Big Data Management

Assignment 2

Description

In this assignment your task is to prepare the batch layer (off-line processing pipeline) of the lambda architecture that will enable us to perform some analytics on a dataset. You are required to use Apache Spark's SQL API to compute some simple analytics.

Dataset

Transport Infrastructure Ireland (TII) (`https://www.tii.ie`) operates and maintains a network of traffic counters on the motorway, national primary and secondary road networks in Ireland. These traffic counters capture data on different parameters. There are currently around 400 of these counters active across the network. For an interactive view of the data they capture, go to the TII Traffic Counter Data Website: (`https://www.nratrafficdata.ie`). On this website, green dots display the individual traffic counter locations around the country. Summary traffic data information can be obtained from each site by clicking on the green dot. Upon clicking a dot, a pop-up window will appear that summarises the traffic data and provides a link to more detailed data. You can click on the **list of sites** button to view description of each counter.

The traffic counter data set is a valuable source of information on vehicle movements across the national road network and is made available publicly in its raw form, in order to provide researchers, public bodies, engineering companies, as well as the general public, with the opportunity to analyse and query the data independently for their own specific purposes. The first row of each file contains headers which describe each field. However, the meaning of some of these may not be apparent to consumers. The following explains some of the less obvious column headers:

cosit: The unique identifier for the traffic counter device. In conjunction with the site's dataset, this can be used to determine the location and route of the counter, used to record the vehicle movement.

lane: The Id of the lane in which the movement was recorded, which is specific to each counter.

straddlelane: If a value is present, this indicates that the vehicle may have been changing lanes as it passed over the counter.

class/classname: This indicates the category of vehicle that was recorded e.g. car, bus, etc.

length: The approximate length of the vehicle recorded.

headway: The approximate distance between the front of the recorded vehicle and the vehicle behind.

gap: The approximate distance between the rear of the vehicle and the front of the vehicle behind.

weight: This is available on (Weigh-in-Motion) WIM sites only and indicates the approximate weight of the vehicle.

temperature: If available, this indicates the approximate surface temperature of the road at the location of the device.

numberofaxles: This is available on WIM sites only and indicates the number of axles detected for the vehicle.

axleweights: This is available on WIM sites only and expresses as an array of real numbers, the weight over each axel in order.

axlespacing: This is available on WIM sites only and expresses as an array of real numbers, the distance between each of the axles.

Setting Up

Follow the Cassandra setup instructions provided on Moodle. You can also follow:

Official guidelines on setting up Apache Cassandra are available at:
`http://cassandra.apache.org/download/`
To configure Apache Cassandra to work with Apache Spark, see:
`https://github.com/datastax/spark-cassandra-connector`

Questions

You answer similar questions as in assignment 1 but this time you are required to use Apache Spark's SQL API. Prepare Cassandra structures and the Spark code that saves the computed batch views into these structures:

1. Show the busiest site in Ireland.

2. Show the average distance between vehicles on all M50 sites.

3. What site has recorded the highest temperature? Show the hour of the day.

4. Show total number of WIM sites available in the dataset?

5. Compute the average speed for each site on M50.

6. Show total number of counts by vehicle class. Order results in descending.

7. List the top 3 busiest sites on M50.

8. What is the busiest site on M6?

9. What site reports the highest number of HGVs?

10. Calculate the total number of vehicles on each site on M7.

Submission

- Submit your solution on Moodle before the deadline.

- Acceptable file format: Python notebook - name it `assignment2.ipynb`. The notebook should be exported as iPython Notebook with *.ipynb extension. If the code in your notebook does not run, it will result in 20% penalty.

- Take two screenshots of your solution to each question (spark code + its output into a Cassandra table) and insert it in a word document, generate a pdf of this document.

- Zip both files together and submit your solution on Moodle by the deadline.

- Do not submit work that's not your own and do not let others copy work that is your own. Both Copyier and Copyee will get ZERO marks.