Inteligencia de negocios
4A. Business Intelligence para OLAP
Por David Aarón Ramírez Olmeda
Programa: Maestría en Ciencia de Datos e Información

Introducción

En la unidad anterior, se planteó una estrategia preliminar para la implementación de un Data Warehouse (DW) basado en datos públicos que se actualizan aproximadamente una vez al año. Este contexto nos lleva a definir un enfoque adecuado para los procesos ETL (Extracción, Transformación y Carga), así como a seleccionar las herramientas y tecnologías más adecuadas para maximizar el valor de los datos a través de la Inteligencia de Negocios (BI).

Contexto de la Unidad

En esta unidad se exploraron los conceptos fundamentales de los modelos OLAP (Online Analytical Processing), destacando su principal característica de cumplir con el principio FASMI (Fast Analysis of Shared Multidimensional Information). Se discutieron diferentes arquitecturas de almacenamiento, como ROLAP, MOLAP y HOLAP, y se analizó la relevancia de los cubos OLAP en comparación con otras soluciones, como los Data Warehouses.

El Modelo Multidimensional fue uno de los enfoques centrales, permitiendo la representación de datos en múltiples dimensiones, lo que facilita un análisis más detallado y eficiente. Además, se examinaron las diferencias entre Data Warehouses y cubos OLAP, destacando la capacidad de los Data Warehouses para manejar grandes volúmenes de datos transaccionales y su integración en el apoyo a la toma de decisiones empresariales. Por otro lado, los cubos OLAP fueron presentados como herramientas clave para el análisis de datos en estructuras tridimensionales, permitiendo un procesamiento analítico en línea altamente eficiente.

Desarrollo

Definición de los procesos ETL

Dado que los datos utilizados son de carácter público y su actualización es esporádica, una vez al año, es fundamental diseñar un proceso ETL que sea eficiente y manejable.

Extracción: Los datos se obtienen a partir de archivos en formatos CSV, XLS, y en algunos casos, PDF. Para automatizar la extracción, se puede desarrollar un script en Python que realice web scraping, lo cual permite obtener los datos directamente desde las fuentes originales. No obstante, esta opción conlleva el riesgo de que un cambio en la estructura del sitio web pueda interrumpir el proceso, requiriendo ajustes en el código. Como alternativa, los datos pueden ser descargados manualmente, minimizando así la necesidad de

mantenimiento continuo. Posteriormente, el script en Python limpiaría y transformaría los datos para prepararlos para su integración en el DW.

Transformación: Una vez extraídos, los datos necesitan ser transformados para asegurar su calidad y consistencia. Este paso incluye la normalización de valores, el tratamiento de datos faltantes y la conversión de formatos. En este caso, se utilizarán bibliotecas de Python como pandas para el manejo de archivos CSV y XLS, y PyPDF2 para la extracción de datos de PDFs.

Carga e integración: Los datos transformados se cargarán en un DW implementado con un esquema de copo de nieve, que facilita la normalización de las dimensiones y asegura la integridad de los datos. Este esquema es ideal cuando las tablas de datos y sus relaciones no siempre guardan una correspondencia directa, como se discutió en la unidad anterior. Además, se debe considerar que la actualización de los datos no depende directamente de nuestro control, lo que agrega una capa de complejidad que es necesario gestionar con técnicas adecuadas de ETL.

Selección de productos y su implementación

Para llevar a cabo las tareas descritas en los procesos ETL, se pueden explorar dos opciones principales: una implementación on-premise o en la nube.

Opción on-premise: Esta opción implica el desarrollo de un script en Python que se ejecutaría en una estación de trabajo o servidor local. Los datos se almacenarían en una base de datos como PostgreSQL o MySQL, permitiendo un control total sobre el proceso ETL y el almacenamiento de datos. Esta opción es viable si se cuenta con los recursos y la infraestructura necesaria, y es ideal para organizaciones que prefieren mantener los datos dentro de su propio entorno de TI.

Opción en la nube: Alternativamente, se puede optar por una solución en la nube que ofrezca mayor flexibilidad y escalabilidad. Por ejemplo, AWS ofrece herramientas como AWS Glue, que permite automatizar el proceso ETL sin la necesidad de escribir líneas de código, y AWS Redshift para almacenar el DW. La integración con otros servicios de AWS facilita la gestión de los datos y permite ajustar los recursos según las necesidades del proyecto. Esta opción, aunque más costosa y compleja, es ideal para proyectos que requieren escalabilidad y una gestión simplificada.

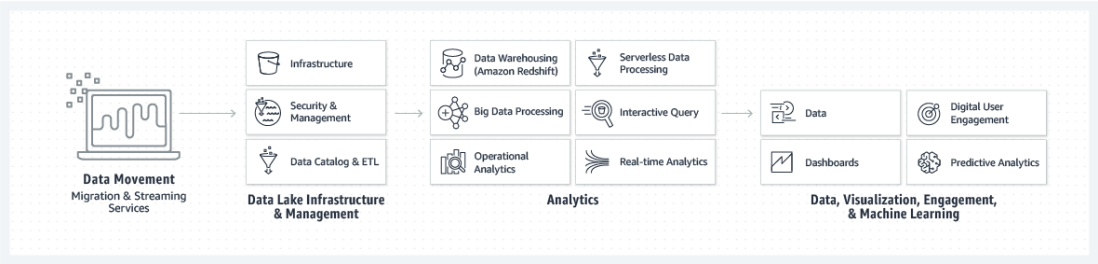

Ilustración 1 Ejemplo proceso ETL y DW en AWS

Herramientas de Inteligencia de Negocios

Una vez que los datos están almacenados y organizados en el DW, el siguiente paso es maximizar su valor mediante herramientas de Inteligencia de Negocios (BI).

Análisis Exploratorio de Datos (EDA): Antes de definir métricas y KPIs, es esencial realizar un análisis exploratorio de los datos para comprender su naturaleza y detectar patrones o tendencias. Este análisis preliminar se puede realizar utilizando herramientas como pandas y matplotlib en Python, que permiten visualizar y analizar los datos de manera eficiente.

Visualización y análisis: Para transformar los datos en información útil, se pueden utilizar herramientas de visualización como Power BI o Tableau, que facilitan la creación de dashboards y scorecards. Estas herramientas permiten a los usuarios finales interactuar con los datos y tomar decisiones informadas. Si el proyecto se implementa en la nube, opciones como Looker en Google Cloud Platform (GCP) o Grafana en AWS son alternativas viables que ofrecen capacidades avanzadas de visualización y análisis.

Consultas e informes: La generación de informes y consultas detalladas es fundamental para extraer KPIs clave. Utilizando SQL en bases de datos como PostgreSQL o Redshift, se pueden realizar consultas que extraigan información crítica para el negocio. Por ejemplo, en el caso de una aseguradora, se podrían generar KPIs relacionados con el costo de siniestros, mientras que en una institución gubernamental, se podrían enfocar en aspectos sociales, como la incidencia de accidentes relacionados con el consumo de alcohol.

Conclusión

En esta actividad, se han integrado los conceptos vistos en la unidad para desarrollar procesos ETL, selección de tecnologías y aplicación de herramientas de Inteligencia de Negocios. La inclusión de estas metodologías y herramientas no solo proporciona un marco técnico, sino que también asegura que se aproveche al máximo el contexto proporcionado por la unidad.

Bibliografía

Amazon Web Services. (n.d.). What is a Data Warehouse? Amazon Web Services, Inc. Retrieved September 12, 2024, from https://aws.amazon.com/what-is/data-warehouse/

IBM. (n.d.). What is OLAP (Online Analytical Processing)? IBM. Retrieved September 12, 2024, from https://www.ibm.com/topics/olap

Amazon Web Services. (n.d.). What is OLAP (Online Analytical Processing)? Amazon Web Services, Inc. Retrieved September 12, 2024, from https://aws.amazon.com/what-is/olap/