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.

Spotify basic ETL Process

Proyecto realizado para tener un primer contacto con la nube de AWS.

Tecnologías utilizadas: AWS S3, IAM, Glue, Athena.

Descripción

- Creación de usuario y rol IAM para empezar a usar los diferentes servicios de AWS.
- Configuración de bucket en S3 y subida de datos (en un caso real, deberían de salir de una bbdd por ejemplo)
- Configuración de pipeline con AWS Glue. Transformaciones como joins o eliminación de duplicados y definición de destino de vuelta en S3.
- Creación y ejecución de un crawler en AWS Glue para descubrir metadatos en los datos transformados.
- Realización de consultas en AWS Athena y almacenamiento de los resultados en S3.