Commercial airplane flight delays in the United States 2011-2020 after EDA

The contains flight statistics for all airports in the United States from January 2011 to December 2020. Each observation is reported by month, year, airport, and airline. Flights can be categorized as on time, delayed, canceled, or diverted. Flight delays are attributed to five causes: carrier, weather, NAS, security, and late aircraft.

What flights does the reporting cover?
The rule requires carriers to report on domestic operations to and from U.S. airports.

A flight is considered delayed when it arrived 15 or more minutes than the schedule. Delayed minutes are calculated for delayed flights only.

When multiple causes are assigned to one delayed flight, each cause is prorated based on delayed minutes it is responsible for. The displayed numbers are rounded and may not add up to the total.

The marketing carrier networks are:
Alaska Airlines (AS)*
Allegiant Air (G4)
American Airlines (AA)*
Delta Air Lines (DL)*
Frontier Airlines (F9)
Hawaiian Airlines (HA)*
JetBlue Airways (B6)
Southwest Airlines (WN)
Spirit Airlines (NK)
United Airlines (UA)*

*Includes branded code-share partners

The reporting airlines are:
Alaska Airlines (AS)
Allegiant Air (G4)
American Airlines (AA)
Delta Air Lines (DL)
Endeavor Air (9E)
Envoy Air (MQ)
Frontier Airlines (F9)
Hawaiian Airlines (HA)
Horizon Air (QX)
JetBlue Airways (B6)
Mesa Airlines (YV)
PSA Airlines (OH)
Republic Airlines (YX)
SkyWest Airlines (OO)

Southwest Airlines (WN)
Spirit Airlines (NK)
United Airlines (UA)

The airlines report the causes of delays in five broad categories:

- Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.
- Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

Data Set

Variable	Definition	Key
year	year of the observation	The value is in the range between [2011, 2020]
month	month of the observation	The value is in the range between [1, 12]
carrier	IATA two-letter airline designator	
Carrier_*name*	*name of the airline*	
airport	*three-letter airport code*	
*Airport_*name	city and airport full name	
Arr*_flights*	*total number of arriving flights in the observation*	The value is in the range between [1, 22000]
*Arr*_del15	total number of delayed flights in the observation	The value is in the range between [0, 5,270]
Carrier_*ct*	*number of flight delays attributed to carriers*	The value is in the range between [0, 1,240]
*Weather_*ct	number of flight delays attributed to weather	The value is in the range between [0, 295]
Nas_*ct*	*number of flight delays attributed to the national advisory service*	The value is in the range between [0, 2400]
*Security_*ct	number of flight delays attributed to security concerns	The value is in the range between [0, 26.1]

Late_*aircraft*_ct	number of flight delays because the aircraft was late on its previous trip	The value is in the range between [0, 1850]
Arr_*cancelled*	*total number of canceled flights in the observation*	The value is in the range between [0, 4,950]
*Arr*_diverted	total number of diverted flights in the observation	The value is in the range between [0, 256]
Min_*delay*	*total number of minutes delayed in that month by an airline in an airport*	The value is in the range between [0, 429000]
*Carrier*_delay	total number of minutes delayed attributed to the airline	The value is in the range between [0, 197000]
Weather_*delay*	*total number of minutes delayed attributed to the weather*	The value is in the range between [0, 32000]
*Nas*_delay	total number of minutes delayed attributed to the national advisory service	The value is in the range between [0, 137000]
Security_*delay*	*total number of minutes delayed attributed to security concerns*	The value is in the range between [0, 3190]
*Late*_aircraft_*delay*	*total number of minutes delayed attributed to the flight being late on its previous trip*	The value is in the range between [0, 148000]
*Arr*_ontime	total number of flights on time	The value is in the range between [0, 20500]
city	airport city	
state	airport state	
perontime	percentage of flights that are on time	The value is in the range between [0, 1]
perdelay	percentage of flights that are delayed	The value is in the range between [0, 1]
percancelled	percentage of flights that are canceled	The value is in the range between [0, 0.99]
perdiverted	percentage of flights that are diverted	The value is in the range between [0, 0.67]
Per_*car*_delay	percentage of delayed minutes attributed to the carrier	The value is in the range between [0, 1]
Per_*wea*_delay	percentage of delayed minutes attributed to weather	The value is in the range between [0, 1]

Per_*nas*_delay	percentage of delayed minutes attributed to the national advisory service	The value is in the range between [0, 1]
Per_*sec*_delay	percentage of delayed minutes attributed to security issues	The value is in the range between [0, 1]
Per_*late*_delay	percentage of delayed minutes attributed to the flight being late on its previous trip	The value is in the range between [0, 1]

Tasks

Question 1

Design a SQL database using the above airplane flight delays data set.

1. Use Visual Paradigm to draw conceptual diagram
2. Use Visual Paradigm to draw ERD

Question 2

Consider the following two relations for a firm:

Employee(employeeID, employeeName, contact, email)

Performance(employeeID, departmentID, rank)

The following is a typical query against these relations:

```
select Employee_T.employeeID, employeeName, departmentID, grade

from Employee_T, Performance_T

where Employee_T.employeeID = Performance_T.employeeID AND rank == 1.0

order by employeeName;
```

1. By what attributes should indexes be defined to speed up this query? Give the reasons for each attribute selected.
2. Write SQL commands to create indexes for each attribute you identified in the previous question.

Question 3

Create a table named students with the following columns:

studentID (integer, PK, auto-increment), firstName (varchar(30)), lastName (varchar(30)), dob (date), enrol_date (date)

To insert the following records into the students table:

```
('John', 'Doe', '2000-05-15', '2018-09-01'),
('Jane', 'Smith', '1999-07-20', '2017-09-01'),
('Michael', 'Johnson', '2001-03-10', '2019-09-01'),
('Emily', 'Davis', '2002-11-25', '2020-09-01'),
('William', 'Brown', '2000-08-30', '2018-09-01'),
('Olivia', 'Wilson', '2001-12-12', '2019-09-01'),
```

To add a new column email to the students table and ensure that it can store unique values

To perform a transaction that inserts a new student and then commits the transaction.

To perform a transaction that inserts a new student but then rolls back the transaction before committing.

To create a savepoint, perform an insert, and then roll back to the savepoint, then commit the transaction.