Portfolio

Proyectos

NYC Taxi Project

Proyecto de pipeline ETL en batch para analizar grandes cantidades de datos en entornos distribuidos.

Tecnologías utilizadas : Jupyter notebook, PySpark, CloudStorage, Dataproc, BigQuery, LookerStudio.

Descripción

- Creación de template en local del proceso de transformación con PySpark en un Jupyter Notebook.
- Añadir código para conectar datos de entrada con Cloud Storage y salida a BigQuery
- Crear y desplegar un cluster de Dataproc para mandar dicho job, contenido en un archivo .py
- Analizar datos con interfaz SQL y guardarlos en un data warehouse mediante BigQuery. Creación de tablas para análisis.
- Creación de reportes con LookerStudio para obtener información valiosa de los datos.