Dataset and Problem Statement

This is a compulsory assignment on SQL based data processing. You will get a document with instructions to code. You need to write the Hive commands and your comments about those commands below each instruction and submit the document (only the commands and your explanation about the commands used is needed).

The deadline for submission of this assignment is **January 22, 2017, 11:59 PM**. For submissions obtained within 1 week of the deadline, there will be a 30% penalty. Submissions beyond 1 week of the deadline will not be accepted.

Airline Dataset

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. The department has made the collection of all domestic flight records for the years 1987 to 2008.

There is a different file for each year, which includes the details about all the flights scheduled by all air carriers, on all routes. Please refer the following table containing the metadata/data dictionary of this dataset.

Metadata/Data Dictionary

Sr. No.	Field / Column Name	Description
1.	Year	1987 to 2008
2.	Month	1 to 12
3.	DayofMonth	1 to 31

4.	DayofWeek	1 (Monday) to 7 (Sunday)
5.	DepTime	Actual departure time (local, hhmm)
6.	CRSDepTime	Scheduled departure time (local, hhmm)
7.	ArrTime	Actual arrival time (local, hhmm)
8.	CRSArrTime	Scheduled arrival time (local, hhmm)
9.	UniqueCarrier	This is the carrier number allocated to each air carrier. It is unique for air carriers.
10.	FlightNum	This is the flight number for each flight operating under an air carrier.
11.	TailNum	Aeroplane's tail number
12.	ActualElapsedTime	Elapsed time of flight, it is measured in Minutes.
13.	CRSElapsedTime	Scheduled elapsed time of flight, it is measured in Minutes.
14.	AirTime	Flight time, it is measured in Minutes.
15.	ArrDelay	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
16.	DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
17.	Origin	Origin Airport, IATA airport code
18.	Dest	Destination Airport, IATA airport code
19.	Distance	Distance between origin and destination. It is measured in miles.
20.	TaxiIn	Taxi in time, in minutes
21.	TaxiOut	Taxi out time, in minutes
22.	Cancelled	Cancelled Flight Indicator (1=Yes)
23.	CancellationCode	The reason for cancellation. (A = carrier, B = weather, C = NAS, D = security)
24.	Diverted	Diverted Flight Indicator (1=Yes)

25.	CarrierDelay	Delay due to carrier services, measured in minutes
26.	WeatherDelay	Delay occurred due to weather conditions, measured in minutes
27.	NASDelay	National Air System Delay, measured in minutes
28.	SecurityDelay	Delay occurred due to security reasons, measured in minutes
29.	LateAircraftDelay	Delay due to delayed inbound aircraft, measured in minutes

Using this dataset, we wish to perform some analysis. The general checkpoints are given below:

Checkpoint 1: Data Understanding

This is the most important stage of solving any problem. In this checkpoint, you are supposed to get well versed with the dataset. There is a question in your assignment, which is based on this understanding.

Checkpoint 2: Creating the data tables for analysis

In this checkpoint, you are expected to use DDL statements and create the necessary data tables. You may be needed to create managed as well as external tables. This step needs very keen observation about the datatypes and models to be used while creating tables. With the thorough understanding of data in hand, you can easily decide and formulate the DDL statements for the same.

Checkpoint 3: Performing basic analysis

In this checkpoint, you are expected to use DML statements to get basic insights from the airline data. In this section, the commands will be simple and straightforward. These commands will form the base for complex analysis, that is the focus of next checkpoint.

Checkpoint 4: Performing complex analysis

In this checkpoint, you are expected to use DML statements and your prior knowledge about SQL commands to get more complex insights. The tasks under this checkpoint will include the results from multiple commands to get final insights. This checkpoint will demand the use of a chain of commands to arrive at the expected solution.

NOTE: Keep on writing about each and every command that you are using to arrive at final insight. Your comments/explanation about each command will help you doing so.

Assignment Submission

The deadline for submission of this assignment is **January 22, 2017, 11:59 PM**. For submissions obtained within 1 week of the deadline, there will be a 30% penalty. Submissions beyond 1 week of the deadline will not be accepted.

Instructions

Download the HiveAssignment.docx file from the bottom of this page and follow the instructions to complete your submission.

Marks Distribution

This assignment is **worth 700 points.**

There are 15 tasks in all in this assignment for which you need to write the applicable Hive command.

Submission

Submit the completed **.docx** file through the submission link below. Name it **DDA_Hive.docx**.

IMPORTANT NOTE: Please name your submission file correctly using the above naming convention to avoid confusion during the grading process.

SUBMIT CODE HERE (filename = DDA_Hive.docx)