Curso

Data Engineer

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Centro de e-Learning UTN.BA

Módulo 1:

Extracción y almacenamiento de datos

Unidad 1:

Almacenamiento de datos

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Presentación

En esta unidad hablaremos sobre diferentes aspectos del almacenamiento de datos, en el contexto de Data Engineering.

En primer lugar, trataremos sobre **formatos de almacenamiento**, como los **basados en columnas** y los **basados en fila**. Haremos foco en el primer tipo de formato, como por ejemplo Parquet, y su capacidad para comprimir el tamaño de los datos y permitir un análisis óptimo de los mismos.

Luego, describiremos dos tipos de **sistemas de almacenamiento centralizado**, **Data Lake** y **Data Warehouse**. Un Data Lake permite almacenar datos en su forma cruda y sin procesar. Además admite la ingesta de datos de diferentes fuentes y formatos sin necesidad de definir un esquema rígido de antemano. En un Data Lake se sigue un esquema de **ELT** (Extract, Load, Transform) que consiste en recolectar los datos, cargarlos en un repositorio y luego procesarlos.

Por otro lado, un **Data Warehouse** es un sistema diseñado para **almacenar, organizar y procesar datos estructurados** con el propósito de respaldar el análisis de negocios y la toma de decisiones. A diferencia de un Data Lake, un Data Warehouse sigue un enfoque estructurado y requiere **definir un esquema antes de la carga de datos.** En este caso, se sigue el esquema de **ETL** (Extract, Transform, Load), donde los datos se extraen de las fuentes, se procesan y luego se cargan y almacenan.

Por último, hablaremos sobre Big Data y el almacenamiento distribuido. Big Data se define en término de 3 V: Volumen, Velocidad y Variedad. El almacenamiento distribuido es una técnica utilizada para almacenar y procesar grandes volúmenes de datos sobre clusters (servidores o computadores interconectadas entre sí), lo que permite una mayor escalabilidad y paralelismo en las operaciones.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Bloques temáticos

1. OLAP y formatos analíticos
2. Data Lake
3. Data Warehouse
4. Big data

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

OLAP y formatos analíticos

Luego de recolectar y extraer datos de diferentes fuentes, los almacenamos en algún sistema centralizado para poder procesarlos, analizarlos y obtener valor. Los datos se almacenan, en primera instancia, de forma cruda. A medida que los procesamos, limpiamos, transformamos, etc. iremos haciendo usos de formatos que optimicen el espacio y permitan un análisis eficiente.

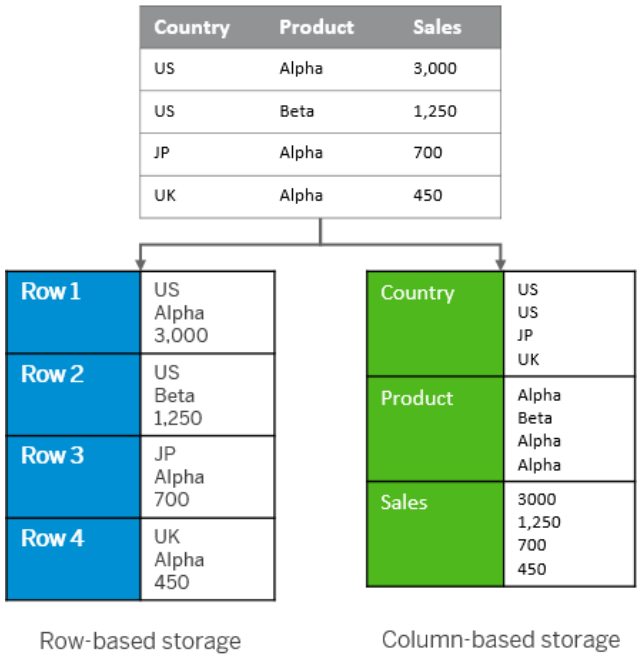
Es muy probable que vayamos a manejar grandes volúmenes de datos y que nos encontremos con datos de millones de filas o registros, lo que se traduce en muchos gigabytes o terabytes de almacenamiento. En primer lugar, es posible reducir los tamaños de los datos trabajando con formatos binarios que comprimen los mismos. Seguramente, muchos de nosotros estamos acostumbrados a trabajar con formatos como .ZIP o .RAR. En el campo de la ingeniería, existen formatos similares que ofrecen varias ventajas. Los más populares son: Parquet, Avro, ORC, etc.

Avro, Parquet y ORC son formatos de datos muy utilizados en Data Engineering y en Big Data. Todos tienen en común lo siguiente:

- **Compresión:** reducen significativamente el tamaño de los archivos de datos.
- **Rendimiento:** ofrecen un buen rendimiento tanto en la lectura como en la escritura de datos, frente a formatos de texto plano como CSV, JSON, TXT, etc.
- **Autodescriptivos:** El esquema se almacena dentro del archivo de datos, de modo que las aplicaciones pueden entender los datos sin tener que depender de metadatos externos.
- **Evolución del esquema:** el esquema puede modificarse sin romper la compatibilidad con los archivos de datos existentes.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Ahora bien, además de estas similitudes, presentan una diferencia en cuanto al formato de almacenamiento. Avro es un formato **basado en filas** (también denominado como **row-based**), mientras que Parquet y ORC son **formatos columnares**, o basados en columnas (**column-based**). Los formatos columnares suelen ser más eficientes para cargas de trabajo analíticas, las cuales se caracterizan por leer o consultar grandes cantidades de registros. Mientras que los formatos basados en filas son más eficientes para cargas de trabajo operativas o transaccionales, las cuales consisten en muchas operaciones de escrituras de datos, que insertan o agregan nuevos registros.



Ejemplo de formato basado en filas y en columnas

Volviendo al formato Parquet, se trata de un formato de almacenamiento orientado a columnas. Una de sus características es que permite acelerar y optimizar los tiempos de consulta sobre los datos, ya que sólo pueden leer las columnas necesarias para una consulta concreta. Por eso, está pensado para cargas de trabajo analítica donde, generalmente, se realizan cálculos sobre un

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

conjunto de columnas en particular. Por ejemplo, si quisiéramos calcular el total de ventas, a partir del campo Sales:

- En el formato “row-based”, hay que iterar por cada fila, obtener el campo “Sales” e ir acumulando ese número.
- Mientras que en un formato “column-based”, los valores de Ventas ya están disponibles en el mismo espacio y no es necesario iterar, solo es cuestión de aplicar la operación deseada.

En la práctica, los lenguajes de programación y las librerías ya cuentan con métodos y funcionalidades predefinidas para manipular formatos como Parquet. Por ejemplo, la librería Pandas de Python cuenta con el método `to_parquet` para el almacenamiento de los datos en dicho formato. Esto es posible, ya que son **formatos abiertos** y permiten la **interoperabilidad** con diferentes tipos de herramientas.

Los formatos columnares se utilizan a menudo en bases de datos OLAP (procesamiento analítico en línea). Las bases de datos OLAP están diseñadas para cargas de trabajo analíticas, como la consulta de grandes cantidades de datos para identificar tendencias y patrones. Los formatos columnares pueden mejorar el rendimiento de las consultas OLAP almacenando los datos de forma que sea más fácil leerlos y procesarlos.

Data Lake

Un Data Lake es un **repositorio centralizado** que permite **almacenar datos de cualquier tipo de estructura**, desde estructurados a no estructurados, a cualquier escala. Almacena los datos de forma cruda sin modificarlos y sin tener que estructurarlos primero. Esto permite que los datos sean accesibles para diferentes análisis y aplicaciones en una etapa posterior.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Está pensado para contener datos de fuentes heterogéneas, desde tablas de bases de datos, pasando por mediciones de sensores, publicaciones de redes sociales, imágenes, documentos, etc.

Los Data Lakes se implementan comúnmente utilizando servicios en la nube, como Amazon S3, Azure Data Lake Storage o Google Cloud Storage. Estos servicios ofrecen una serie de ventajas. En primer lugar, proporcionan una escalabilidad prácticamente ilimitada, lo que significa que los Data Lakes pueden crecer y adaptarse fácilmente a medida que se agregan más datos. Además, estos servicios generalmente ofrecen modelos de precios flexibles, lo que resulta en un menor costo en comparación con la implementación y mantenimiento de infraestructuras físicas.

ELT

ELT es el acrónimo de "**Extract, Load, and Transform**" (extraer, cargar y transformar). Describe las tres etapas de un tipo de pipelines de datos. Consiste en **extraer datos** de los sistemas de origen y **cargarlos inmediatamente en un sistema de destino para transformarlos posteriormente, según sea necesario**. El sistema de destino, donde se depositan los datos, suelen ser un Data Lake.

El Data Lake **no requiere aplicar una estructura** o un esquema definido de antemano en el momento de la carga. Permite la captura de datos de manera rápida, económica y altamente escalable.

ELT aprovecha las capacidades del Data Lake para almacenar los datos extraídos de diferentes fuentes en su forma original, sin requerir una transformación previa. Esto significa que los datos se cargan directamente en el Data Lake y luego se transforman o procesan dentro de él, aprovechando las capacidades de procesamiento distribuido y escalable del entorno de almacenamiento.

Un último aspecto a destacar del vínculo entre ELT y Data Lake es la **flexibilidad.** Al cargar los datos en su forma original en el Data Lake, se pueden

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

aplicar transformaciones y análisis en cualquier momento posterior. Esto permite un enfoque más flexible en el procesamiento de datos, ya que las transformaciones se pueden ajustar y adaptar según las necesidades y los requisitos cambiantes.

Capas de un Data Lake

No basta con tener los datos crudos en un repositorio, los Data Lakes se dividen típicamente en zonas o capas para organizar y gestionar los datos de manera eficiente. Estas zonas representan diferentes **niveles de procesamiento y gobernanza sobre los datos almacenados.**

Dependiendo de la bibliografía, los nombres de las zonas pueden variar pero lo importante es destacar el propósito de cada una. A continuación se describen cada una de las capas de un Data Lake:

1. **Landing:** La capa de “Landing”, también conocida como zona de aterrizaje o de “**raw data**”, es la primera capa en un Data Lake. Aquí es donde los datos aterrizan desde diversas fuentes, por medio de pipelines, sin realizar transformaciones significativas. Los datos se almacenan en su forma bruta, preservando su integridad original. En esta capa, se pueden incluir datos estructurados, semi-estructurados y no estructurados, por eso los formatos a usar aquí pueden variar.

2. **Trusted:** La capa Trusted es donde los datos crudos se transforman y preparan para un uso más amplio. Aquí se aplican procesos de **limpieza, normalización, de-duplicación, validación** y otras transformaciones para mejorar la **calidad** y la **estructura** de los datos. Esta capa tiene como objetivo ofrecer datos de calidad, más estructurados y listos para los equipos de análisis y ciencia de datos. En esta capa, se suele trabajar con formatos columnares como Parquet.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

3. **Refined:** La capa Refined es la capa final en un Data Lake, donde se ofrecen **datos enriquecidos y de valor** para la organización. Aquí se aplican **reglas de negocio** y se realizan **agregaciones, cálculos, transformaciones más avanzadas** y se optimizan los datos para casos de uso específicos. Esta zona se utiliza para generar informes, paneles de control, visualizaciones y otros productos de datos que brinden información valiosa y respalden la toma de decisiones empresariales. Los datos en esta zona están altamente estructurados y están diseñados para ser utilizados por aplicaciones empresariales y usuarios finales de manera eficiente y rápida.

La división en capas proporciona una **estructura lógica** para gestionar y organizar los datos en el Data Lake. Cada capa tiene un propósito específico y ofrece distintos niveles de procesamiento y calidad de los datos. Las tres zonas mencionadas proporcionan una estructura común para organizar, preparar y utilizar los datos en un data lake. Ahora bien, es posible sumar nuevas capas de acuerdo a la necesidad de la organización. A continuación se describen dos capas adicionales o complementarias.

- **Sensitive**
  En esta zona, se almacenan y se gestionan datos sensibles que requieren un tratamiento especial debido a su naturaleza **confidencial** o regulaciones de privacidad. Esta zona está diseñada para **garantizar la seguridad y el cumplimiento normativo** de los datos sensibles. Aquí se aplican reglas y medidas específicas de seguridad, como el cifrado, el control de acceso restringido para proteger la confidencialidad y la integridad de los datos. Además, se pueden implementar técnicas de **anonimización** o pseudonimización para garantizar la privacidad de los datos personales.

- **Sandbox**
  Es un espacio dedicado a la **experimentación** y la colaboración para los equipos de Data Science. Aquí, los/as profesionales de Data Science pueden acceder a una copia de los datos en el data lake para realizar **pruebas, prototipos y experimentos** sin afectar los datos en las capas

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

principales. La zona de sandbox proporciona un entorno seguro para probar nuevas ideas, explorar modelos de machine learning, desarrollar algoritmos y realizar investigaciones sin riesgo de impacto en las otras capas del Data Lake. Los datos en esta zona se utilizan para el desarrollo y la validación de modelos antes de ser implementados. Además, los equipos pueden colaborar y compartir sus resultados y descubrimientos en esta zona sin interferir con los flujos de trabajo principales.

Organización por dominios y sistemas

Además de las capas o zonas que describimos anteriormente, un data lake también puede organizarse por dominios de la organización y por los sistemas o fuentes de datos de la compañía. Esto implica estructurar y categorizar los datos en función de su origen y su relación con diferentes áreas de la organización.

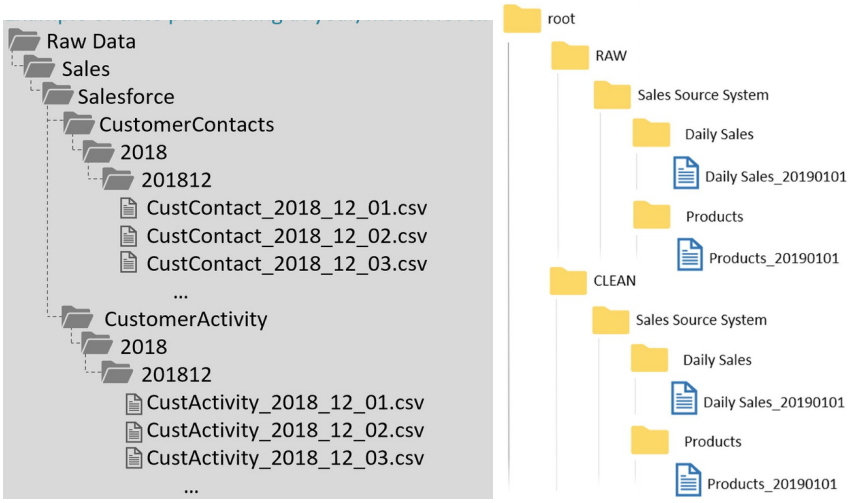
Organizar el data lake por dominios de la organización implica agrupar los datos según las diferentes áreas funcionales o departamentos de la empresa, como Ventas, Finanzas, Recursos Humanos, Marketing, Operaciones, entre otros. Cada dominio puede tener su propia sección o espacio dentro del data lake donde se almacenan y gestionan los datos relevantes para esa área específica. Esto facilita el acceso y la colaboración de los equipos dentro de cada dominio, ya que pueden trabajar con los datos específicos que son relevantes para sus funciones y responsabilidades.

Por otro lado, organizar el data lake por sistemas o fuentes de datos implica agrupar los datos según las diferentes aplicaciones, sistemas o fuentes de donde provienen. Esto puede incluir sistemas de gestión de clientes (CRM), sistemas de gestión de recursos humanos (HRMS), sistemas de gestión financiera (ERP), sistemas de seguimiento de marketing, entre otros. Cada sistema o fuente de datos puede tener su propio espacio dentro del data lake, lo que permite una gestión y un acceso más eficiente de los datos provenientes de diferentes fuentes. Esto facilita el seguimiento y la trazabilidad de los datos

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

hasta su origen y permite un análisis más específico en función de los sistemas o aplicaciones utilizados por la organización.

Organizar el data lake tanto por dominios de la organización como por sistemas o fuentes de datos proporciona una visión más holística y estructurada de los datos, permitiendo a los equipos acceder, analizar y utilizar la información de manera más efectiva y enfocada en sus respectivas áreas y fuentes de origen.



Ejemplo de capas de un Data Lake, organizadas por dominios y/o sistemas

Data Warehouse

Un Data Warehouse es un sistema de **almacenamiento centralizado**, optimizado para almacenar datos históricos y actuales de una organización. Estos datos se estructuran y organizan de acuerdo con un esquema predefinido para facilitar el **análisis y la generación de informes.**

El objetivo principal de un Data Warehouse es proporcionar una fuente única y confiable de datos para la toma de decisiones empresariales y el análisis. A diferencia de otras bases de datos transaccionales, que están diseñadas para

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

admitir operaciones de transacciones diarias, el Data Warehouse se enfoca en admitir consultas analíticas complejas y exhaustivas.

Un Data Warehouse **integra** datos de múltiples **fuentes estructuradas**, como sistemas de gestión de relaciones con clientes (CRM), sistemas de gestión de recursos empresariales (ERP) y bases de datos transaccionales. Estos datos se extraen, transforman y cargan (**ETL**) en el Data Warehouse, donde se organizan y estructuran en un formato que facilite su análisis.

El esquema de datos en un Data Warehouse se establece de antemano y generalmente sigue un modelo dimensional o un modelo de estrella. Esto significa que los datos se organizan en dimensiones (características descriptivas) y hechos (medidas cuantitativas). Este enfoque facilita la realización de consultas y análisis eficientes mediante herramientas de business intelligence (BI).

Los Data Warehouses también almacenan datos históricos a lo largo del tiempo, lo que permite realizar análisis retrospectivos y comparativos. Esto es esencial para identificar tendencias, patrones y comportamientos en los datos, lo que a su vez ayuda a tomar decisiones informadas y estratégicas.

ETL

Como mencionamos anteriormente, ETL es el acrónimo de Extract, Transform, Load (extraer, transformar, cargar). Es un proceso fundamental en la ingeniería de datos que se utiliza para extraer datos de diversas fuentes, transformarlos de acuerdo con las necesidades específicas y cargarlos en un destino de almacenamiento, como un Data Warehouse.

El vínculo entre ETL y Data Warehouse es estrecho y crucial para garantizar la integridad y la calidad de los datos en el Data Warehouse. Luego de la extracción, se realizan una serie de transformaciones para limpiar, filtrar, combinar y estructurar los datos de acuerdo con el modelo de datos predefinido

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

del Data Warehouse. Las transformaciones pueden incluir la normalización de datos, el cálculo de nuevas variables, la agregación de datos y la resolución de inconsistencias. El objetivo es garantizar la calidad y coherencia de los datos antes de cargarlos en el Data Warehouse.

Una vez finalizado el proceso de transformación, los datos transformados se cargan en el Data Warehouse de una forma estricta, siguiendo un esquema predefinido y tabular.

Modelado dimensional

El modelado dimensional es una técnica utilizada en la ingeniería de datos para diseñar la estructura de un Data Warehouse de manera eficiente y optimizada para el análisis de datos. Se basa en la creación de modelos que representan las dimensiones y los hechos de un conjunto de datos, siguiendo un enfoque intuitivo y fácil de entender.

En el modelado dimensional, se distinguen dos conceptos principales: dimensiones y hechos.

1. Dimensiones: Las dimensiones representan las características descriptivas y contextuales de los datos. Pueden incluir elementos como fecha, producto, ubicación, cliente, entre otros. Cada dimensión tiene una tabla asociada en el modelo dimensional, donde se almacenan los atributos y las jerarquías correspondientes. Las tablas de dimensiones contienen columnas que describen los diferentes niveles de granularidad y los atributos relacionados con cada dimensión.

2. Hechos: Los hechos representan las medidas cuantitativas y numéricas que se analizan en el Data Warehouse. Pueden ser valores monetarios, cantidades, recuentos o cualquier otra medida relevante para el análisis. Los hechos se almacenan en una tabla de hechos en el modelo dimensional y se relacionan con las tablas de dimensiones mediante claves externas.

El enfoque central del modelado dimensional es la creación de esquemas de estrella y copo de nieve.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

- Esquema de estrella (star schema): En el esquema de estrella, una tabla de hechos central se conecta directamente a múltiples tablas de dimensiones. La tabla de hechos contiene las claves primarias de las dimensiones y las medidas numéricas. Este enfoque simplifica las consultas y los análisis, ya que se accede directamente a los datos en una única tabla de hechos.

- Esquema de copo de nieve (snowflake schema): En el esquema de copo de nieve, las tablas de dimensiones se normalizan en múltiples niveles, lo que resulta en una estructura en forma de copo de nieve. Esto puede ayudar a reducir la redundancia de datos, pero también puede complicar las consultas al requerir más joins entre tablas para acceder a los datos.

El modelado dimensional tiene varias ventajas:

- Facilidad de comprensión: El modelado dimensional utiliza una estructura intuitiva y fácil de entender, lo que facilita la interpretación de los datos y la construcción de consultas.

- Rendimiento optimizado: El modelado dimensional permite un acceso rápido a los datos y un rendimiento eficiente en las consultas analíticas, ya que se minimiza la cantidad de joins necesarios y se evita la duplicación de datos.

- Flexibilidad y escalabilidad: El modelado dimensional es altamente flexible y se adapta bien a los cambios y adiciones futuras en las dimensiones y medidas. También es escalable, lo que significa que puede manejar grandes volúmenes de datos sin perder rendimiento.

Big data

El término "big data" se refiere a conjuntos de datos que son tan grandes y complejos que no pueden ser gestionados ni procesados fácilmente con

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

herramientas tradicionales de procesamiento de datos. Estos conjuntos de datos grandes suelen caracterizarse por las llamadas "tres V":

- **Volumen**: se refiere a la gran cantidad de datos que se generan. Estos datos pueden alcanzar tamaños enormes, desde terabytes hasta petabytes o incluso exabytes.

- **Velocidad**: se refiere a la tasa a la cual se generan los datos. En algunos casos, los datos se generan en tiempo real, como las mediciones de sensores. El procesamiento de datos en tiempo real requiere tecnologías y enfoques especiales para garantizar que los datos se capturen y procesen en el menor tiempo posible.

- **Variedad**: se refiere a la diversidad de tipos y formatos de datos, desde estructurados y tabulares a semiestructurados y no estructurados, como textos, imágenes, videos, etc. Manejar esta variedad de datos requiere técnicas de procesamiento y análisis específicas.

Además de las "tres V", a menudo se mencionan dos V adicionales:

- **Veracidad:** La veracidad se refiere a la calidad y confiabilidad de los datos, ya que pueden estar sujetos a problemas de calidad, como ruido, errores o inconsistencias. Es necesario aplicar técnicas de limpieza, normalización y validación de datos para garantizar su veracidad antes de su análisis.

- **Valor:** El valor se refiere al potencial de obtener información y conocimientos significativos a partir de los datos. El análisis de grandes volúmenes de datos puede revelar patrones, tendencias y correlaciones que pueden ser utilizados para tomar decisiones informadas, descubrir oportunidades de negocio y mejorar la eficiencia operativa.

Dado el tamaño, la velocidad y la variedad de los datos, como Data Engineers debemos utilizar técnicas y herramientas especiales para gestionar y procesar estos datos. Esto incluye tecnologías como el almacenamiento y procesamiento distribuido, por medio de clusters como infraestructura, y el uso de frameworks

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

y plataformas diseñados específicamente para el big data, como Apache Hadoop, Apache Spark, etc.

Cluster y Almacenamiento distribuido

Un cluster se refiere a un **grupo de computadoras** o servidores conectados entre sí, que trabajan en conjunto para **procesar y almacenar grandes volúmenes de datos,** de forma eficiente.

Un cluster es la **infraestructura** o el hardware con el que trabajaremos como Data Engineers.

El **almacenamiento distribuido** es una técnica clave utilizada en un cluster. En lugar de almacenar todos los datos en una sola máquina, los datos se dividen en fragmentos más pequeños llamados **particiones**. Además se crean copias o **réplicas** de cada partición y se distribuyen y almacenan en diferentes nodos del clúster. La distribución equilibrada de las particiones permite un acceso rápido y eficiente a los datos.

La principal tecnología de Big Data que implementa el almacenamiento distribuido es **Apache Hadoop** y su sistema de archivos distribuido (**HDFS**: Hadoop Distributed File System)

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Conclusión

Al explorar los diversos aspectos del almacenamiento de datos, hemos descubierto la importancia de considerar diferentes factores para elegir la estrategia de almacenamiento más efectiva.

En primer lugar, es fundamental tener en cuenta el tipo de pipeline que vamos a utilizar: **ETL** o **ELT**. Si seguimos un enfoque tradicional de ETL, donde los datos se transforman antes de cargarlos en un Data Warehouse, es importante contar con un **esquema estructurado y definido previamente**. Por otro lado, si optamos por el enfoque más **flexible** de ELT, donde los datos se cargan primero en un Data Lake y luego se transforman, podemos manejar datos **no estructurados o semi estructurados** sin restricciones.

Así mismo, los **formatos columnares**, como **Parquet**, han demostrado ser una opción altamente eficiente para consultas analíticas en grandes volúmenes de datos. Estos formatos **optimizan el acceso a los datos** y **reducen el uso de recursos**, lo que se traduce en un rendimiento mejorado y una mayor eficiencia en las consultas.

Además, en un contexto de Big Data, donde los volúmenes de datos son masivos y superan la capacidad de los sistemas tradicionales, el almacenamiento distribuido se vuelve fundamental. La capacidad de distribuir y procesar datos en clústers de servidores interconectados nos permite aprovechar al máximo los recursos y obtener resultados escalables y en menor tiempo. El almacenamiento distribuido nos brinda la flexibilidad necesaria para abordar los desafíos del Big Data y realizar análisis complejos en grandes conjuntos de datos.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Bibliografía utilizada y sugerida

- A complete overview of dimensional data modeling. (2023, 31 mayo). ThoughtSpot. https://www.thoughtspot.com/data-trends/data-modeling/dimensional-data-modeling

- Data lake zones and containers - Cloud Adoption Framework. (2023, 6 junio). Microsoft Learn. https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/best-practices/data-lake-zones

- Marr, B. (2018, 27 agosto) What Is A Data Lake? A Super-Simple Explanation For Anyone. Forbes. https://www.forbes.com/sites/bernardmarr/2018/08/27/what-is-a-data-lake-a-super-simple-explanation-for-anyone/?sh=6b699ecc76e0

- MongoDB. (s. f.). Databases Vs. Data Warehouses Vs. Data Lakes. https://www.mongodb.com/databases/data-lake-vs-data-warehouse-vs-database

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning