

Diseño e implementación de plataforma ETL Cygnus

PRACTICANTE: Juan Pablo Yarce Escobar

ASESORES: Jeysson Perez Gomez

PROGRAMA: Ingeniería de Sistemas

Semestre de la práctica: 2024-1

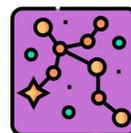
Se desarrolló una plataforma tecnológica llamada Cygnus centrada en procesos ETL utilizando Apache Airflow para optimizar la gestión de inventarios en una empresa logística.

La ETL automatiza la extracción, transformación y carga de datos desde múltiples fuentes, mejorando la escalabilidad y eliminando inconsistencias en el flujo de trabajo.

ETL (Extract, Transform, Load) es un proceso fundamental en la ingeniería de datos, que consiste en la extracción de datos de diversas fuentes, su transformación para ajustarlos a un formato coherente y útil, y su posterior carga en sistemas de almacenamiento, como bases de datos o data warehouses.

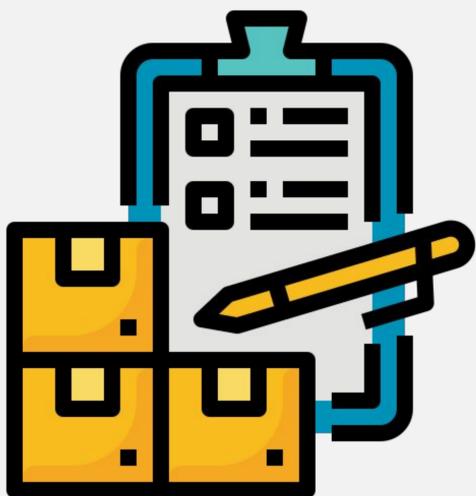


Introducción



Plataforma Cygnus

El sistema de inventarios actual de una empresa centrada en operaciones logística, enfrentaba problemas de inconsistencias y escalabilidad debido a su arquitectura monolítica en PHP Laravel. Para resolver estas deficiencias, se diseñó e implementó una ETL con Apache Airflow, que automatiza el manejo de datos de diversas fuentes y asegura la integridad de la información en un entorno dinámico, controlado y escalable.



Metodología



Se utilizó la metodología ágil "Dragonfly", que permite un enfoque iterativo y colaborativo. Las fases incluyeron la exploración de requerimientos, el diseño y la planificación colaborativa, construcción con integración continua, pruebas de integración y despliegue incremental.



Resultados



- Automatización completa del flujo de datos mediante Apache Airflow, eliminando la intervención manual.
- Integración exitosa de múltiples fuentes de datos.
- Creación de la plataforma Cygnus
- Reducción de errores y mayor consistencia en el manejo de datos, asegurando la integridad de la información en tiempo real.

The ETL Process Explained



Objetivos

- ✓ Diseñar y desarrollar una ETL para optimizar el manejo de datos en la gestión de inventarios.
- ✓ Automatizar los procesos de extracción, transformación y carga de datos desde diversas fuentes.
- ✓ Asegurar la integridad y consistencia de la información en todas las etapas del proceso.
- ✓ Realización de pruebas y monitoreo en un ambiente de producción



Conclusiones

- ✓ La ETL contribuyó a una mejora sustancial en la calidad de los datos procesados,
- ✓ El uso de Apache Airflow proporcionó una solución escalable y flexible, permitiendo que el sistema se adapte a futuros incrementos en los volúmenes de datos y a nuevos requerimientos empresariales.
- ✓ La ETL integró con éxito múltiples fuentes de datos (EDI, XML, Excel) y las unificó en un proceso coherente, asegurando la consistencia de la información en todas las etapas.
- ✓ La implementación de la ETL con Apache Airflow permitió la automatización total del flujo de datos, eliminando la intervención manual y reduciendo significativamente los errores en la gestión de inventarios.