Big Data Management

Assignment

# Description

In this assignment your task is to prepare the batch layer (o pro- cessing pipeline) of the lambda architecture that will enable us to per- form some analytics on a dataset. You are required to use **Apache Spark’s SQL API** to perform the required computation and save the results to the Cassandra structures.

# Dataset

Transport Infrastructure Ireland (TII) (https://www.tii.ie) operates and maintains a network of tra and secondary road networks in Ireland. These tra data on diparameters. There are currently around counters active across the network. For an interactive view of the data they capture, go to the TII Trahttps://[www.](http://www/) nratrafficdata.ie). On this website, green dots display the individual tra counter locations around the country. Summary tra data infor- mation 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 tra 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 tra hicle movements across the national road network and is made available publicly in its raw form, in order to provide researchers, public bodies, en- gineering companies, as well as the general public, with the opportunity to analyse and query the data independently for their own speci The row of each contains headers which describe each How- ever, 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.

# Questions

You will use the same data as for assignment You answer similar ques- tions as in assignment Spark’s SQL API. Prepare Cassandra structures and the Spark code that saves the computed batch views into these structures:

Calculate the usage of Irish road network in terms of percentage grouped by vehicle category.

Calculate the highest and lowest hourly on M - show the hours and total number of vehicle counts.

Calculate the evening and morning rush hours on M - show the hours and the total counts.

Calculate average speed between each junction on M (e.g., junction

- junction junction - junction etc.).

Calculate the top locations with highest number of counts of HGVs (class). Map the COSITs with their names given on the map.