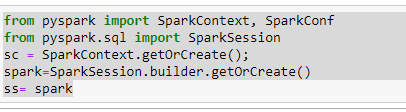
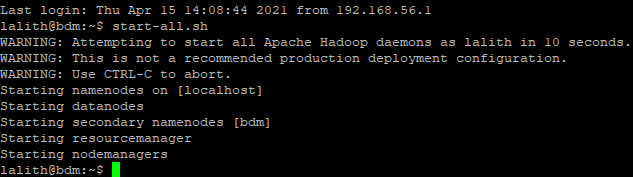
**BDM Assignment 1**

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.

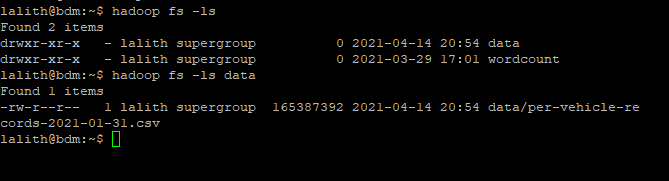
Importing all sparkcontext, sparkconf,sparksession from pyspark.

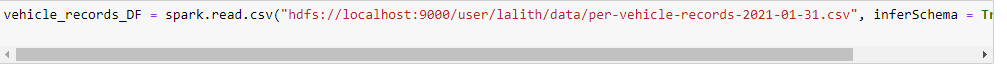


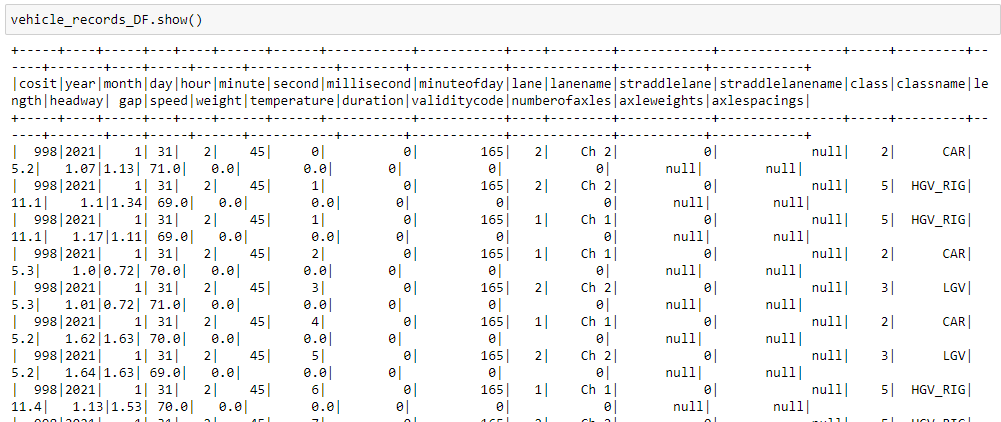
**Loading the csv file from hdfs requires the hadoop services to be started.**



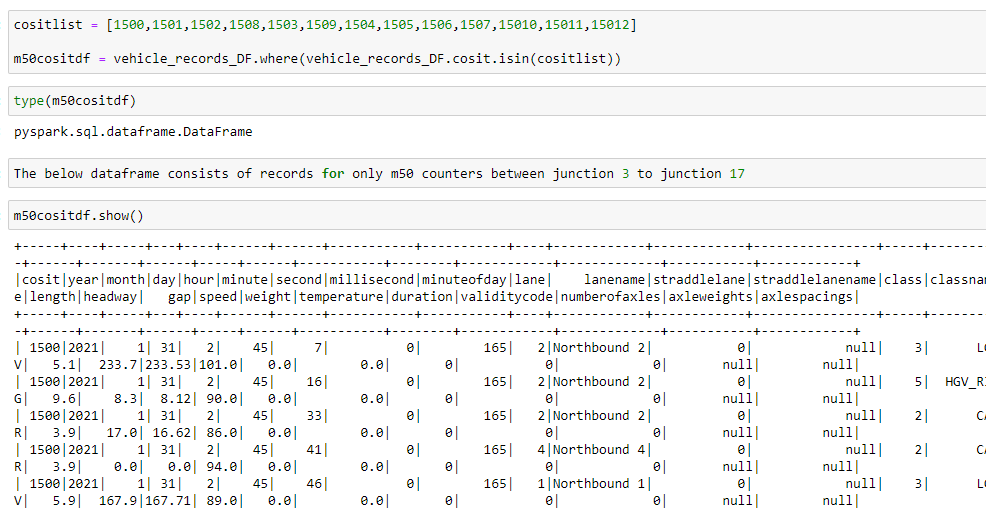
1. **Loading the csv file from hdfs to spark dataframe**

****

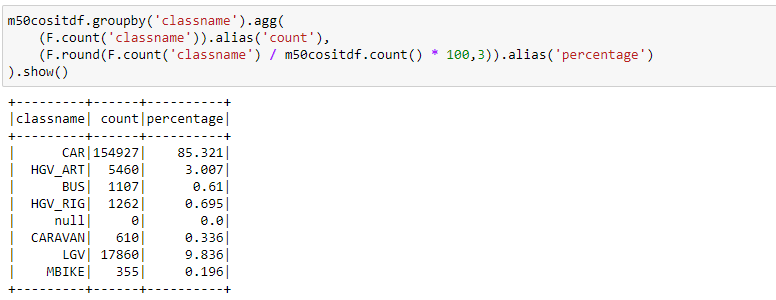
****

****

**Filtering and creating a new dataframe for M50 highway cosits from junction 3 and junction 17**

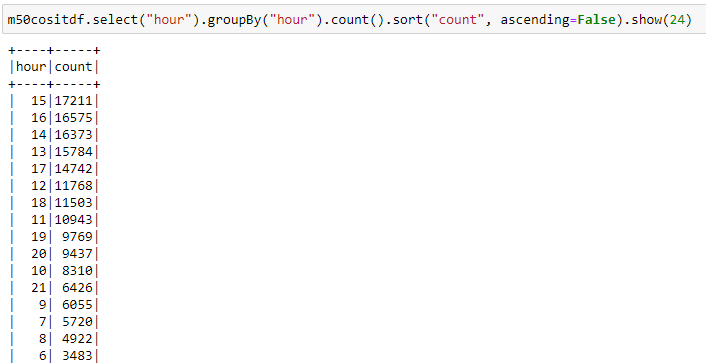
****

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

****

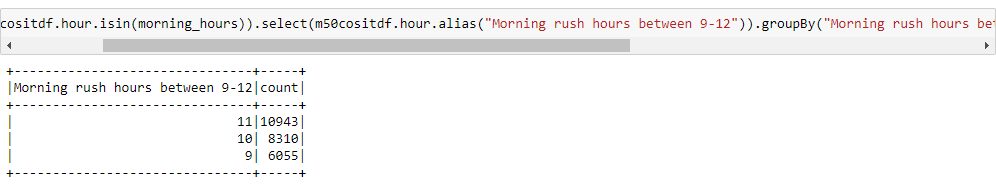
**From the above query we can find the usage of irish road network by each vehicle class. From the result we can observe that classname CAR has the highest uasge with percentage of 85.321**

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

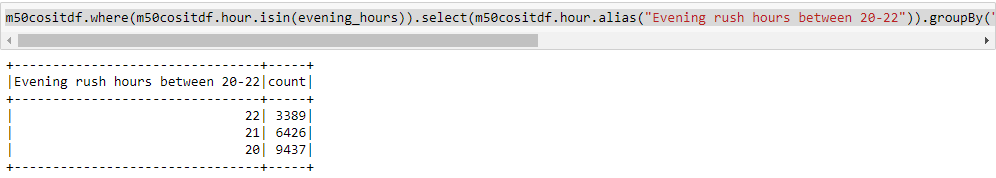
****

**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**

1. **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.

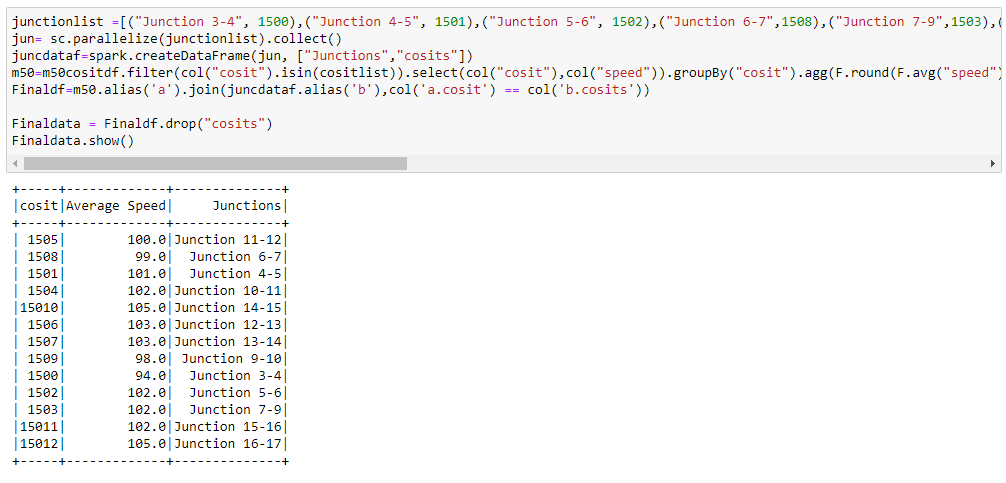
****

The above query identifies the Evening rush hours and the count of vehicles per that hour. We can observe that evening rush hours are between 20 -22

1. **Calculating the average speed between each junctions on M50**

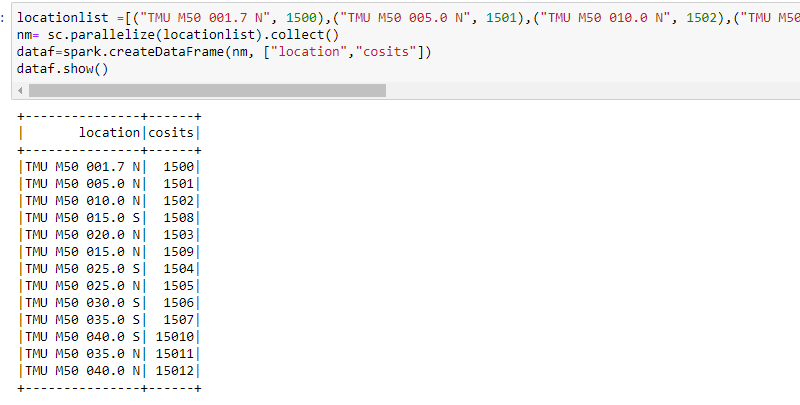
****

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

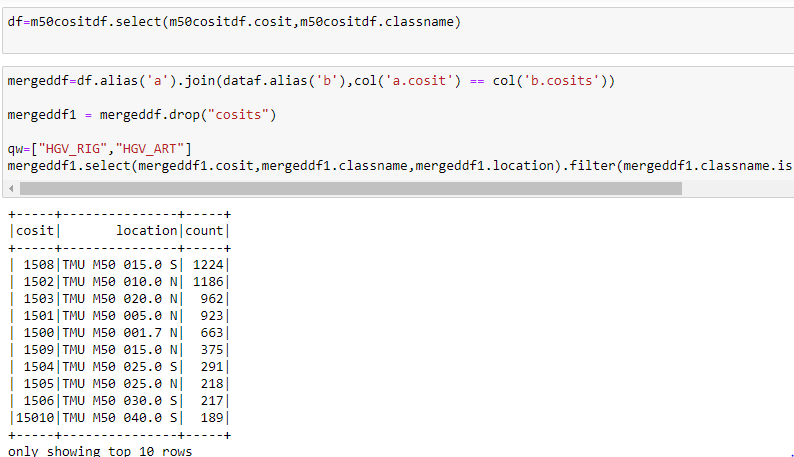
****

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**.**

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

****

I have a created a list of location names and its cosits and converted it to a dataframe.



I have joined the two data frames to the get the location name next to the cosits, The above query gives the top 10 locations and its cosits with highest counts of vehicles.