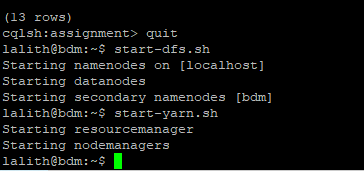
**BDM Assignment 2**

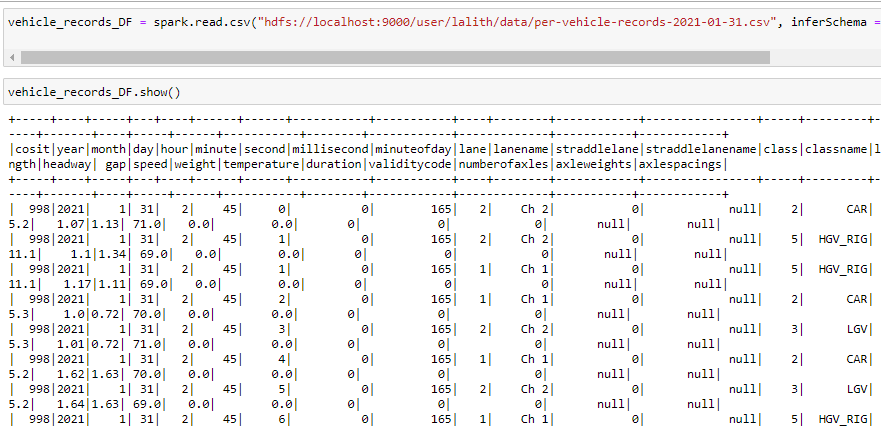
The task of this assignment is to prepare the batch layer (off-line processing pipeline) of the lambda architecture that will enable to perform some analytics on a dataset.

We answer the same questions which we had answered in Assignment 1 using Data frame API. But in this Assignment 2 we'll answer these questions using Apache Spark's SQL API. We had defined Cassandra structures and the Spark code that saves the computed batch views into these structures.

**Loading the csv file from the Hdfs file system**

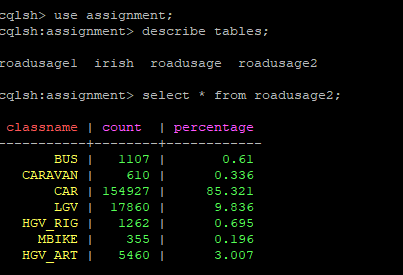
Loading the csv file from hdfs requires the hadoop services to be started.





**1) Calculating the usage of Irish road network in terms of percentage by vehicle category**



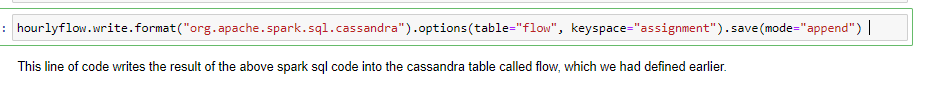


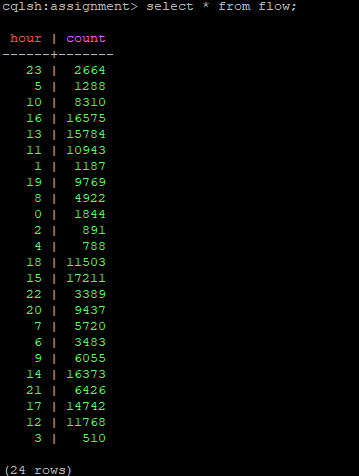
The above codes writes the data to this defined Cassandra table.

**2) Calculating the highest and lowest hourly fows on M50 - show the hours and total number of vehicle counts**

****

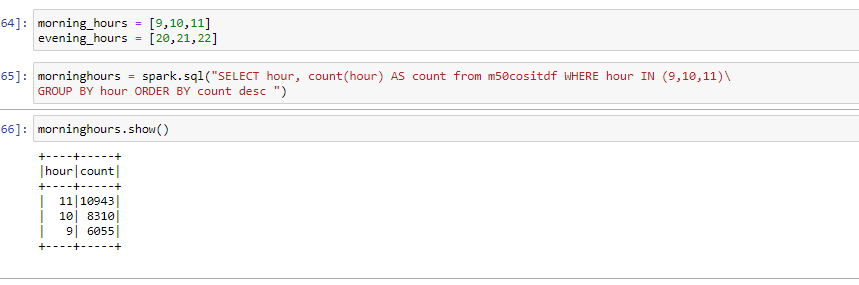
The above query gives the count of vehicles per hour, we can observe that hours 13 -17 has highest uage, while hours 1 - 5 has least usage



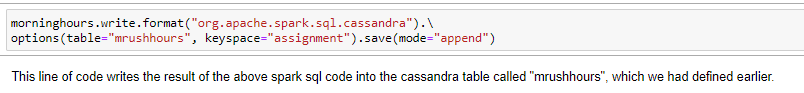


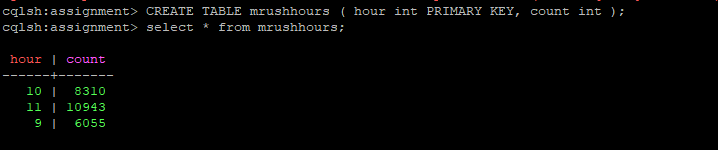
This line of code writes the result of the above spark sql code into the cassandra table called flow, which we had defined earlier.

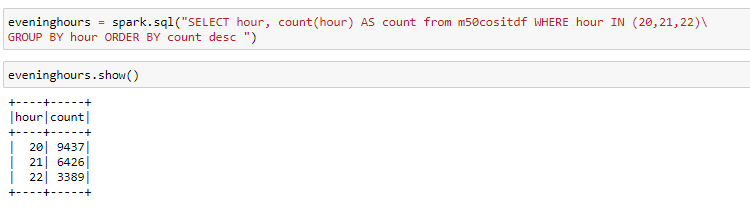
**3) Calculating the evening and morning rush hours on M50 - show the hours and the total counts**

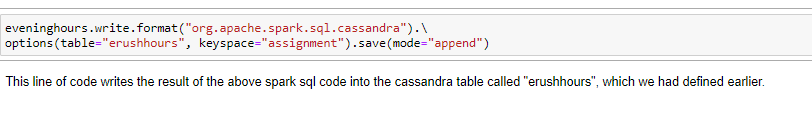


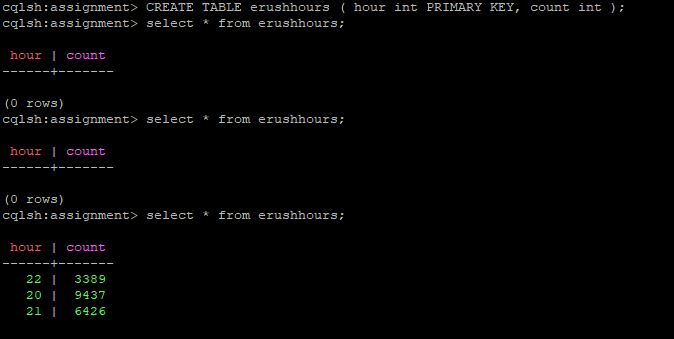
The above query identifies the morning rush hours and the count of vehicles per that hour. We can observe that morning rush hours are between 9 -12.









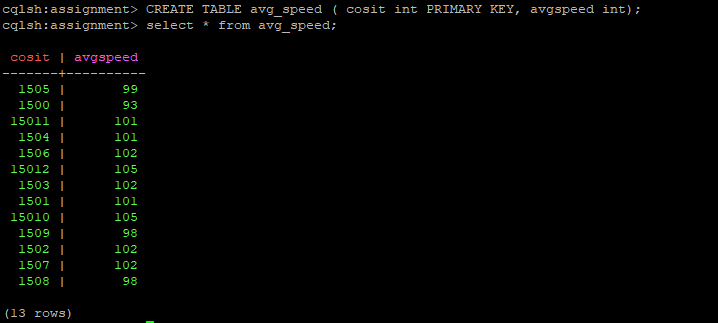


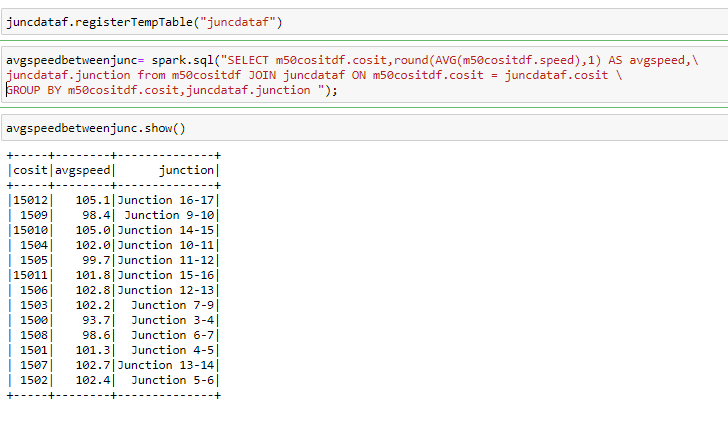
This line of code writes the result of the above spark sql code into the cassandra table called "erushhours", which we had defined earlier.

**4) Calculating the average speed between each junctions on M50**

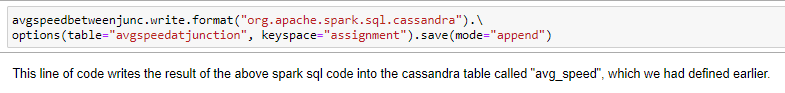
****

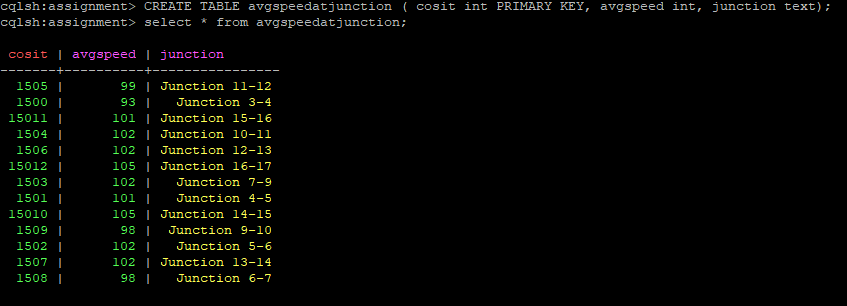
The above the query gives the average vehicle speed recorded at each counter.



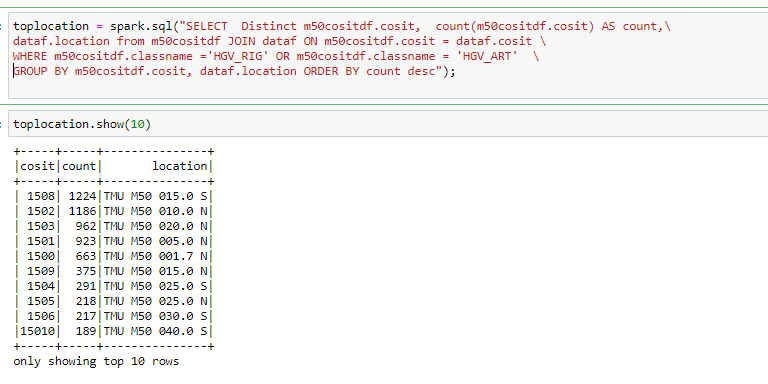


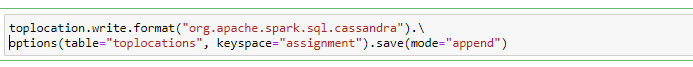
This is a refined output for this question, I have added the junction numbers at which the respective counter is located and the percentage is rounded to 1 decimal.

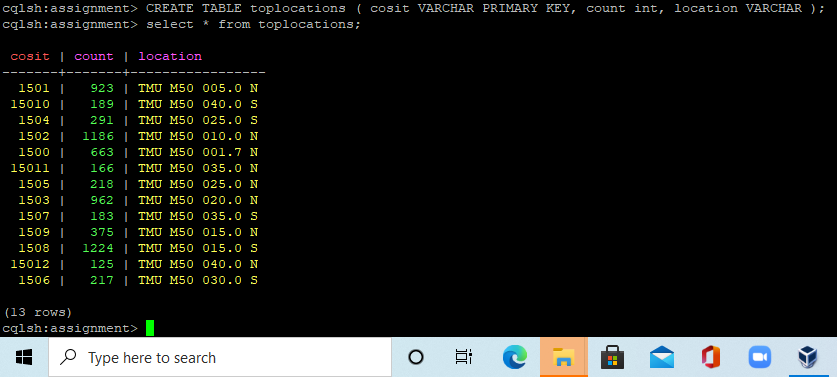




**5) Calculating the top 10 locations with highest number of counts of HGVs(class)and Mapping the COSITs with their names given on the map**







We have joined the two temporary tables to the get the location name next to the cosit, The above query gives the top 10 locations and its cosits with highest counts of vehicles class "HGV".

This line of code writes the result of the above spark sql code into the cassandra table called "toplocations", which we had defined earlier.