Rohaan Advani – 111903151

BDA: Assignment 1: Analyze a Dataset using Apache Pig

The Dataset selected is that of flight data. We will use this data in order to determine the following:

1. In which month flights are cancelled the most.
2. Top 10 maximum diversions.
3. Top 5 destinations visited.

Dataset has following columns:

1. year
2. month
3. flightno
4. origin
5. dest
6. cancelled
7. cancellation code
8. diverted

Pig Latin Script:

```
file = load '/Downloads/flightdata.csv' using PigStorage(',') as (year:int, month:int, flight_no:int,
origin:chararray, destination:chararray, cancelled:int, cancellation_code:chararray, diversion:int);

cancelled = filter file by cancelled == 1 AND cancellation_code == 'B';
groupbymonth = group cancelled by month;
total = foreach groupbymonth generate group as month, COUNT(cancelled)as totalno;
ordered_total = order total by totalno DESC;
most_cancelled = limit ordered_total 1;
STORE most_cancelled INTO '/Downloads/most_cancelled_flights';

filt = filter file by diversion == 1;
groupbyroute = group filt by (origin, destination);
relation = foreach groupbyroute generate group.origin as origin, group.destination as destination,
COUNT(filt.diversion) as diversion;
ordered_relation = order relation by diversion DESC;
result = limit ordered_relation 10;
store result into '/Downloads/top_10_diversions';

groupbydest = group file by destination;
relation2 = foreach groupbydest generate group as destination,COUNT(file.destination) as
frequency;
ordered_relation2 = order relation2 by frequency DESC;
result = limit ordered_relation2 5;
store result into '/Downloads/top_5_dest';
```