Homework 4

Mini-project

For the mini-project we use a public data set, the Airline On-Time Statistics and Delay Causes data set, published by the United States Department of Transportation at http://www.transtats.bts.gov/. The On-Time Performance dataset records flights by date, airline, originating airport, destination airport, and many other flight details. Data is available for flights since 1987. The FAA uses the data to calculate statistics such as the percent of flights that depart or arrive on time by origin, destination, and airline.

Goals of the project:
1. Get experience to work with real data
2. Perform main data related tasks: explore the data, create schema, load the data, analyze the data using SQL.
3. Look up for required technical information in the external sources – manuals and internet at large.

1. Data for the project

The data is organized in a so-called star schema. The star schema consists of one (or more) fact tables referencing any number of dimension tables.

Fact Table

Fact tables record measurements or metrics for a specific event. In our case it is one table containing data about flights by date, airline, originating airport, destination airport, and many other flight details. Fact tables generally consist of numeric values, and foreign keys to dimensional data where descriptive information is kept. Glossary of Terms used in the On-Time Performance dataset can be found here:
https://www.transtats.bts.gov/Glossary.asp

Dimension tables

Dimension tables usually have a relatively small number of records compared to fact tables, each record may have a attributes to describe the fact data.

You will have the following dimension data describing some attributes of the fact table:

Dataset	Attribute
L_DISTANCE_GROUP_250.csv	DistanceGroup
L_AIRLINE_ID.csv (ID, Name)	DOT_ID_Reporting_Airline
L_AIRPORT.csv (Code, Name)	Dest
L_AIRPORT.csv (Code, Name)	Origin
L_AIRPORT_ID.csv (ID, Name)	DestAirportID
L_AIRPORT_ID.csv (ID, Name)	OriginAirportID
L_CANCELATION.csv (Code, Reason)	CancellationCode
L_ONTIME_DELAY_GROUPS.csv (Code, Description)	ArrivalDelayGroups
L_ONTIME_DELAY_GROUPS.csv (Code, Description)	DepartureDelayGroups
L_WEEKDAYS.csv (Code, Day)	DayOfWeek

2. Download CSV file with fact table data

The data for the fact table can be downloaded from the web site
https://www.transtats.bts.gov/

- Go to “Passenger Travel” in the “By Subject” list
- Then click on “Airline On-Time Performance Data”
- Finally click on "Download" link below "Reporting Carrier On-Time Performance (1987-present)"
- Check "Prezipped File" check box
- Filter on your Year and Month
- Download a zip file, that contain a csv file with the name

On_Time_Reporting_Carrier_On_Time_Performance_(1987_present)_<YYYY>_<MM>

Rename the file to "al_perf.csv" for easier handling.

3. Create a schema

Create a new schema called 'FAA' for your project in the Workbench

4. Load the data

You will use different methods to load the data into your database.

1) Create and Load fact table using

MySQL provides a utility mysqlimport to load large data sets into tables. Follow these steps to load the fact table data:

- Create table 'al_perf' in schema FAA using CreateFactTable.sql script

- Create EC2 Instance on AWS. Document "Create_EC2_Instance_on_AWS_instructions.docx" contains instructions.

- Secure copy(scp) your csv file **from your laptop** to your home directory of the EC2 instance. For example:

  $scp ~/al_perf.csv ec2-user@<your_EC_instance_public_IP>:/home/ec2-user

- Run the following command to install mysql on your EC2 instance:

  $sudo yum install mysql

- Run mysqlimport utility to move the file AWS RDS. Notice options that are used below. You can read about their meaning in MySQL documentation. Example:

```
$ mysqlimport
--local \
--compress \
--user=admin
--password=<your_password> \
--host=<your_mysql_database_aws_server>.rds.amazonaws.com  \
--fields-terminated-by=',' \
--fields-optionally-enclosed-by='"' \
<name_of_your_workbench_schema> al_perf.csv
```

2) Create and Load dimension tables

L_AIRLINE_ID.csv
L_AIRPORT.csv
L_AIRPORT_ID.csv
L_DISTANCE_GROUP_250.csv
L_WEEKDAYS.csv

using Table Data Import Wizard on the Workbench. The Wizard does not require to create tables in advance, it creates a table if it does not exist. However, if it takes too long you can load the tables using mysqlinport. In that case you will need to create the dimension tables first.

3) Create dimension table L_CANCELATION using CREATE TABLE statement. Load data into dimension tables using INSERT statements.

5. Analyze the data

Create and run SQL queries to do the following.

1) Find maximal departure delay in minutes for each airline. Sort results from smallest to largest maximum delay. Output airline names and values of the delay.

2) Find maximal early departures in minutes for each airline. Sort results from largest to smallest. Output airline names.

3)Rank days of the week by the number of flights performed by all airlines on that day (1 is the busiest). Output the day of the week names, number of flights and ranks in the rank increasing order.

4) Find the airport that has the highest average departure delay among all airports. Consider 0 minutes delay for flights that departed early. Output one line of results: the airport name, code, and average delay.

5) For each airline find an airport where it has the highest average departure delay. Output an airline name, a name of the airport that has the highest average delay, and the value of that average delay.

6) a) Check if your dataset has any canceled flights.
b) If it does, what was the most frequent reason for each departure airport? Output airport name, the most frequent reason, and the number of cancelations for that reason.

7) Build a report that for each day output average number of flights over the preceding 3 days.