Aviation Data Analysis Using Apache Hive

First cleanse the data using pig:

REGISTER '/home/training/Desktop/pig/piggybank-0.11.0.jar';

A = load '/home/training/Desktop/Hive/Aviation\_HIVE.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage as (year, month, flight\_num, origin, dest, cancel

**Loading pre-processed data from pig to hive using HCatalog.**

**Create an external table:**

CREATE EXTERNAL TABLE AVIATION (Year int, Month int, **FlightNum** int, Origin string, Dest string, **Cancelled** int, CancellationCode int, diversion int)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION ‘/home/training/Desktop’;

Load data directly from Pig to HIVE table – Aviation:

grunt> STORE A INTO 'hdfs://localhost:54310/home/training/Desktop' USING PigStorage(',');

**LOAD DATA FROM HDFS TO HIVE TABLE;**

### PROBLEM STATEMENT 1:

Which month have seen the most number of cancellation due to bad weather?

SELECT month,COUNT(canceled) as t FROM aviation

WHERE canceled = 1 AND canel\_code = ‘B’

GROUP BY month

ORDER BY t DESC

LIMIT 1;

***Top 10 route(origin and dest) that has seen maximum diversions?***

Select Origin, Dest, count(Diversion) as cnt from aviation where Diversion = 1 order by cnt desc limit 10;

***Top 5 visited destination.***

Select Dest, count(Dest) as cnt from aviation group by Dest order by cnt desc limit 5;