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 EXTERNAL TABLE airports\_external (Airport int, Name string, City string, Country string, IATA string, ICAO string, Latitude float, Longitude float, Altitude int, Timezone float, DST string, Tz string, Type string, Source string)**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE**

**LOCATION '/user/cloudera/test/airports';  
Table

Description automatically generated**

**Warehouse dir is empty   
Graphical user interface, text

Description automatically generated with medium confidence**

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\_internal**

**row format delimited fields terminated by '|'**

**stored as textfile**

**as SELECT airport as airport\_id,**

**Name,City,Country,Timezone,Tz**

**FROM airports\_external**

**WHERE iata IS NOT NULL**

Text

Description automatically generated with medium confidence

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

**SET hive.exec.max.dynamic.partitions=3000;**

**SET hive.exec.max.dynamic.partitions.pernode=3000;**

**SET hive.exec.dynamic.partition=true;**

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

**CREATE EXTERNAL TABLE airports\_partitioned**

**(airport\_id INT,name STRING,city STRING,timezone STRING,tz STRING)**

**PARTITIONED BY (country STRING)**

**CLUSTERED BY (city) INTO 4 BUCKETS**

**ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'**

**STORED AS TEXTFILE**

**LOCATION '/user/hive/warehouse/airports\_partitioned'**

**FROM airports\_internal AS t**

**INSERT OVERWRITE TABLE airports\_partitioned PARTITION(country)**

**SELECT t.airport\_id AS airport\_id,**

**t.name AS name,**

**t.city AS city,**

**t.timezone AS timezone,**

**t.tz AS tz,**

**t.country AS country**Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

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

**INSERT INTO airports\_partitioned PARTITION(country='Unknown Country')**

**VALUES (7777,**

**'To the moon',**

**'Moscow',**

**3.0,**

**'"Europe/Moscow"');**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, table, email

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 stored AS AVRO AS**

**SELECT \***

**FROM airports\_external**

**CREATE TABLE airports\_parquet stored AS PARQUET AS**

**SELECT \***

**FROM airports\_external**Text, table

Description automatically generated

**Avro= 1025909B**

**Parquet= 603739B  
Dat= 1127225B**

Task 6:

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

**avro-tools tojson --pretty hdfs://192.168.88.240/user/hive/warehouse/airports\_avro/000000\_0 > /home/cloudera/Downloads/air.json**

Graphical user interface, text, application, email

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 EXTERNAL TABLE ipinyou\_external (ip STRING,**

**region\_id INT, city INT)**

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

**WITH SERDEPROPERTIES ("input.regex"="(?:)\\s(\\S+\\.\\S+\\.\\S+\\.\\S)\\s(\\d{1,3})\\s(\\d{1,3})\\t")**

**STORED AS TEXTFILE**

**LOCATION '/user/cloudera/try';**

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**

Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

Task 9:

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

**Data files were deleted, although there were some traces. Probably due to my manual mapipulations.**

Text

Description automatically generated

**Expected output:**

Provide Hive scripts, execution plan dump/screenshtot and analysis.