**Data Processing**

Pig commands:

Data Cleaning using pig:

* Create a cluster in GCP and open ssh.
* Upload the csv files.
* Replace new lines and format file: sed ':a;N;$!ba;s/\n/ /g' cleanedairlines.csv> cleanedairlines\_4dec.csv
* Create directory in Hadoop: hadoop fs –mkdir airlines
* Put files into newly created directory: hadoop fs -put cleanedairlines\_4dec.csv /airlines
* Go to pig using- pig to do ETL operations.
* tmp = LOAD '/ airlines' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES\_MULTILINE', 'NOCHANGE', 'SKIP\_INPUT\_HEADER') as (airline\_name:chararray, title:chararray, author:chararray, author\_country:chararray, flowm\_date:chararray, content:chararray, date\_visit:chararray, type\_traveller:chararray, overall\_rating :float, seat\_rating:float, staff\_rating:float, food\_rating:float, entertainment\_rating:float, value\_money\_rating:float, recommended:int);
* Clean the data and store it in new variable:

tmp1 = FOREACH tmp GENERATE airline\_name,flowm\_date, overall\_rating, staff\_rating, food\_rating, entertainment\_rating, value\_money\_rating, recommended, REPLACE (REPLACE(body,'\n',''),',','') AS airline\_name,flowm\_date, overall\_rating, staff\_rating, food\_rating, entertainment\_rating, value\_money\_rating, recommended;

tmp2 = FOREACH tmp1 GENERATE FLATTEN ((airline\_name,flowm\_date, overall\_rating, staff\_rating, food\_rating, entertainment\_rating, value\_money\_rating, recommended));

* Store cleaned data into file:

STORE tmp2 INTO 'hdfs://cluster-fcd6-m/ cleanedairlines\_5dec.csv' USING PigStorage (',');

* Quit pig using \q; and get the above generated file into hdfs using:

hadoop fs -get / cleanedairlines\_4dec.csv

**\*Repeat the same procedure for Airport and lounge to clean the Dat.**

Query On Hive:

1. Create a table ‘Airlines’ with the necessary columns

CREATE table airlines (airline\_name STRING, flown\_date DATE, cabin\_flown STRING, overall\_rating FLOAT,seat\_comfort\_rating FLOAT, cabin\_staff\_rating FLOAT, food\_beverages\_rating FLOAT, inflight\_entertainment\_rating FLOAT, value\_money\_rating FLOAT, recommended INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

TBLPROPERTIES("skip.header.line.count"="1");

1. Load the date from original CSV file into the table created

LOAD DATA LOCAL INPATH '/home/chethana\_hosurguruprasad2/cleanedairlines\_4dec.csv' INTO TABLE airlines;

1. Run query to get average of ratings

select airline\_name, avg(overall\_rating), AVG(seat\_comfort\_rating), AVG(cabin\_staff\_rating), AVG(food\_beverages\_rating), AVG(inflight\_entertainment\_rating), AVG(value\_money\_rating), sum(recommended) from airlines group by airline\_name;

1. Create a new table to store the average values

CREATE TABLE new\_airlines

row format delimited

fields terminated by '|'

STORED AS RCFile

AS select airline\_name, avg(overall\_rating), AVG(seat\_comfort\_rating), AVG(cabin\_staff\_rating), AVG(food\_beverages\_rating), AVG(inflight\_entertainment\_rating), AVG(value\_money\_rating), sum(recommended) from airlines group by airline\_name;

1. Save the file with average values into CSV file

hive -e "insert overwrite local directory '/home/chethana\_hosurguruprasad2/Airlines' row format delimited fields terminated by ',' select \* from assignment.new\_Airlines" cat /home/chethana\_hosurguruprasad2/lounge/\* > /home/chethana\_hosurguruprasad2/lounge/mean\_Airlines.csv

1. Create a table ‘Lounge’ with the necessary columns

CREATE table lounge (airline\_name STRING, flown\_date DATE, airport STRING, lounge\_type STRING, overall\_rating FLOAT, comfort\_rating FLOAT, cleanliness\_rating FLOAT, bar\_beverages\_rating FLOAT, catering\_rating FLOAT, washrooms\_rating FLOAT, wifi\_connectivity\_rating FLOAT, staff\_service\_rating FLOAT, recommended INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

TBLPROPERTIES("skip.header.line.count"="1");

1. Load the date from original CSV file into the table created

LOAD DATA LOCAL INPATH '/home/chethana\_hosurguruprasad2/lounge/cleanedlounge\_4dec.csv' INTO TABLE lounge ;

3.Run query to get average of ratings

select airline\_name, avg(overall\_rating), AVG(comfort\_rating), AVG(cleanliness\_rating), AVG(washrooms\_rating), AVG(wifi\_connectivity\_rating), AVG(staff\_service\_rating), sum(recommended) from lounge group by airline\_name;

4.Create a new table to store the average values

CREATE TABLE new\_lounges

row format delimited

fields terminated by '|'

STORED AS RCFile

AS select airline\_name, avg(overall\_rating), AVG(comfort\_rating), AVG(cleanliness\_rating), AVG(washrooms\_rating), AVG(wifi\_connectivity\_rating), AVG(staff\_service\_rating), sum(recommended) from lounge group by airline\_name;

5.Save the file with average values into CSV file

hive -e "insert overwrite local directory '/home/chethana\_hosurguruprasad2/lounge' row format delimited fields terminated by ',' select \* from assignment.new\_lounges" cat /home/chethana\_hosurguruprasad2/lounge/\* > /home/chethana\_hosurguruprasad2/lounge/mean\_lounges.csv

1. Create a table ‘Airport’ with the necessary columns

CREATE table Airport (airport\_name STRING,flown\_date DATE, overall\_rating FLOAT, queuing\_rating FLOAT, terminal\_cleanliness\_rating FLOAT, recommended INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

TBLPROPERTIES("skip.header.line.count"="1");

1. Load the date from original CSV file into the table created

LOAD DATA LOCAL INPATH '/home/chethana\_hosurguruprasad2/lounge/cleanedairport\_4dec.csv' INTO TABLE lounge ;

3.Run query to get average of ratings

select airport\_name, avg(overall\_rating), AVG(queuing\_rating), AVG(terminal\_cleanliness\_rating), sum(recommended) from airport group by airport\_name;

4.Create a new table to store the average values

CREATE TABLE new\_airport

row format delimited

fields terminated by '|'

STORED AS RCFile

AS select airport\_name, avg(overall\_rating), AVG(queuing\_rating), AVG(terminal\_cleanliness\_rating), sum(recommended) from airport group by airport\_name;

5.Save the file with average values into CSV file

hive -e "insert overwrite local directory '/home/chethana\_hosurguruprasad2/airport' row format delimited fields terminated by ',' select \* from assignment.new\_airport" cat /home/chethana\_hosurguruprasad2/lounge/\* > /home/chethana\_hosurguruprasad2/lounge/mean\_airport.csv