Apache Spark WordCount

Introducción

Este proyecto recopila diferentes tweets sobre el coronavirus y hace un conteo de las todas las palabras que aparecen en los tweets. Creando una clave valor por palabra (word, number).

Datos

Se han recopilado 2 GB de datos utilizando la API de Twitter. Estos tweets eran tweets en inglés con los hashtags #coronavirus, #covid19 y #covid. Para un manejo más cómodo y rápido de los datos, se adjuntará un fichero .txt llamado sample_english_covid_tweets el cual solo contendrá una muestra de dichos 2GB de tweets.

Funciones

- map-reduce-spark.py
    - Se hace un mapeo por final para obtener solamente los tweets y eliminar los símbolos
    - Se filtran las líneas vacías
    - Se hace un flatMap para conseguir una lista con todas las palabras de todos los tweets
    - Se hace un map para convertir estas palabras en una estructura de clave valor como (word, 1)
    - Se hace un reduceByKey para sumar todos los valores de las claves valor que tengan la misma clave.
    - Se guarda el resultado en una carpeta específica

Tiempos de Ejecución con Archivo de 2GB

- python map-reduce-spark.py
    - 0m54,384s

Ejecutar Demo Funciones con Hadoop

1. sudo apt install python-pip
2. pip install pyspark
3. python map-reduce-spark.py