## Project Practice – Hive and Spark

**Hive practice**

Story: Downloading the data, running some code to move data to HDFS.

Places to get data:

NYC public open Data

Kaggle

Google Datasets

There scripts that went to NYC public data written in Python, that downloaded the csv. And then I uploaded the CSV to HDFS.

Aim: The goal of this project is to find out who earns the highest in NYC using NYC public data and use HIVE and save the results on HDFS.

1. Go to NYC Public data in the link and download the csv (comma separated values) files
2. csv files are comma seperated. Clean them using excel in Ubuntu and have the columns we ned also rename the columns
3. Using Openlib (MS excel of ubuntu) remove all the unwated rows so that you have profession and salary in the columns
4. After you have cleaned the data into columns that you want then give names to column and import the data
5. After importing the data run group by and find max salary grouped by profession

Download the csv from:

Remove all the columns and rename the columns in the excel as:

Job ID Agency Business Title Civil Service Title Salary Range From Salary Range To Division/Work Unit

Create table in Hive:

create table NYC(JobID INT, Agency STRING, BusinessTitle STRING, CivilServiceTitle STRING,SalaryRangeFrom INT,SalaryRangeTo INT,DivisionWorkUnit STRING) row format delimited fields terminated by ',' stored as textfile;

Now load the data:

load data local inpath '/home/joshi/Downloads/NYC.csv' into table NYC;

Check the data:

select \* from NYC;

select \* from NYC limit 5;

select state ,COUNT(\*) from students group by state ;

Run Aggregations:

select BusinessTitle,avg(SalaryRangeFrom) from NYC group by BusinessTitle;

select BusinessTitle,avg(SalaryRangeFrom) from NYC group by BusinessTitle order by avg(SalaryRangeFrom) limit 5 ;

select BusinessTitle,avg(SalaryRangeFrom) from NYC group by BusinessTitle order by BusinessTitle limit 5 ;

select BusinessTitle,avg(SalaryRangeFrom) as avgs from NYC group by BusinessTitle order by avgs limit 5 ;

select BusinessTitle,avg(SalaryRangeFrom) as avgs from NYC group by BusinessTitle order by avgs desc limit 5 ;

**Spark Practice / Pyspark**

**Part 1**

The goal of this project is to use Spark to run map reduce with different delimeters

Run the spark command with any news page from internet

The default word count with me similar to other map reduce

Run the count

Refer to the Spark code what we did in the class:

<https://docs.google.com/document/d/1lY4XUAVjfdJXfnqrwqTRnb8TdNXbQ-r5Ah6_RzQYVog/edit?usp=sharing>

Sample code:

val inputfile = sc.textFile("file:///home/hadoopuser/input2.txt")

val counts = inputfile.flatMap(line => line.split(" ")).map(word => (word,1)).reduceByKey(\_+\_);

counts.saveAsTextFile("file:///home/hadoopuser/output2")

**Part 2**

Import the simple lines

Change the delimiter to commas and redo the same process

Create a input.txt file and have space as the delimiter and run the code below.

val inputfile = sc.textFile("file:///home/hadoopuser/input2.txt")

val counts = inputfile.flatMap(line => line.split(" ")).map(word => (word,

1)).reduceByKey(\_+\_);

counts.saveAsTextFile("file:///home/hadoopuser/output2")

**PySpark Practice**

The goal of this project is to run python commands and lambda commands and create data frames

Make sure we have sparksession alive and then we can create data frames

Also we need to learn how to import the data

Reference codes from the slack that were given to you.

Using Databricks run some practice code.

**SCOOP to move mysql to hdfs**

Attendees

Move into HDFS