Conceptos y aplicaciones en Big Data

Trabajo Práctico 2 – Modalidad de cursada presencial

Spark

Pautas generales

- La entrega consiste en la implementación de scripts de Spark, resolviendo todas las consignas presentes en este enunciado.
- Los alumnos pueden conformar grupo de no más de dos integrantes y hacer una única entrega grupal.
- La entrega se realiza por la mensajería del curso en IDEAS.
- La fecha límite de entrega es el 30 de noviembre de 2022.
- La calificación obtenida en este TP será tenida en cuenta en la nota final de la materia.

Enunciado

Una empresa de remises almacena la geoposición de cada uno de sus vehículos en cada uno de los viajes realizados. La ciudad por donde realizan los viajes los vehículos de la empresa es un cuadrado conformado por 100 avenidas verticales y 100 calles horizontales. Y la geolocalización se almacena cada vez que el vehículo llega a una intersección entre una avenida y una calle.

El dataset posee los siguientes atributos: (ID_vehículo, Avenida, Calle, Timestamp, Destino)

Un viaje comienza en una determinada esquina y finaliza en otra, la finalización del viaje está marcada con el lugar de destino.

Durante el trayecto se van registrando todas las esquinas por la cual transita el vehículo hasta llegar a su destino.

Ejemplo:

ID	Ave	Calle	Timestamp	Destino
905	28	28	8079	
905	28	27	8088	
905	27	27	8097	
905	26	27	8106	
905	25	27	8115	
905	24	27	8124	Escuela

NOTA: Se ha de suponer que los vehículos son “Big Data” y que cada vehículo puede hacer “Big Data” viajes.

NOTA 2: Si bien el archivo con el dataset tiene los viajes de cada vehículo “ordenados”, este orden solo debería tener propósito de depuración. Para la resolución de los ejercicios no se puede asumir ningún orden en el dataset.

1) Implemente un script de Spark que permita conocer cuántos viajes realizó cada vehículo. Recordar que un viaje es una serie de coordenadas que finalizan en un destino determinado.
   Ejemplo:

ID Vehículo	Cantidad de viajes
511	4
697	11
151	1
653	3

2) Implemente un script de Spark que permita conocer cual es el top 3 de los "tipos" de destinos más visitados.
   Los "tipos" de destino válidos son "Hospital", "Escuela", "Plaza", "Ferretería", "Farmacia", "Supermercado", "Museo", NO interesa contar a los destinos "Otro".
   Ejemplo:

Tipo de destino	Cantidad de visitas
Supermercado	316
Museo	197
Escuela	193

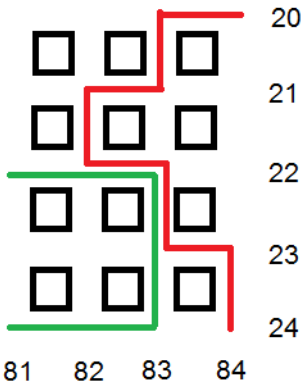
3) Implemente un script de Spark que permita conocer la cantidad de vehículos en movimiento por franja horaria. La duración de la franja horaria es un parámetro de la consulta.
   Ejemplo (franja horaria de duración 3000 (parámetro de la consulta)):

Franja horaria	Vehículos en movimiento
0 - 3000	843
3000 - 6000	673
6000 - 9000	449
9000 - 12000	248

4) Implemente un script de Spark que permita conocer cuál es el top 10 de las esquinas (avenida, calle) más transitadas por vehículos diferentes. En esta consulta cada vehículo cuenta como paso de una esquina una única vez, independientemente de que el mismo vehículo haya pasado por la misma esquina varias veces en diferentes viajes. Ejemplo:

Esquina	Cantidad de pasos (vehículos distintos)
(79,25)	38
(60,44)	37
(58,61)	33
(5,58)	33

5) Implemente un script de Spark que permita conocer la avenida y la calle más recorrida. La avenida (y también la calle) más recorrida es aquella por la que transitaron más vehículos en cualquiera de sus tramos. En esta consulta, cada vehículo puede sumar más de una vez si pasó por la misma cuadra varias veces.
Ejemplo:



La calle (horizontal) más recorrida es la 22 ya que se recorren tres cuadras (dos por el recorrido verde más uno por el rojo).
La avenida (vertical) más recorrida es la 83 ya que se recorren cuatro cuadras (dos por el verde y dos por el rojo).