JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

Non-Graded Assignment

PIG – 2

01.

Write a UDF function in Pig to reverse the sequence of given last name and first name. For example if the name passed is *Agarwal,Amit* then the UDF should return *Amit Agarwal* as the return value.

You are given a file Name-Data.txt. Write a UDF *reverseName* and use it a Pig script to reverse all the names in the input file.

02.

For all Open Georgia Salary/Travel data provided as CSV file with this assignment for the Fiscal Year 2010 and Organization Type of Local Boards of Education, produce a distinct list of all Job Titles along with the total number of employees aligned with each Job Title & the minimum/maximum/average salaries for each of the identified Job Titles.

Expected Steps:

1. Store the given input file salaryTravelReport.csv into the HDFS Location and save the given piggybank.jar file in your disk.
2. Load up the base UDF library – piggybank and get a handle on the REPLACE function.
    - This jar file `piggybank.jar` contains a class `org.apache.pig.piggybank.evaluation.string.REPLACE` which can replace any substring from the given string.
    - Open the grunt console and execute the following commands:

```
register "<Path>/piggybank.jar";
define REPLACE
org.apache.pig.piggybank.evaluation.string.REPLACE();
```

3. Create your Pig commands and execute in the console, hints are given below:
    - Load the salary file and declare its structure
    - Loop through the input data to clean up the number fields. Take out the commas from the salary and travel fields and cast to a float
    - Trim down to just Local Boards of Education
    - Further trim it down to just be for the year in question
    - Bucket them up by the job title
    - Loop through the titles and check how many are there under each title
    - Determine the minimum, maximum and average salaries for every title
    - Guarantee the order on the way out
    - Dump the results on the console
    - Save results back to HDFS

03.

Airline Data Analysis with Pig: Travel industry generates huge amounts of data every day. Refer Airline_data_schema for the Table description.

Write Pig scripts for the following Queries:

Question 3A: Find out top 20 airports by total volume of flights. What are the busiest airports by total flight traffic? JFK will feature, but what are the others? For each airport code compute the number of inbound, outbound and all flights. Compute the top 20 airports per month per year based on total traffic (inbound + outbound).

Question 3B: Carrier Popularity – Some carriers come and go, others demonstrate regular growth. Compute the (log base 10) volume -- total flights -- over each year, by each carrier. The carriers are ranked by their median volume (over the 10-year span).