Práctica de BD - Processing - Keep Coding 2018

Objetivos

- Trabajar, a través de Spark, con fuentes de datos generadas por el comportamiento de uno o varios sistemas, a veces en forma de batch, a veces en forma de tiempo real

- Saber tratar (transformar y accionar) los datos ‘ingestados’ con Spark, a través del conocimiento del lenguaje SQL (SparkSQL)

- Aprender a trabajar con fuentes de datos en tiempo real, cuando el análisis de los datos generados es relevante que suceda en el mismo momento de la generación de la información

- Saber optimizar (tunning) un sistema procesamiento de datos masivo siguiendo algunas ‘buenas prácticas’ comentadas

- Consolidar conocimiento del lenguaje Scala (nativo Spark)

Enunciado

En una galaxia lejana a nuestra Via Láctea hay un sistema solar llamado Andrómeda, con dos soles. Éste da cobijo a numerosas colonias de humanos que tuvieron que huir de La Tierra por los peligros que representaba el Cambio Climático para la vida.

En esta galaxia existe un material muy preciado a la par que escaso como es la coltanita. Se trata de un mineral extraño que permite a las naves de transporte interestelar viajar a lo largo de galaxias (trayectos) sin escalas, incluídos los viajes al planeta azul (todavía hay vida). Este material combustible se está agotando y es necesario hacer estudios y optimizaciones de los datos que se obtuvieron durante algunos años gracias a los sistemas automáticos de recogida de información de cada una de las naves y que fue almacenada en algunas de las colonias.

Actualmente esta información es recogida en tiempo real y sirve para planificar las rutas más optimas, ya que el combustible es escaso. Los puertos de atraque también han generado durante varios años información que también debe ser analizada (batch)
El comandante ha emitido un decreto ley de contrucción urgente de naves de transporte: aquellas que tengan una mejor relación entre el consumo y los trayectos realizados serán utilizadas como modelo para construir las nuevas generaciones de transbordadores espaciales. Para conseguirlo, no sólo serán necesarios los datos obtenidos durante años en los puertos y las naves, sino también el rendimiento en cada uno de los viajes en tiempo real. Los modelos seleccionados serán aquellos donde la diferencia de las medias de consumo por trayecto entre los datos históricos almacenados y las medias del consumo en tiempo real sean menores a cierto umbral U (parametrizable).

Se pide:

- Ingesta de los datos almacenados durante años por los sistemas de navegación de las naves y los puertos de atraque (modo batch)
- Cleaning de datos (nos quedamos sólo con lo relevante)
- Cálculo de medias de consumo de todas las naves de la flota agrupadas por nave (cada nave tiene un identificador)
- Ingesta de datos de consumo por trayecto en tiempo real (Spark Streaming)
- Cálculo de medias de consumo de todas las naves de la flota agrupadas por nave (cada nave tiene un identificador) obtenidas en tiempo real.
- Proceso sobre ambos datasets obteniendo la diferencia entre consumos medios.
- Obtención de una colección (List) de elementos tupla (identificación nave,modelo) con los tres mejores transportes