Task 1:

* Create external Hive table named airports\_external from the airports.dat file from <https://openflights.org/data.html> dataset
  + Use proper data type for every specific column
  + Handle NULL values correctly
  + Notice and explore if there are any changes in /user/hive/warehouse dir in HDFS (attach screenshot to the report)

--CREATE TABLE AIRPORTS\_EXT

CREATE EXTERNAL TABLE Airports\_external(

Airport\_ID bigint,

Name String,

City STRING,

Country STRING,

IATA STRING,

ICAO STRING,

Latitude DOUBLE,

Longitude DOUBLE,

Altitude BIGINT,

Timezone DOUBLE,

DST STRING,

Tz\_database\_timezone STRING,

Type STRING,

Source STRING)

COMMENT 'This is Airports table stored as textfile'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE LOCATION '/user/student/airports\_table/airports\_t.dat';

--LOAD DATA

LOAD DATA INPATH '/user/student/airports/airports.dat'

OVERWRITE INTO TABLE Airports\_external;

Graphical user interface, application, table, Excel

Description automatically generated

No changes in /user/hive/warehouse dir in HDFS (table location was in user/student/)

Graphical user interface, text, application

Description automatically generated

Task 2:

* Create managed table named airports\_internal with columns airport\_id,name,city,country,timezone,tz as select from airports\_external
  + airports\_internal table must not have rows which IATA code is NULL
  + The table must be stored as a text file with ‘|’ character as a column separator
  + Notice and explore if there are any changes in /user/hive/warehouse dir in HDFS (attach screenshot to the report)

--CREATE TABLE AIRPORTS\_INT

CREATE TABLE airports\_internal

ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'

AS

SELECT airport\_id, name, city, country, timezone, tz\_database\_timezone as tz

FROM airports\_external

where IATA is not NULL;

Graphical user interface, application, table

Description automatically generated

/user/hive/warehouse dir in HDFS – new file

Graphical user interface, application

Description automatically generated

Task 3:

* Create external table airports\_partitioned dynamically partitioned by country and bucketed by city from the airports\_internal table
  + External table’s data must be stored at /user/hive/warehouse/airports\_partitioned dir
  + Table must be bucketed by 4 buckets
  + Notice and explore if there are any changes in /user/hive/warehouse dir in HDFS (attach screenshot to the report)

Note: table is not bucketed actually

--CREATE TABLE AIRPORTS\_DP

SET hive.exec.dynamic.partition.mode=nonstrict;

SET hive.exec.dynamic.partition=true;

SET hive.enforce.bucketing=true;

CREATE EXTERNAL TABLE Airports\_partitioned

(Airport\_ID BIGINT,

Name STRING,

City STRING,

Timezone DOUBLE,

Tz STRING

)

PARTITIONED BY (Country STRING)

CLUSTERED BY (City) INTO 4 BUCKETS

STORED AS TEXTFILE LOCATION '/user/hive/warehouse/airports\_partitioned';

--LOAD DATA

INSERT OVERWRITE TABLE Airports\_partitioned PARTITION(country)

SELECT Airport\_ID,Name,City,Timezone,Tz,Country FROM airports\_internal limit 100;

Graphical user interface, text, application, email

Description automatically generated

/user/hive/warehouse dir in HDFS – new files with partitions:

Table

Description automatically generated

NOTE: Limit 100 was used due to the following error, hence only 4 countries appeared as partitions

Graphical user interface, text, application, email

Description automatically generated

Task 4:

* Update table airports\_partitioned by inserting new row
  + Row should be statically partitioned by country column using “Unknown Country” value
  + Other fields’ values may be random
  + Notice and explore if there are any changes in /user/hive/warehouse dir in HDFS (attach screenshot to the report)
  + Check if the inserted record available for select

set hive.mapred.mode = strict;

INSERT INTO TABLE Airports\_partitioned PARTITION(country="Unknown country")

VALUES (666,"Unknown name","Unknown city","6","Unknown tz");

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Task 5:

* Create table airports\_avro stored in AVRO format and airports\_parquet stored in Parquet format from the airports\_external table
  + Notice and explore if there are any changes in /user/hive/warehouse dir in HDFS (attach screenshot to the report)
  + Compare size of resulting Parquet/AVRO files and source CSV file

CREATE TABLE airports\_avro

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS AVRO

AS

SELECT Airport\_ID,Name,City,Country,IATA,ICAO,Latitude,Longitude,Altitude,Timezone,DST,Tz\_database\_timezone

FROM airports\_external;

CREATE TABLE airports\_parquet

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS PARQUET

AS

SELECT Airport\_ID,Name,City,Country,IATA,ICAO,Latitude,Longitude,Altitude,Timezone,DST,Tz\_database\_timezone

FROM airports\_external;

AVRO:

Graphical user interface, application, email

Description automatically generated

Parquet:

Graphical user interface, application

Description automatically generated

CSV:

Graphical user interface, text, application, email

Description automatically generated

CSV file is the largest one.

Task 6:

* Convert AVRO file stored in /user/hive/warehouse/airports\_avro to JSON file
  + Hint: use avro-tools

hadoop fs -get /user/hive/warehouse/airports\_avro/000000\_0 Documents/airports\_avro

avro-tools tojson Documents/airports\_avro > Documents/airports\_avro.json

Graphical user interface, application

Description automatically generated

Task 7:

* Create external table with name ipinyou\_external and schema ip string, region\_id int, city int from **iPinYou** dataset using RegexSerDe
  + See **iPinYou** dataset description at Table 3 of the page 4 here: <http://contest.ipinyou.com/ipinyou-dataset.pdf>

-- CREATE TABLE

CREATE EXTERNAL TABLE ipinyou\_external

(ip string, --6

region\_id int, --7

city int --8)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'

WITH SERDEPROPERTIES

( "input.regex" = "^[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t([\\S ]\*)\\t([\\S ]\*)\\t([\\S ]\*)\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t[\\S ]\*\\t\\d$")

STORED AS TEXTFILE LOCATION '/user/student/ipinyou/';

--LOAD DATA

LOAD DATA INPATH '/user/student/ipinyou\_dataset.txt'

OVERWRITE INTO TABLE ipinyou\_external;

Graphical user interface, text

Description automatically generated with medium confidence

Task 8:

* Compute statistics for table airports\_internal and then:
  + Display statistics for the whole table
  + Display statistics for city column

analyze table airports\_internal compute statistics;

analyze table airports\_internal compute statistics for columns city;

DESCRIBE EXTENDED airports\_internal;

col\_name data\_type comment

1 airport\_id bigint

2 name string

3 city string

4 country string

5 timezone double

6 tz string

7 NULL NULL

8 Detailed Table Information Table(tableName:airports\_internal, dbName:default, owner:cloudera, createTime:1615649763, lastAccessTime:0, retention:0, sd:StorageDescriptor(cols:[FieldSchema(name:airport\_id, type:bigint, comment:null), FieldSchema(name:name, type:string, comment:null), FieldSchema(name:city, type:string, comment:null), FieldSchema(name:country, type:string, comment:null), FieldSchema(name:timezone, type:double, comment:null), FieldSchema(name:tz, type:string, comment:null)], location:hdfs://quickstart.cloudera:8020/user/hive/warehouse/airports\_internal, inputFormat:org.apache.hadoop.mapred.TextInputFormat, outputFormat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compressed:false, numBuckets:-1, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe, parameters:{serialization.format=|, field.delim=|}), bucketCols:[], sortCols:[], parameters:{}, skewedInfo:SkewedInfo(skewedColNames:[], skewedColValues:[], skewedColValueLocationMaps:{}), storedAsSubDirectories:false), partitionKeys:[], parameters:{numFiles=1, transient\_lastDdlTime=1615659716, COLUMN\_STATS\_ACCURATE=true, totalSize=453214, numRows=6074, rawDataSize=447140}, viewOriginalText:null, viewExpandedText:null, tableType:MANAGED\_TABLE)

Graphical user interface, table

Description automatically generated

DESCRIBE FORMATTED airports\_internal city;

Graphical user interface, application, table

Description automatically generated

Task 9:

* Drop all external and managed tables. What happened to related data files in Hive warehouse dir?

All files and folders deleted except for Airports\_partitioned folder and files because this table was external