Data Lake & Data Warehouse

Introducción

En el ámbito de la gestión de datos, dos conceptos clave son los Data Lakes y los Data Warehouses. Estas dos tecnologías desempeñan roles fundamentales en el almacenamiento y análisis de datos en el contexto empresarial. En este artículo, exploraremos en detalle cada uno de estos conceptos, y analizaremos cómo se pueden utilizar las herramientas de Postgres, SQL y Python para implementarlos de manera efectiva.

Con la creciente cantidad de datos producidos, la nube ofrece muchos beneficios para el procesamiento y análisis de datos, como la escalabilidad, confiabilidad y disponibilidad. Además, existen diversas herramientas y tecnologías para el procesamiento y análisis de datos en el ecosistema de la nube.

En cualquier diseño de plataforma de análisis, el procesamiento y almacenamiento son fundamentales para el rendimiento de las plataformas de datos. Hay tres categorías principales de plataformas de análisis: data warehouses, data lakes y data lakehouses. Veamos brevemente cada una de ellas.

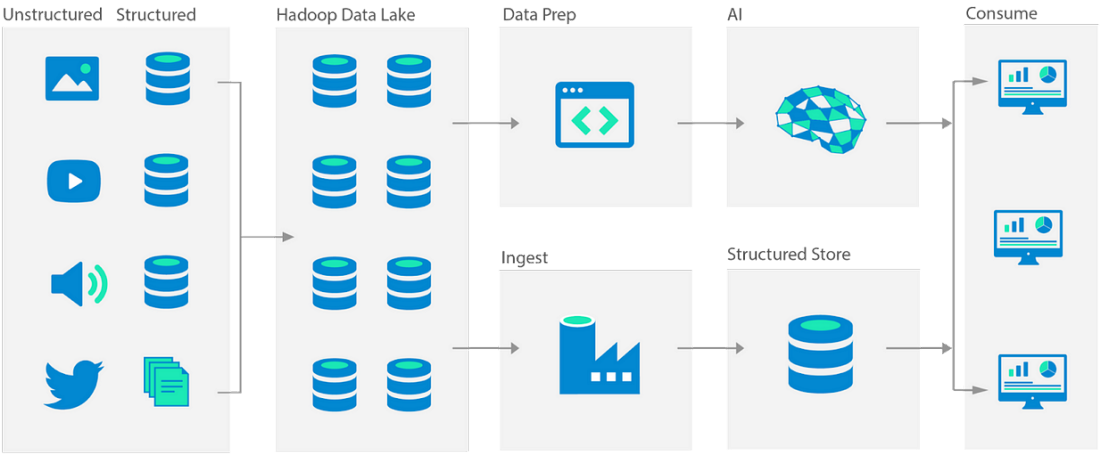
Data Lakes

Comencemos por comprender qué es un Data Lake. Un Data Lake es un repositorio centralizado y escalable que almacena grandes volúmenes de datos estructurados y no estructurados en su formato original. A diferencia de los sistemas tradicionales de almacenamiento de datos, los Data Lakes no requieren una estructura definida antes de la carga de datos, lo que permite la captura de información en su forma bruta y su posterior transformación según las necesidades.

Un Data Lake utiliza un enfoque de almacenamiento basado en archivos, donde los datos se almacenan en su forma nativa, sin imponer una estructura predefinida. Esto permite una mayor flexibilidad en la gestión y el análisis de los datos, ya que no se requiere un esquema rígido antes de su almacenamiento. Además, los Data Lakes pueden almacenar una amplia variedad de datos, como datos transaccionales, datos de sensores(IoT), datos de registros de aplicaciones y más.

Para implementar un Data Lake utilizando Postgres, es posible utilizar herramientas como Amazon Web Services: S3 y Apache Hadoop. Postgres (RDS=DDBB) proporciona funcionalidades de almacenamiento y acceso a datos escalables, mientras que Amazon S3 y Apache Hadoop permiten la gestión eficiente de grandes volúmenes de datos no estructurados y semiestructurados.

Los data lakes resuelven muchos desafíos de los data warehouses, pero tienen una baja calidad de datos y un rendimiento de consultas que no es lo suficientemente eficiente. Además, requieren herramientas adicionales para ejecutar consultas SQL para usuarios de negocio. Si un data lake no está bien organizado, puede llevar a un problema de estancamiento de datos.



Data Warehouses

Pasemos ahora a los Data Warehouses. Estos se basan en un sistema utilizado para la recopilación, almacenamiento y análisis de datos de diferentes fuentes con el objetivo de apoyar la toma de decisiones empresariales. A diferencia de los Data Lakes, los

Data Warehouses están diseñados para almacenar datos estructurados, normalizados y listos para su análisis.

La estructura de un Data Warehouse se basa en un modelo dimensional, que consiste en tablas de hechos y tablas de dimensiones. Las tablas de hechos contienen las medidas numéricas y cuantitativas, mientras que las tablas de dimensiones contienen atributos descriptivos para proporcionar contexto a los datos. Esta estructura facilita las consultas y análisis eficientes de los datos almacenados en el Data Warehouse.

Postgres es una base de datos relacional ampliamente utilizada que puede utilizarse para implementar un Data Warehouse. Utilizando Postgres junto con SQL, es posible modelar la estructura dimensional y cargar los datos en las tablas correspondientes. SQL proporciona un lenguaje de consulta potente para extraer información de manera eficiente de un Data Warehouse.

Data Lake (capa raw(sin formato) → capa stage(con formato, normalizamos) AWS

Data Warehouse (SnowFlake)

Integración con Python

Python, por su parte, es un lenguaje de programación versátil y ampliamente utilizado en el ámbito del análisis de datos. Al combinar Python con Postgres y SQL, podemos lograr una integración poderosa para trabajar con Data Lakes y Data Warehouses.

Python ofrece una amplia gama de bibliotecas y herramientas que facilitan la manipulación y análisis de datos. Por ejemplo, pandas es una biblioteca popular que proporciona estructuras de datos flexibles y funciones para la limpieza, transformación y análisis de datos. Al cargar datos desde un Data Lake o Data Warehouse en Python utilizando pandas, podemos aprovechar su funcionalidad para preparar los datos antes de su análisis.

Para acceder a los datos almacenados en un Data Lake o Data Warehouse, podemos utilizar el conector de bases de datos de Postgres en Python. Este conector permite establecer una conexión con la base de datos y ejecutar consultas SQL para extraer los datos necesarios. La biblioteca psycopg2 es ampliamente utilizada para este propósito y proporciona una interfaz fácil de usar para interactuar con Postgres desde Python.

Una vez que hemos extraído los datos en Python, podemos aplicar diversas técnicas y análisis para obtener información valiosa. Por ejemplo, podemos utilizar algoritmos de aprendizaje automático de bibliotecas como scikit-learn o TensorFlow para realizar análisis predictivos o clustering en los datos. También podemos utilizar bibliotecas de visualización, como Matplotlib o Seaborn, para representar gráficamente los resultados y comunicarlos de manera efectiva.

En resumen, los Data Lakes y los Data Warehouses desempeñan un papel crucial en la gestión y análisis de datos en el entorno empresarial. Mientras que los Data Lakes se centran en almacenar datos en su formato bruto(raw) y no estructurado, los Data Warehouses se utilizan para almacenar datos estructurados(normalizados) y listos para el análisis. Postgres, SQL y Python son herramientas poderosas que se pueden utilizar para implementar estas tecnologías y aprovechar al máximo los datos almacenados.

En conclusión, la combinación de Postgres, SQL y Python proporciona una plataforma sólida para trabajar con Data Lakes y Data Warehouses. Estas tecnologías permiten almacenar, acceder, transformar y analizar datos de manera eficiente. Al comprender y utilizar adecuadamente estas herramientas, los profesionales pueden aprovechar al máximo la información contenida en los Data Lakes y Data Warehouses para tomar decisiones informadas y obtener una ventaja competitiva en el mundo empresarial actual.

Bonus:

Data Lakehouses

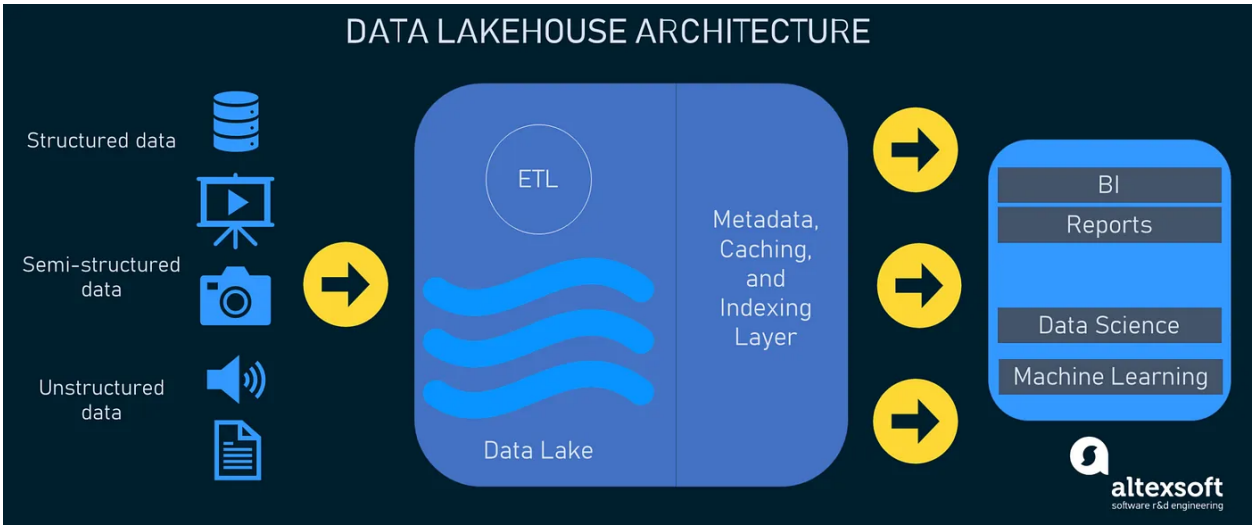
Un data lakehouse ofrece una mayor confiabilidad de los datos al reducir las transferencias de datos de ETL pero ofrecer almacenamiento de datos en bruto. Los datos no se duplicarán en múltiples sistemas. Debido a la reducción de los procesos de ETL(Extract, Transform & Load) y la eliminación de duplicados, también se reduce el costo. Además, ofrece una mejor gestión de datos y abre los datos para múltiples casos de uso.

Algunas de las características clave del data lakehouse son las siguientes:

- Soporte de transacciones ACID: ACID (Atomicidad, concurrencia, aislamiento y durabilidad) garantiza la consistencia de las transacciones y la integridad de los datos. Ayuda a mantener la integridad de los datos cuando diferentes componentes realizan operaciones concurrentes o en caso de fallas. Es una propiedad fundamental de un data warehouse y se hereda en un data lakehouse.
- Formato de datos en bruto o no estructurado: Un data warehouse solo admitía datos estructurados, pero ahora tenemos soporte para tipos de datos en bruto, incluidos audio, video, etc.
- Soporte de transmisión continua: Ahora, los datos se generan como un flujo sin límites, por lo que el data lake tiene soporte para transmitir los datos y generar información en tiempo real.
- Almacenamiento y procesamiento desacoplados: El almacenamiento y el procesamiento están desacoplados, lo que permite escalarlos de forma independiente según las necesidades del caso de uso. También te permite ejecutar consultas utilizando nodos de cómputo diferentes, mientras otros acceden al almacenamiento directamente.

En conclusión, los data warehouses han existido durante mucho tiempo y han madurado, pero no están diseñados para las necesidades modernas de procesamiento de datos. Por otro lado, los data lakes resuelven la mayoría de los desafíos, pero eliminan algunas de las mejores características de los data warehouses. Por lo tanto, el concepto de data lakehouse surgió y trajo lo mejor de ambos mundos. Sin embargo, la arquitectura del data lakehouse aún es relativamente nueva y llevará tiempo madurarla y compartir las mejores prácticas por parte de los primeros adoptantes. Mientras tanto, los data warehouses y los data lakes todavía se implementan para casos de uso

específicos, y en la mayoría de los casos, coexisten y se complementan bastante bien para resolver el problema en cuestión.



Resumen de diferencias entre cada arquitectura:

Data Warehouse:

- Almacenamiento: Los data warehouses almacenan datos en un formato estructurado y predefinido, con esquemas y tablas definidas de antemano.
- Procesamiento: Los data warehouses están optimizados para consultas y análisis de datos estructurados. Utilizan un modelo de procesamiento en lotes para extraer, transformar y cargar los datos (ETL) antes de que estén disponibles para su análisis.
- Uso de datos: Los data warehouses son ideales para aplicaciones de inteligencia empresarial (BI), informes regulares y consultas ad hoc. Son más adecuados para usuarios de negocio y analistas.
- Calidad de datos: Los data warehouses tienden a tener una calidad de datos más alta, ya que los datos se estructuran y transforman antes de su carga.

Data Lake:

- Almacenamiento: Los data lakes almacenan datos en su formato original, sin estructurar, incluyendo datos estructurados, semiestructurados y sin procesar. No se requiere un esquema definido de antemano.

- Procesamiento: Los data lakes permiten un enfoque más flexible y ágil para el procesamiento de datos. Utilizan herramientas de procesamiento distribuido, como Apache Spark o Apache Hadoop, para procesar datos a gran escala.
- Uso de datos: Los data lakes son más adecuados para aplicaciones de ciencia de datos, análisis exploratorio y descubrimiento de patrones. Son utilizados principalmente por científicos de datos y analistas técnicos.
- Calidad de datos: Los data lakes pueden tener una calidad de datos más baja, ya que los datos no se transforman ni estructuran previamente. Requieren una mayor limpieza y preparación antes de su análisis.

Data Lakehouse:

- Almacenamiento: Los data lakehouses combinan las características de los data warehouses y data lakes, almacenando tanto datos estructurados como sin estructurar en un único repositorio.
- Procesamiento: Los data lakehouses adoptan un enfoque más unificado y convergente para el procesamiento de datos. Permiten tanto consultas SQL tradicionales como análisis de datos a gran escala utilizando herramientas como Spark.
- Uso de datos: Los data lakehouses ofrecen flexibilidad para una amplia gama de casos de uso, desde análisis de BI hasta ciencia de datos avanzada. Pueden ser utilizados por usuarios de negocio, científicos de datos y analistas técnicos.
- Calidad de datos: Los data lakehouses pueden combinar la calidad de datos más alta de los data warehouses con la flexibilidad y variedad de datos de los data lakes. Sin embargo, la calidad de los datos dependerá de la gestión y preparación adecuadas.

En resumen, los data warehouses se centran en datos estructurados y predefinidos, los data lakes se enfocan en datos sin estructurar y los data lakehouses buscan combinar lo mejor de ambos mundos, almacenando y procesando datos estructurados y sin estructurar en un solo lugar.