aumentar la calidad de los datos.

Explicación

Spark tiene como fundamento y predecesor a MapReduce (dando un marco para escribir funciones que permiten procesar grandes cantidades de datos en paralelo, mediante grandes grupos de hardware de forma estable) (Mulchandani, 2020), pero mantiene algunas bondades, como la escalabilidad lineal y la tolerancia a fallos, e incorpora las siguientes:

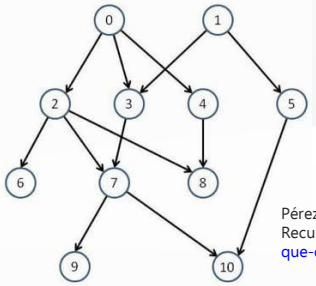


Kadam, G. (2018). *Beneath RDD (Resilient Distributed Dataset) in Apache Spark.* Recuperado de https://medium.com/@gkadam2011/beneath-rdd-resilient-distributed-dataset-in-apache-spark-260c0b7250c6

DAG (Directed Acyclic Graph): tipo de grafo que puede representar una serie de datos relacionados, los cuales se simbolizan mediante nodos y cada uno de estos representa un cierto conjunto de datos dentro de todo el grupo. Los nodos se conectan mediante líneas.

Este tipo de grafo no cuenta con ciclos, por tanto, no hay un camino directo que empiece y termine en un mismo nodo. En consecuencia, un vértice se puede conectar con otro, pero no consigo mismo (Maldonado, 2020).

RDD (Resilient Distributed Datasets): posibilita el desarrollo de operaciones con grandes volúmenes de datos en clúster de forma rápida. Se pueden realizar dos tipos de operaciones: las transformaciones y las acciones.

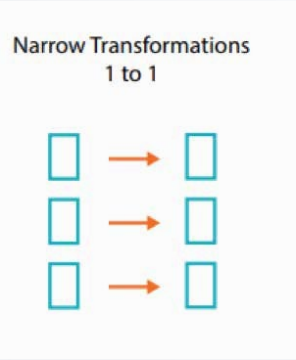


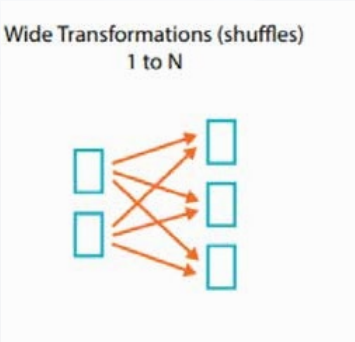
Pérez, M. (2016). *Apache Spark: qué es y cómo funciona.* Recuperado dehttps://geekytheory.com/apache-spark-que-es-y-como-funciona

Explicación

Una **transformación** es un tipo de operación, en la cual los datos RDD se transformarán de un formulario a otro de Spark. Por ende, cuando este proceso se realice, se obtendrá un nuevo RDD con datos transformados (en Spark los RDD son inmutables).

Según Omar (2020), Spark define dos tipos de operaciones de transformación: *narrow* y *wide transformation*.





Databricks. (s.f.). *Transformations*. Recuperado de https://databricks.com/glossary/what-are-transformations

Narrow transformation

Las transformaciones se ubican en la partición que les corresponde. Algunos ejemplos de esta transformación son las funciones filter(), sample(), map() o flatMap().

Las transformaciones combinan datos de diferentes particiones. Algunos ejemplos de esta transformación son groupByKey() o reduceByKey().

Wide transformation

Explicación

La **manipulación de datos** consiste en ajustarlos y alterarlos con la finalidad de organizarlos y facilitar su lectura. Entre los principales beneficios para las organizaciones se pueden señalar los siguientes (Srinivasan, 2020):

- Creación de valor mediante los datos.
- Proyección de los datos.
- Coherencia de los datos.
- Eliminación de datos innecesarios.

Para conocer un ejemplo de uso de Spark en Azure, ve el siguiente video:

Dirección de Producción de Contenidos. (2021, 3 de febrero). *Tema10* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=aun-vI2qkxw

Cierre

A través de este tema se enfatizó en la importancia de las herramientas que existen para la transformación y manipulación de datos, las cuales se utilizan en organizaciones de diversos giros, por ejemplo, el financiero, el hospitalario, las ciencias ambientales, los deportes, entre otros.

Además, incursionamos en las características de la transformación y de la manipulación de datos.

Apache Spark es una gran herramienta para la transformación de datos, ya que mejora y otorga varias ventajas a Hadoop.

Bibliografía

- Databricks. (s.f.). *Transformations.*
Recuperado de https://databricks.com/glossary/what-are-transformations

- IBM Cloud Education. (2020). *ETL (Extract, Transform, Load).*
Recuperado de https://www.ibm.com/cloud/learn/etl

- Kadam, G. (2018). *Beneath RDD (Resilient Distributed Dataset) in Apache Spark.*
Recuperado de https://medium.com/@gkadam2011/beneath-rdd-resilient-distributed-dataset-in-apache-spark-260c0b7250c6

- Maldonado, J. (2020). *¿Qué es un DAG y cómo funcionan?*
Recuperado de https://es.cointelegraph.com/explained/what-is-a-dag-and-how-do-it-work

- Mulchandani, D. (2020). *Difference Between MapReduce and Apache Spark.*
Recuperado de https://www.geeksforgeeks.org/difference-between-mapreduce-and-apache-spark/

- Omar, S. (2020). *Wide Vs Narrow dependencies in Apache Spark.*
Recuperado de https://sauravomar01.medium.com/wide-vs-narrow-dependencies-in-apache-spark-2cd33bf7ed7d

- Pérez, M. (2016). *Apache Spark: qué es y cómo funciona.*
Recuperado de https://geekytheory.com/apache-spark-que-es-y-como-funciona

- Srinivasan. (2020). *A Comprehensive Guide to Data Manipulation for Experts and Beginners.*
Recuperado de https://www.digitalvidya.com/blog/data-manipulation/

- Talend. (s.f.). *Data Transformation Defined.*
Recuperado de https://www.talend.com/resources/data-transformation-defined/#:~:text=Data%20transformation%20is%20the%20process,data%20wrangling%20and%20data%20warehousing.