Rohaan Advani – 111903151

BDA: Assignment 2: Perform Data Querying and Analysis using Apache Hive:

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

1. Which month has seen the greatest number of cancellation due to bad weather?
2. Top 10 routes that has seen maximum diversions.
3. Top 5 visited destinations.

Dataset has following columns:

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

Hive Script:

```
create table aviation (year INT, month INT, flight_num INT, origin STRING, dest STRING, cancelled INT, cancel_code INT, diversion INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

load data inpath "/user/cloudera/pig_avaitation_output_hive/" into table aviation;

SELECT month,COUNT(canceled) as t FROM aviation
WHERE canceled = 1 AND canel_code = ‘B’
GROUP BY month
ORDER BY t DESC
LIMIT 1;

SELECT origin,dest,COUNT(diversion) as t FROM aviation
WHERE diversion = 1
GROUP BY origin,dest
ORDER BY t DESC
LIMIT 10;

SELECT dest,COUNT(dest) as x FROM aviation
GROUP BY dest
ORDER BY x DESC
LIMIT 5;
```