**Zeppelin-Spark Assignment Big Data 1**

This assignment is based on some data for worldwide sales that has been given to you to provide some analysis on. The object of the exercise is that you will use Zeppelin and HDFS to ingest this data and query it using spark basic scala commands and SQL .

The Customer who has given you this data would like a Zeppelin notebook returned with the following breakdown . Each item should be represented by a paragraph in the notebook. Also please use the %md to describe the activity that you are going to do in the next paragraph as you will need to do a screen shot of every activity. Place your screenshots on a document with your name and student number prior to submitting. The data provided is call worldsales.csv

1. Show how you loaded the data into a dataframe
2. Show the dataframe
3. Print the dataframe schema
4. Filter the dataframe to show units sold greater than 8000 and unit cost greater than 500
5. Show the dataframe in group by “Region” and count
6. Create a separate dataframe with the group by results
7. Save this new subset dataframe as a csv file into HDFS – make sure its is saved as a single file in HDFS
8. Create two views using the “createOrReplaceTempView” command
   1. View on “Salesview” from the first dataframe
   2. View on “Regionview” from the second dataframe
9. Using SQL select all from “Regionview” view and show in a line graph .
10. Using SQL select from the “Salesview” view – the region and sum of units sold and group by region and display in a data grid view
11. Using SQL select from the “Salesview” view – the region and sum of total\_profit and group by region and display in a Bar chart
12. Using SQL select from the “Salesview” view – show the total profit as profit, the total revenue as revenue and the total cost as cost from “Salesview” group by region – The client wants to see this data in a line graph so as to see the correlation between cost ,revenue ,profit between regions.
13. The customer is in the process of opening up a new store an they are looking at the best location to do so, they need to see the avg profit in each region as a percentage (pie chart) compared to other regions, please use both views created to demonstrate answer also point out the region where it is most profitable.