**Importing CSV Files into Hive**

**Hive**:

Apache Hive is an open-source data warehouse system for querying and analysing large datasets stored in Hadoop files. Hadoop is a framework for handling large datasets in a distributed computing environment.

**CSV File Format:**

CSV stands for Comma Separated Values. It consists of a data that is separated by ‘,’.

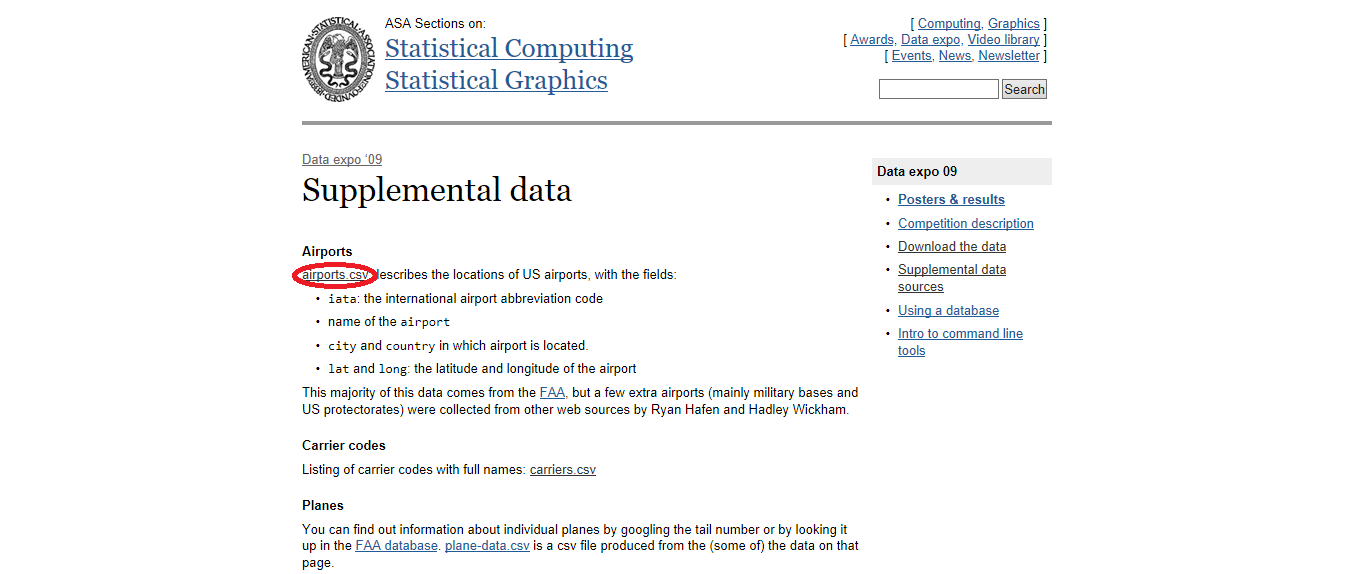
**Sample CSV data:**

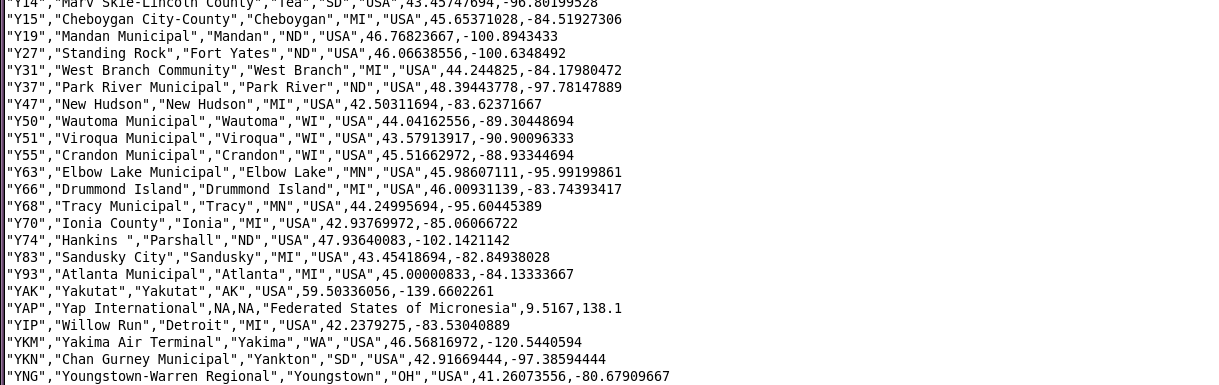
A, B, C, D, E, F

**Procedure:**

**Step 1:**

Download the required CSV files.

(I have used a file downloaded from http://stat-computing.org/ dataexpo/2009/supplemental-data.html)

****

**Step 2:**

Move the dataset into hive using the following command.

Command:

Hadoop fs –put Downloads/airports.csv /user/CSV

**Step 3:**

Open Hive using the following command.

Command:

Hive

**Step 4:**

Create a Database if needed using the following command.

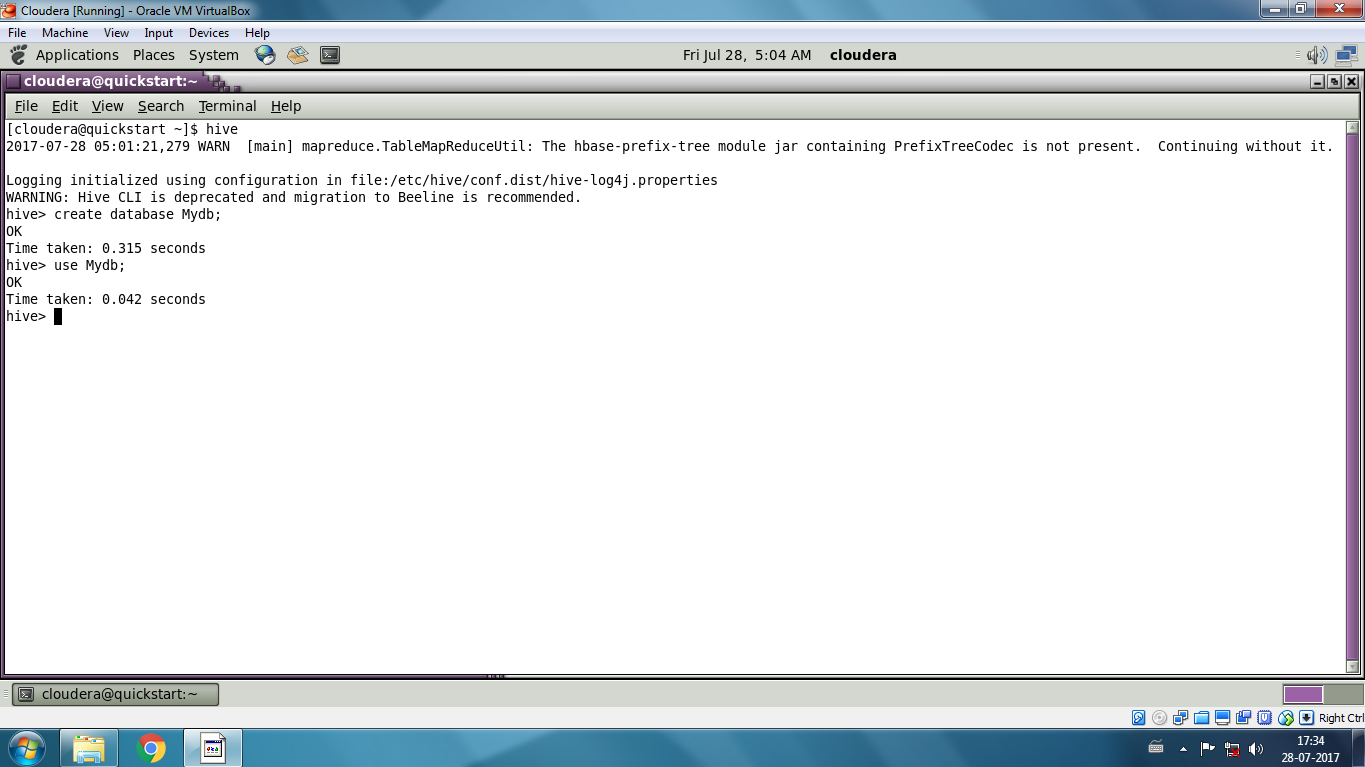
Command:

Create database <dbname>;

**Step 5:**

Use the database using the following command.

Command:

useMydb;

**Step 6:**

Create an external table with a schema that matches the data in the CSV file.

The dataset that I have used consists of the following data schema.

Id -> String

Airport name ->String

City ->String

State ->String

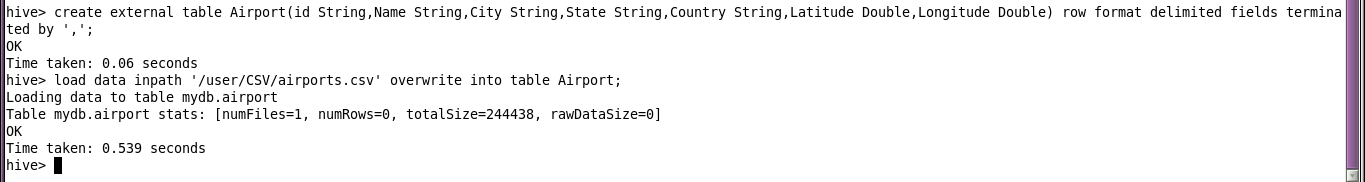
Country ->String

Latitude ->Double

Longitude ->Double

The schema for the dataset is created using the following command

Command:

Create external table Airport (id String, Name String, City String, State String, Country String, Latitude Double, Longitude Double) row format delimited fields terminated by ‘,’;

**Step 7:**

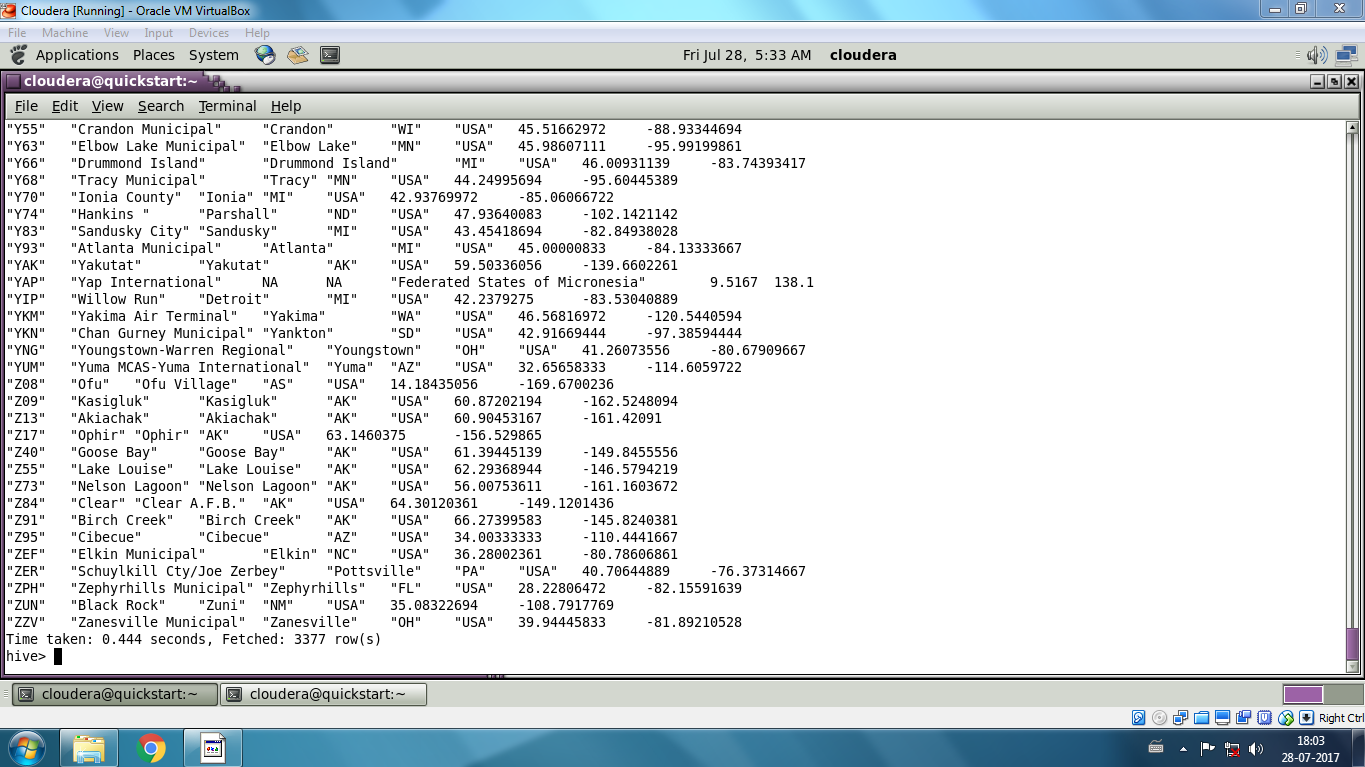
Load data into the table using the command.

Command:

Load data inpath ‘/user/CSV/airports.csv’ overwrite into table Airport;

**Step 8:**  
 The data in the table can be viewed using the following command.

Command:

Select \* from Airport; 

(or)

Select distinct city from Airport;