Exercise, Pig 2

This Exercise will go through more complex concepts within Pig to calculate values using flight information for 2015

Login to your sandbox.

Step 1 Pull down the repo at the following location into your sandbox

```
$git clone https://github.com/redgianttx/BUAN6346.git
```

Step 2 Placing the data

Run:

```
$ cd BUAN6346/Pig/PigExercise2/
```

Then:

```
$ unzip data.zip
```

Next Place the data into Hadoop by doing the following

```
$hadoop fs -mkdir /data
$hadoop fs -put airports.csv /data
$hadoop fs -put flights.csv /data
```

Step 3 A pig script stub has been provided that contains the general framework.

Open the file **flight_delay_exercise.pig** using VIM

Notice the load statements and the schemas defined.

Step 4 Filter airports to those only in Texas

Using FILTER, enter the following after airports_filtered

```
airports_filtered = FILTER airports BY STATE==‘TX’
```

Step 5 Next we'll calculate the flight departure delay. Notice the **flight_with_delay** has already been defined with question marks at the end as place holders for the fields DEPARTURE_DELAY, ARRIVAL_DELAY.

Add the necessary projection in the place holders to calculate the delay

Step 6 Next we need to filter out null values in calculated DEPARTURE DELAY (This will be important later)

Add the filter function to relation that filters out any rows that have null for DEPARTURE_DELAY

```
FILTER flight_with_delay BY DEPARTURE_DELAY is not null;
```

Step 7 We want to calculate based upon YEAR, MONTH, ORIGIN_AIRPORT, and AIRLINE

Write the **group by** statement for relation **flight_airport_departure** to do this

Step 8 Next we'll use a nested FOREACH and non-linear processing to calculate or derived data set. First let's calculate the average DEPARTURE_DELAY

```
average_delay = AVG(flight_airport_departure.DEPARTURE_DELAY);
```

What is the function **AVG?**

Why did we have to put 'flight_airport_departure'?

Step 9 Next, We'd like to calculate the number of delayed flights for each group that are over 20 minutes in delays

First let's limit the data set. Finish the relation **limited** with the proper **FILTER** statement

Next, Finish the **count** relation and count the number of in the relation **limited**

Step 10 Next, to demonstrate the non-linear data creation, finish the relations **max** and **min** to calculate the max and min DEPARTURE_DELAY. Hint: Look at step 8.

Step 11 Finally, notice the generate statement is already provided.

What does the projection **group** reference?

What is the significance of **FLATTEN?** What if we didn't use **FLATTEN**?

Step 12 Replicated Join

We want to now add airport information to the flight delay data. We'll do this through a join.

Finish the **joined_data** relation as follows:

```
joined_data = JOIN summary_data BY ORIGIN_AIRPORT,
airports_filtered BY IATA_CODE using 'replicated';
```

Remember, airports have been filtered to only include airports in Texas. What happens to the rows in **summary_data** that are not for airports in Texas after the join?

What does the replicated join do?

Step 13 Finally, let's sort the data for output.

Finish the relation **final** using **ORDER** and sort **ORGIN_AIRPORT, YEAR, MONTH**

Step 14 Save and Exit your file.

Run your script using

```
$pig flight_delay_exercise.pig
```

Step 15 Parameter Substitution

What if we wanted to run this for delay times other than 20 minutes?

Create a copy of your script file

```
$cp flight_delay_exercise.pig flight_delay_exercise2.pig
```

Open the script file **flight_delay_exercise2.pig** in **VIM**

Replace the value of 20 in the nested for each with **$DELAY**

Next replace the dump line with

```
store final into '/data/${DELAY}_final' using PigStorage(‘,’)
```

Save and Exit your file

Step 16 Run the script as:

```
$pig -p DELAY=30 flight_delay_exercise2.pig
```

After it completes

Check that the folder **30_final** exists in the folder **/data** on HDFS.

Run the following:

```
$hadoop fs -cat /data/30_final/part*
```

Run the program again and change the delay to 40.

Step 17 UDFs

Create a copy of the flight_delay_exercise2.pig script

```
$ cp flight_delay_exercise2.pig flight_delay_exercise3.pig
```

Add the following to the top of flight_delay_exercise3.pig in VIM

```
register datafu.jar

define define Median datafu.pig.stats.StreamingMedian();
```

In the **foreach**

Add a relation called **median** that uses the UDF to calculate the median for the DEPARTURE_DELAY.

Add **median** to the generate within the **foreach**

Save and Exit.

Run the script as follows

```
pig -p DELAY=10 flight_delay_exercise2.pig
```

After it completes, run
$hadoop fs -cat /data/10_final/part*

There should be an extra column indicating the median.

Do you notice anything about this column that is different than the others?
How do you get it to look like the other columns?