75.06/95.58 Organización de Datos
Finger 2: Análisis Exploratorio con Apache Spark: Trocafone.

Puntos: Se otorgarán los siguientes puntos en orden de entrega 5,3,2,1,1,1.
Tiempo Estimado de Resolución: 3 hs.

Utilizando las herramientas que provee Apache Spark, se quiere realizar un análisis exploratorio del Set de datos del Trabajo Práctico de la materia.

Los datos del TP están disponibles desde el siguiente link:

https://drive.google.com/file/d/1gUddcLLujjFfwZslypUv1LESTM6KiwJn/view?usp=sharing

Considerando la siguiente información del set de datos

- Eventos (events.csv)

Se plantean los siguientes puntos a realizar:

1. Encontrar cuál es el tipo de evento predominante.
2. Encontrar el Top 5 de dispositivos más visitados dentro de la categoría Smartphones.

En http://spark.apache.org/docs/latest/rdd-programming-guide.html puede encontrar la documentación de Spark para apoyar lo dado en clase.

Recomendamos utilizar para la resolución un Notebook de Jupyter Notebook que permita organizar claramente el output y comentar los hallazgos, de tal forma que puedan compartirse por ejemplo en Github.
Los ejercicios deben resolverse utilizando Apache Spark usando e RDDs.
A partir de Spark 2 se pueden leer CSV directamente con los readers integrados en Spark.
Para esto se debe obtener el sqlContext y luego hacer:

dataframe = sqlContext.read.csv('datos.csv')

Con eso se obtiene el dataframe de los datos, que es una abstracción por arriba de los RDD. Se puede obtener el rdd a partir del dataframe haciendo simplemente dataframe.rdd.
Documentación en http://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html

Atención: Es importante considerar en la resolución que no es posible traer toda la información al driver de spark desde el cluster. Es necesario poder reducir la información hasta un cierto punto que pueda traerse a la memoria del mismo, por lo que operaciones de sumarización, etc deben realizarse en el cluster.