Memoria Big Data Processing

Prática Luis Martín Vegas

**Provisioner:**

En primer lugar, creamos una base de datos (PostgreSQL) en Google Cloud, en la que cargaremos la tabla que contiene los datos de los usuarios, a la que después accederemos cuando realicemos las queries en Batch.

Además, será en este PostgreSQL donde se almacenen los resultados de las diferentes queries que ejecutemos, generando así nuevas tablas.

**Simulación de Datos:**

Para la emisión de datos en tiempo real de las antenas, crearemos una instancia en Compute Engine (Google Cloud) en instalaremos y ejecutaremos zookeeper y Kafka. Dicho Kafka recogerá los datos emitidos por un docker cada minuto en un topic, all que accederemos a través de nuestro job.

**Streaming Job:**

En este job, accederemos a los datos recogidos en el topic de Kafka en tiempo real gracias al micro-batching, leyéndolos para después responder a las preguntas planteadas por medio de cálculos agregados en diversas queries.

Dichos resultados se almacenan en el disco local en formato parquet para después acceder a ellos en el job de Batch.

**Batch Job:**

Mediante Spark, leemos la información recogida en los archivos parquet generados en el paso anterior y nuevamente calculamos las métricas que se nos indica por medio de queries similares a las anteriores pero adaptadas al proceso batch.

En este caso, para el cálculo de la última métrica (límite de consumo por usuario), es necesario realizar un join de los datos del disco local con los datos de los usuarios anteriormente cargados a PostgreSQL, por lo que nos conectamos a la base de datos y los cruzamos en base a los “id”s de usuario.