

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 contains records of 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 database related tasks: explore the data, create schema, load the data, analyze the data using SQL.
3. Work in AWS cloud using MySQL RDS.
4. Look up technical information in the external sources, as needed.

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>

You need to download the fact table for this project from the FAA website according to the provided instructions and for **the specific year and month assigned to you**.

Dimension tables

Dimension tables usually have a relatively small number of records compared to fact tables, and each record may have 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_CANCELLATION.csv	(Code, Reason) CancellationCode
L_WEEKDAYS.csv	(Code, Day) DayOfWeek

You need to download the Dimension tables for this project from Canvas, folder Mini-project.

2. Download CSV file with fact table data

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

Details follow:

- In the “By Subject” list, go to “Passenger Travel”, as highlighted below:

Bureau of Transportation Statistics

Ask-A-Librarian A-Z Index

Search

Topics and Geography

Statistical Products and Data

National Transportation Library

Newsroom

About BTS

TranStats

Search this site:

Go

Advanced Search

Resources

Database Directory

Glossary

Upcoming Releases

Data Release History

Data Finder

By Mode

Aviation

Maritime

Highway

Transit

Rail

Pipeline

Bike/Pedestrian

Other

By Subject

Safety

Freight Transport

Passenger Travel

Infrastructure

Economic/Financial

Quick Answers

Carrier Snapshots

Airline Fuel Cost and Consumption

Air Freight Summary

Employment

Airline Activity : National Summary (U.S. Flights)

Enplaned Passengers (million)

Departures (000)

Freight/Mail (million lbs)

Load Factor (%)

Airlines with scheduled service

* 12 months ending June of each year

Airline Domestic Market Share July 2024 - June 2025

Domestic Revenue Passenger Miles (billions)

Airlines Share

Delta

American

Spirit

Southwest

United

Alaska

JetBlue

Spirit

Frontier

SkyWest

Hawaiian

Other

8.700000000000001%

At a Glance

Flight Delays

more...

Percent of U.S. Flights Marking Carrier Network (2024-2025)

Jul

Sep

Oct

Nov

Dec

Jan

Feb

Mar

Apr

May

Jun

Click a bar for details. Mouseover it for percentage.

Average Air Fares

more...

Average Domestic Airline Fares

2015

2017

2019

2021

2023

2016

2018

2020

2022

20...

Click a bar for details. Mouseover it for total.

- Then click on “Airline On-Time Performance Data”.

An official website of the United States government

United States Department of Transportation

Bureau of Transportation Statistics

Ask-A-Librarian A-Z Index

Search

Topics and Geography

Statistical Products and Data

National Transportation Library

Newsroom

About BTS

BTS> TranStats

TranStats

Search this site:

Go

Advanced Search

Resources

Database Directory

Glossary

Upcoming Releases

Data Release History

Data Finder

By Mode

Aviation

Maritime

Highway

Transit

Rail

Pipeline

Bike/Pedestrian

Other

By Subject

Data Library: Passenger Travel

Databases

Summary Tables

Glossary

Filter Mode All Modes

Go

<<Prev Rows 1 to 15 of 25 Next>>

Database Name	Description	Profile
Air Carrier Statistics (Form 41 Traffic)- U.S. Carriers	Monthly data reported by certificated U.S. air carriers on passengers, freight and mail transported. Also includes aircraft type, service class, available capacity and seats, and aircraft hours ramp-to-ramp and airborne.	Profile
Air Carrier Statistics (Form 41 Traffic)- All Carriers	Monthly data reported by certificated U.S. and foreign air carriers on passengers, freight and mail transported. Also includes aircraft type, service class, available capacity and seats, and aircraft hours ramp-to-ramp and airborne.	Profile
Air Carrier Summary Data (Form 41 and 298C Summary Data)	Summary data of the non-stop segment and on-flight market data reported by air carriers on Form 41 and Form 298C.	Profile
Airline On-Time Performance Data	Monthly data reported by US certified air carriers that account for at least one percent of domestic scheduled passenger revenues--includes scheduled and actual arrival and departure times for flights.	Profile
Airline Origin and Destination Survey (DB1B)	Origin and Destination Survey (DB1B) is a 10% sample of airline tickets from reporting carriers. Data includes origin, destination and other itinerary details of passengers transported.	Profile
American Travel Survey (ATS) 1995	National data on the nature and characteristics of long-distance personal travel, from a household survey conducted by BTS about every five years.	Profile
Aviation Support Tables	Provides comprehensive information about U.S. and foreign air carriers, carrier entities, worldwide airport locations, and other geographic data. These data also include information on various aircraft types, their manufacturer and model names.	Profile
Census Transportation Planning Package (CTPP) 1990	The 1990 Census Transportation Planning Package (CTPP) is a collection of summary tables that have been generated from both the 1990 census short and long forms. The tables contain information about population and household characteristics, worker characteristics and characteristics of Journey-to-Work (JTW). The CTPP is organized into a series of parts contained with two elements namely Urban and State. The parts define whether the tables are summarizing information by place of residence, place of work or journey to work.	Profile
Census Transportation Planning Package (CTPP) 2000	The Census Transportation Planning Package (CTPP) is a collection of summary tables that have been generated from the census long form data collected in year 2000. These summary tables contain three sets of tabulations: Part I - PLACE OF RESIDENCE, Part II - PLACE OF WORK, and Part III - JOURNEY-TO-WORK.	Profile

- Finally, in the section for "Reporting Carrier On-Time Performance (1987-present)", click on Download, as highlighted below:

UNITED STATES DEPARTMENT OF TRANSPORTATION

Bureau of Transportation Statistics

Ask-A-Librarian A-Z Index

Search

Topics and Geography Statistical Products and Data National Transportation Library Newsroom About BTS

BTS > TranStats

TranStats

Search this site:

[Advanced Search](#)

Resources

- Database Directory
- Glossary
- Upcoming Releases
- Data Release History

Data Finder

By Mode

- Aviation
- Maritime
- Highway
- Transit
- Rail
- Pipeline
- Bike/Pedestrian
- Other

By Subject

- Safety
- Freight Transport

Database Name: Airline On-Time Performance Data

[Databases](#) [Database Profile](#)

All Rows Shown

Table Name	Description
Marketing Carrier On-Time Performance (Beginning January 2018)	Marketing carriers market flights for themselves and, in some cases, for regional code share partners. Table shows on-time data for the marketing carrier network, if applicable, the marketing carrier that is the reporting carrier and the marketing carrier's code-share affiliates as a group. Reporting carriers that do not market flights are included in the regional code share group. Table shows: on-time arrival and departure data for non-stop domestic flights by month and year, by marketing network, marketing carrier that reports and regional code-share group, by origin and destination airport. Includes scheduled and actual departure and arrival times, canceled and diverted flights, taxi-out and taxi-in times, causes of delay and cancellation, air time, and non-stop distance. Use Download for individual flight data.
Reporting Carrier On-Time Performance (1987-present)	Reporting carriers are required to (or voluntarily) report on-time data for flights they operate: on-time arrival and departure data for non-stop domestic flights by month and year, by carrier and by origin and destination airport. Includes scheduled and actual departure and arrival times, canceled and diverted flights, taxi-out and taxi-in times, causes of delay and cancellation, air time, and non-stop distance. Use Download for individual flight data.

[Table Profile](#) [Carrier Release Status](#) [Download](#)

All Rows Shown

On the screen below, check "Prezipped File" check box and Filter on your Year and Month

The screenshot shows the Bureau of Transportation Statistics website. The main heading is "Bureau of Transportation Statistics". Below it, there are navigation links: "Topics and Geography", "Statistical Products and Data", "National Transportation Library", "Newsroom", and "About BTS". A search bar is located on the right. The "On-Time : Reporting Carrier On-Time Performance (1987-present)" page is displayed. The "Latest Available Data: June 2025" is shown. The "Filter Year" is set to 2025 and the "Filter Period" is set to January. The "Prezipped File" checkbox is checked. The "Download" button is visible. The "Field Name" and "Description" table is shown below.

Field Name	Description	Support Table
Time Period		
<input type="checkbox"/> Year	Year	
<input type="checkbox"/> Quarter	Quarter (1-4)	Get Lookup Table
<input type="checkbox"/> Month	Month	Get Lookup Table
<input type="checkbox"/> DayofMonth	Day of Month	
<input type="checkbox"/> DayOfWeek	Day of Week	Get Lookup Table
<input type="checkbox"/> FlightDate	Flight Date (yyyymmdd)	

Next, click on the blue "Download" button to download a zip file, that contain a csv file with the name On_Time_Reporting_Carrier_On_Time_Performance_(1987_present)_<YYYY>_<MM> Unzip the file and rename the file to "al_perf.csv" for easier handling.

3. Create a schema

Open MySQL Workbench and Create a new schema called 'FAA' for your project. Don't forget to set the charset as UTF8 and make it the default schema.

The screenshot shows the MySQL Workbench Schema Editor window. The "Schema Name" is set to "FAA". The "Character Set" is set to "utf8" and the "Collation" is set to "utf8_bin". The "Apply" button is visible at the bottom right. The "Schema" tab is selected. The "Context Help" panel on the right shows a message: "Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help."

Once you have created the FAA schema, open a new tab and run the following SQL statement to grant access from AWS EC2m which you'll create later on (you can copy and paste it):

GRANT SESSION_VARIABLES_ADMIN ON *.* TO 'admin'@'%';

The screenshot shows the MySQL Workbench interface. On the left, the 'SCHEMAS' sidebar is open, showing a tree view with 'FAA' selected under 'RDS_MySQL_v8_sv642'. The main editor area displays the SQL statement: `GRANT SESSION_VARIABLES_ADMIN ON *.* TO 'admin'@'%';`. Below the editor, the 'Action Output' tab is active, showing a table with execution results.

	Time	Response	Duration / Fetch Time
1	13:40:08	A Changes applied	
2	13:41:02	G 0 row(s) affected	0.067 sec

4. Load the data

Given the size of this dataset, you will need to use different methods to load the data into your database. Follow the steps listed below. **Follow them very carefully.**

4.1) Create and Load fact table

MySQL provides a utility `mysqlimport` to load large data sets into tables that we are going to learn how to use. You will run `mysqlimport` from an AWS EC2. Follow these steps to load the fact table data:

a) Using MySQL workbench, create table 'al_perf' in schema FAA using CreateFactTable.sql script provided in Canvas, folder Mini-project. You may get a few warnings that some features will be deprecated in future releases. Just ignore them.

b) Create EC2 Instance on AWS.

For that, use the **The Document "Create_EC2_Instance_on_AWS_instructions.docx"** available in Canvas, **folder Mini_project, contains the instructions**. Return to this page and section after you have successfully created your EC2 instance on AWS.

c) Secure copy(scp) your csv file from your laptop to your home directory of the EC2 instance:

Note: Make sure you have logged out of your EC2. This command is to be run **from your local computer (Power Shell or Terminal)**

Using the code below from your computer (Power shell on Windows or Terminal on Macs), type

```
scp -i <path_to/your_keypair.pem> <path_to/al_perf.csv> ec2-user@<your_EC_instance_public_IPv4>:/home/ec2-user
```

The structure of this command is explained below with color-code::

scp -i : this is the command to secure copy a file from your computer to the EC2.

<path_to/your_keypair.pem> : This is the path to your key pair file and your keypair file name

<path_to/al_perf.csv> : This is the path and the file name that you want to copy to the EC2. The file name is al_perf.csv but the path will be different for you because it depends on where you saved this file.

Note: Moving the al_perf.csv file to you present working directory may make your life easier here with less characters to type, hence, less chances for typos.

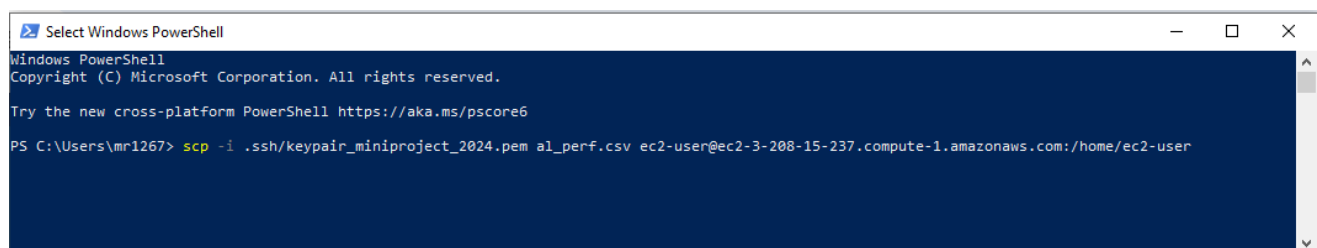
ec2-user@<your_EC_instance_public_IPv4>:/home/ec2-user : is broken down in three parts:

ec2-user@ : This never changes. It is the default user for any EC2 that you created (unless you create another one).

<your_EC_instance_public_IPv4> : This is the Public IPv4 address of your EC2 instance on AWS that you copied before and used previously to connect.

:/home/ec2-user: This is the location on the EC2 where the file will be saved.

This command will copy the csv file from your computer to the EC2 instance on AWS and should look like this one below: (in this case, I saved the al_perf.csv file inside the ssh folder for simplicity so I didn't need to specify the path to it).



```
Select Windows PowerShell
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/powershell

PS C:\Users\mr1267> scp -i .ssh/keypair_miniproject_2024.pem al_perf.csv ec2-user@ec2-3-208-15-237.compute-1.amazonaws.com:/home/ec2-user
```

You should see the following screen, with the status, during the upload:


```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\mr1267> ssh -i .ssh/keypair_miniproject_2024.pem ec2-user@ec2-3-208-15-237.compute-1.amazonaws.com
Last login: Thu Nov  7 00:37:50 2024 from 70.18.239.63

      _#_
     _###_      Amazon Linux 2
    _\#####\
   ~~~~~\###|
   ~~~~~\##/
   ~~~~~V~'~>
   ~~~~~/
   ~~~~~/_/_/
   ~~~~~/_/_'

AL2 End of Life is 2025-06-30.

A newer version of Amazon Linux is available!

Amazon Linux 2023, GA and supported until 2028-03-15.
https://aws.amazon.com/linux/amazon-linux-2023/

[ec2-user@ip-172-31-18-199 ~]$ exit
logout
Connection to ec2-3-208-15-237.compute-1.amazonaws.com closed.
PS C:\Users\mr1267> scp -i .ssh/keypair_miniproject_2024.pem al_perf.csv ec2-user@ec2-3-208-15-237.compute-1.amazonaws.com:/home/ec2-user
al_perf.csv                                     100% 249MB 25.0MB/s 00:09
PS C:\Users\mr1267>
```

d) After the upload is completed, connect to your EC2 instance with:

ssh -i ssh/your_keypair_name.pem ec2-user@your_public_IPv4

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\mr1267> ssh -i .ssh/keypair_miniproject_2024.pem ec2-user@ec2-3-208-15-237.compute-1.amazonaws.com
```

Your screen should look like this below after you have connected with your EC2 again


```
ec2-user@ip-172-31-18-199:~  
Transaction Summary  
=====
```

Install 1 Package	
Total download size: 8.8 M	
Installed size: 49 M	
Is this ok [y/d/N]: y	
Downloading packages:	
mariadb-5.5.68-1.amzn2.0.1.x86_64.rpm	8.8 MB 00:00:00
Running transaction check	
Running transaction test	
Transaction test succeeded	
Running transaction	
Installing : 1:mariadb-5.5.68-1.amzn2.0.1.x86_64	1/1
Verifying : 1:mariadb-5.5.68-1.amzn2.0.1.x86_64	1/1

```
Installed:  
mariadb.x86_64 1:5.5.68-1.amzn2.0.1  
Complete!  
[ec2-user@ip-172-31-18-199 ~]$
```

f) After you have connected to your EC2 instance and installed MySQL, run `mysqlimport` utility to move the file in your EC2 instance to AWS RDS (your MySQL database) with the code provided below. Notice the options that are used below.

\$ 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

The easiest way to do this is to copy this code to a text editor, make the necessary modifications there and when ready, copy and paste it on the terminal. You can use this sample below:

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

Your screen should look like this one below:

```
ec2-user@ip-172-31-18-199:~  
[ec2-user@ip-172-31-18-199 ~]$ mysqlimport --local --compress --user=admin --password=Marcio2024* --host=database-1.cczexdqywykj.us-east-1.rds.amazonaws.com  
--fields-terminated-by=',' --fields-optionally-enclosed-by='"' FAA_al_perf.csv
```

Once you click “enter” you may not see anything happening for a few seconds. When the data transfer gets complete, you will see the following message:

```
Select ec2-user@ip-172-31-18-199:~  
io2024* --host=database-1.cczexdqywykj.us-east-1.rds.amazonaws.com --fields-terminated-by=  
' ' --fields-optionally-enclosed-by='"' FAA_al_perf.csv  
FAA.al_perf: Records: 577263 Deleted: 0 Skipped: 0 Warnings: 157762  
[ec2-user@ip-172-31-18-199 ~]$
```

The number of imported records should be the same as the number of rows in your al_perf.csv file

After completing this step, open MySQL Workbench and look at the table al_perf. The number of records displayed in the screen above should be now in the table al_perf in MySQL Workbench.

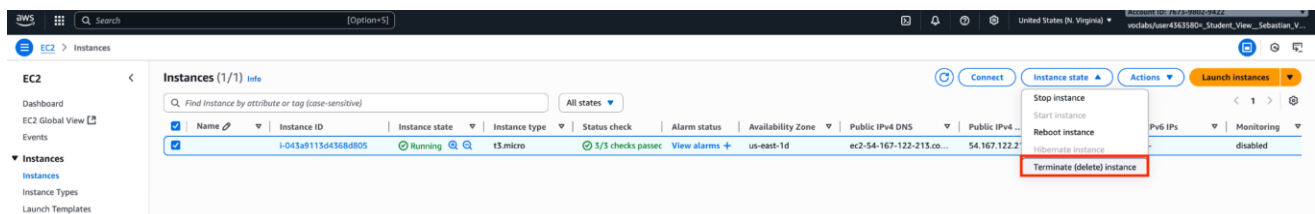
Once you have confirmed that, go back to the AWS console and terminate your EC2. You won’t need the EC2 anymore. This is to prevent unnecessary costs from accruing to your budget.

MAKE SURE YOU TERMINATE THE EC2 AND NOT THE DATABASE (RDS).

If you erroneously terminate your (database RDS) you will have to install everything again.

BE CAREFUL HERE!!!!

Follow the highlighted commands below to terminate the EC2.



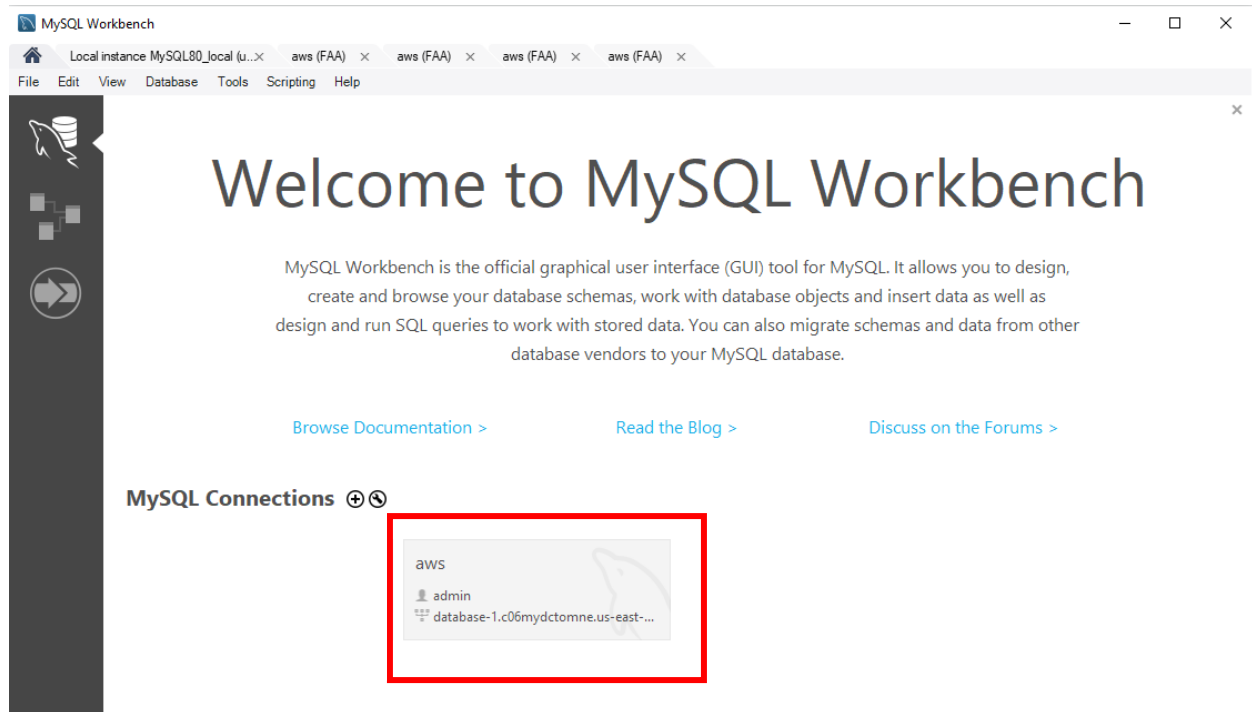
Important:

The host here **IS NOT** the EC2 IPv4 address. It is the database endpoint address. You can get it from the hostname in MySQL workbench, as shown in the next page.

g) Troubleshooting user and password

(you can skip to 4.2 if you successfully logged in):

user = admin will only work if you have set admin as the root user of your MySQL database. If you have used a different root user when setting up your database, you will need to use the same here. To check the user you set up, open MySQL workbench, on the initial screen (below), high click on grey rectangle and select edit connection.



You'll see the screen below. The user you set up before will be in the field "username" (in this case, "admin"). You need to sue the same value in the "user" parameter when running mysqlimport.

Manage Server Connections

MySQL Connections
Local instance MySQL80_...
aws

Connection Name:

Connection Method: Method to use to connect to the RDBMS

Parameters | SSL | Advanced

Hostname: Port: Name or IP address of the server host - and TCP/IP port.

Username: Name of the user to connect with.

Password: The user's password. Will be requested later if it's not set.

Default Schema: The schema to use as default schema. Leave blank to select it later.

Password – This is the password that you used before when setting up your MySQL server.

If for any reason you cannot remember the user or password, you'll need to reset them on AWS. To do so, go to your database on AWS (you will need to start the lab on AWS) and on the screen below, click on "Modify", as shown below.

RDS > Databases

Databases Group resources Modify Actions Restore from S3 Create database

Filter by databases

DB identifier	Role	Engine	Region & AZ	Size	Status	CPU	Current activity	Maintenance	VPC	Multi-AZ
database-1	Instance	MySQL Community	us-east-1d	db.t3.micro	Available	1.03%	0 Connections	available	vpc-028f0335814080b10	No

You will see the screen below. Enter your new master password in the highlighted area.

Modify DB instance: database-1

Settings

DB engine version

Version number of the database engine to be used for this database

8.0.28 ▼

DB instance identifier [Info](#)

Type a name for your DB instance. The name must be unique across all DB instances owned by your AWS account in the current AWS Region.

database-1

The DB instance identifier is case-insensitive, but is stored as all lowercase (as in "mydbinstance"). Constraints: 1 to 60 alphanumeric characters or hyphens. First character must be a letter. Can't contain two consecutive hyphens. Can't end with a hyphen.

New master password [Info](#)

[REDACTED]

Constraints: At least 8 printable ASCII characters. Can't contain any of the following: / (slash), ' (single quote), " (double quote) and @ (at sign).

Once you have entered the new master password, re-enter it for confirmation, go to the end of the page, click on "Continue" and on the next page, select "Apply Immediately".

4.2) Create and Load dimension tables

Use Table Data Import Wizard on the Workbench to import the dimension tables listed below (that you downloaded from Canvas). 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 `mysqlimport`. In that case you will need to create the dimension tables first. You will only need these 6 tables.

L_AIRLINE_ID.csv

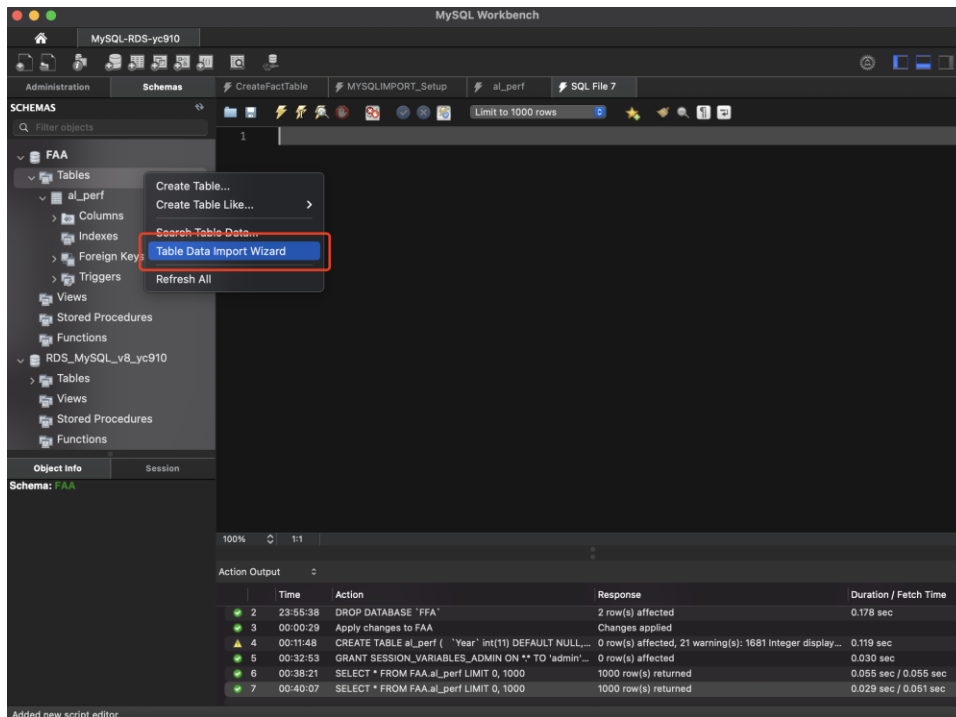
L_AIRPORT.csv

L_AIRPORT_ID.csv

L_DISTANCE_GROUP_250.csv

L_WEEKDAYS.csv

L_CANCELATION



5. Analyze the data

Create and run SQL queries to answer the following questions:

- 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.
- 6a) Check if your dataset has any canceled flights.
- 6b) 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.

6) Submission

Based on the results of your SQL queries and their corresponding results, tell a data-driven story.

Use your data science and visualization skills to demonstrate your findings and insights clearly. Your project should go beyond displaying charts and numbers; it should use data visualizations to build a clear, cohesive narrative.

Think about the following: What key insights do you want your audience to take away?

Design and Communication:

Use thoughtful design choices, clear labeling, and purposeful visualizations to strengthen your message to the audience.

Your final Quarto website should effectively communicate insights through visualization—helping your audience not only see the data but also understand the story behind it and also why your finding/insight matters.

You are encouraged to build a Quarto website and publish it using GitHub Pages.

Submit the URL link to your published website to showcase your work.

Submission Requirements

- **SQL File:**

Submit one .sql file containing:

The question number

Your query

The number of rows returned/affected for each question

(Exactly as you have done in previous assignments.)

- **CSV Files:**

Submit one .csv file for the output of each question.

- **Website Link:**

Submit the URL to your published Quarto website