Aviation data analysis

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights appears in DOT's monthly Air Travel Consumer Report, published about 30 days after the month's end, as well as in summary tables posted on this website. Summary statistics and raw data are made available to the public at the time the Air Travel Consumer Report is released.
You can download the datasets from the following links:
Delayed_Flights.csv

Delayed_Flights.csv Datasets
There are 29 columns in this dataset. Some of them have been mentioned below:
• Year: 1987 – 2008
• Month: 1 – 12
• FlightNum: Flight number
• Canceled: Was the flight canceled?
• CancelleationCode: The reason for cancellation.

For complete details, refer to this link.
Problem Statement 1
Find out the top 5 most visited destinations.
Problem Statement 2
Which month has seen the most number of cancellations due to bad weather?
Problem Statement 3
Which route (origin & destination) has seen the maximum diversion?

a)Loading raw data into the root directory of HDFS.

Command: hadoop fs -put /home/acadgild/DelayedFlights.csv /

b)Pre-processing using Pig.
Loading pre-processed data from pig to hive using HCatalog.

c)start hive metastore service before loading data using HCatalog.
Command: hive –service metastore

```
hive --service metastore
```

d)create a hive table with the same schema as had pre-processed in the Pig.

```
hive> create table aviation(
    > year INT,
    > month INT,
    > flight_num INT,
    > origin STRING,
    > destination STRING,
    > cancelled INT,
    > cancel_code INT,
    > diversion INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 3.997 seconds
```

e)command in pig grunt shell to load the data to hive.
Syntax: STORE *relation_name* INTO *'hive_table'* USING org.apache.hive.hcatalog.pig .HCatStorer();

```
grunt> STORE C INTO 'aviation' USING org.apache.hive.hcatalog.pig.HCatStorer();
```

f)This will load the data into hive table which we had already created.
You can cross check the same using *SELECT * FROM aviation;* in hive shell.

```
hive> select * from aviation LIMIT 10;
```

PROBLEM STATEMENT 1:

Top 5 visited destination.

Source Code:

SELECT dest,COUNT(dest) as x FROM aviation

GROUP BY dest

ORDER BY x DESC

LIMIT 5;

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

```
ORD     108984
ATL     106898
DFW     70657
DEN     63003
LAX     59969
```

PROBLEM STATEMENT 2:

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

Source Code:

SELECT month,COUNT(canceled) as t FROM aviation

WHERE canceled = 1 AND canel_code = 'B'

GROUP BY month

ORDER BY t DESC

LIMIT 1;

```
hive> SELECT month,COUNT(cancelled) as t FROM aviation
    > WHERE cancelled = 1 AND cancel_code = 'B'
    > GROUP BY month
    > ORDER BY t DESC
    > LIMIT 1;
OK
12      250
```

PROBLEM STATEMENT 3:

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

Source Code:

SELECT origin,dest,COUNT(diversion) as t FROM aviation

WHERE diversion = 1

GROUP BY origin,dest

ORDER BY t DESC

LIMIT 1;

```
ORD     LGA     39
```