Homework 4

Mini-project (130 points)

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 mini-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.

3. Look up technical information in the external sources, as needed.

1. Data Model 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 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_CANCELATION.csv	(Code, Reason) CancellationCode
L_WEEKDAYS.csv	(Code, Day) DayOfWeek

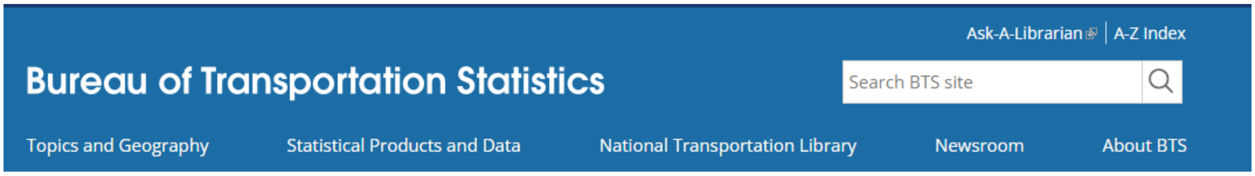
You need to download the Dimension tables for this project from Canvas, folder Mini-project.

2. Download CSV file with fact table data

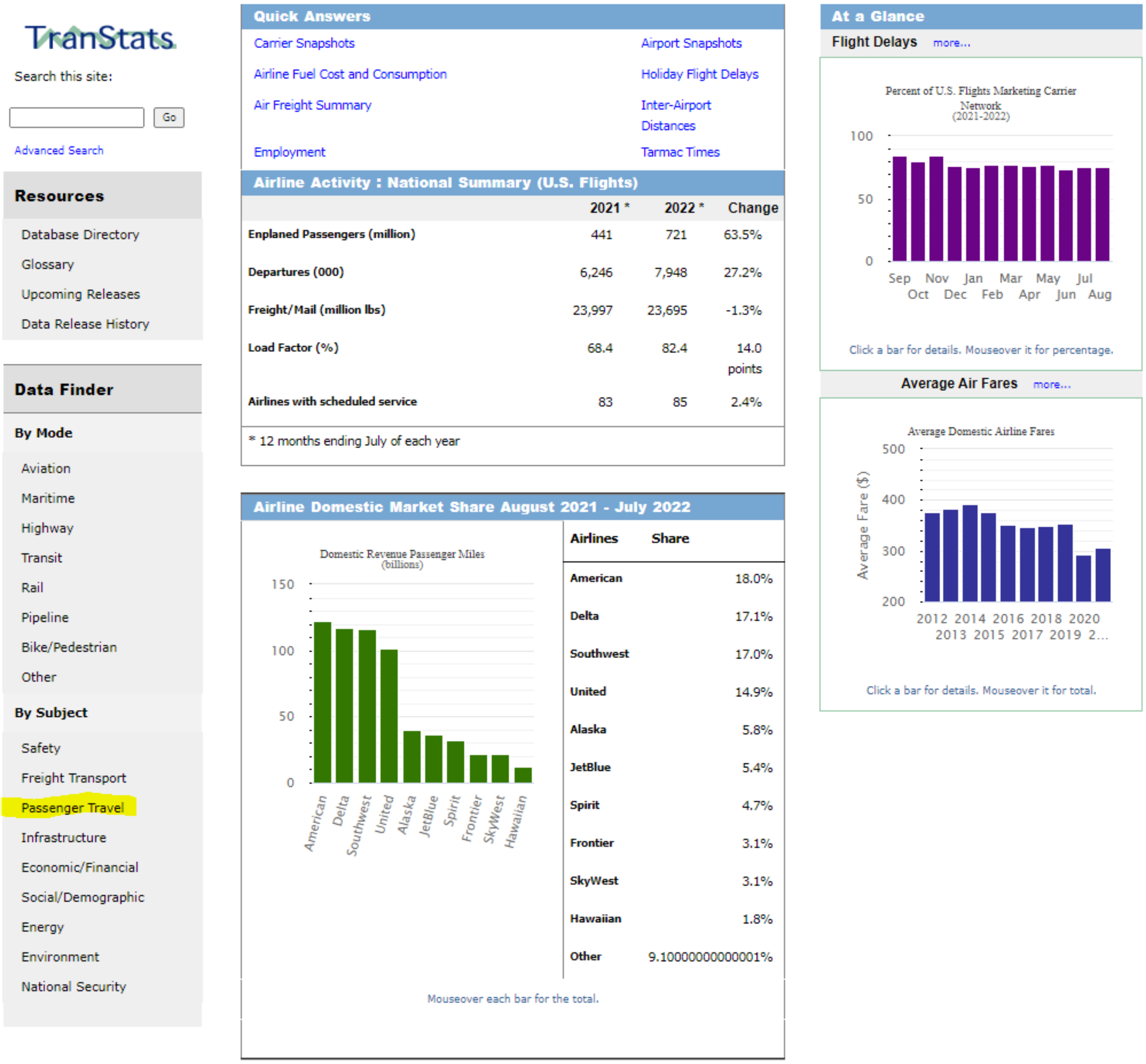
You will be downloading data for the fact table from the web site https://www.transtats.bts.gov/

Details follow:

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



Ask-A-Librarian | A-Z Index
Bureau of Transportation Statistics
Search BTS site
Topics and Geography
Statistical Products and Data
National Transportation Library
Newsroom
About BTS



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
- Social/Demographic
- Energy
- Environment
- National Security

Quick Answers

Carrier Snapshots	Airport Snapshots
Airline Fuel Cost and Consumption	Holiday Flight Delays
Air Freight Summary	Inter-Airport Distances
Employment	Tarmac Times

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

	2021 *	2022 *	Change
Enplaned Passengers (million)	441	721	63.5%
Departures (000)	6,246	7,948	27.2%
Freight/Mail (million lbs)	23,997	23,695	-1.3%
Load Factor (%)	68.4	82.4	14.0 points
Airlines with scheduled service	83	85	2.4%

* 12 months ending July of each year

Airline Domestic Market Share August 2021 - July 2022

Airlines	Share
American	18.0%
Delta	17.1%
Southwest	17.0%
United	14.9%
Alaska	5.8%
JetBlue	5.4%
Spirit	4.7%
Frontier	3.1%
SkyWest	3.1%
Hawaiian	1.8%
Other	9.10000000000001%

Mouseover each bar for the total.

At a Glance

Flight Delays more...

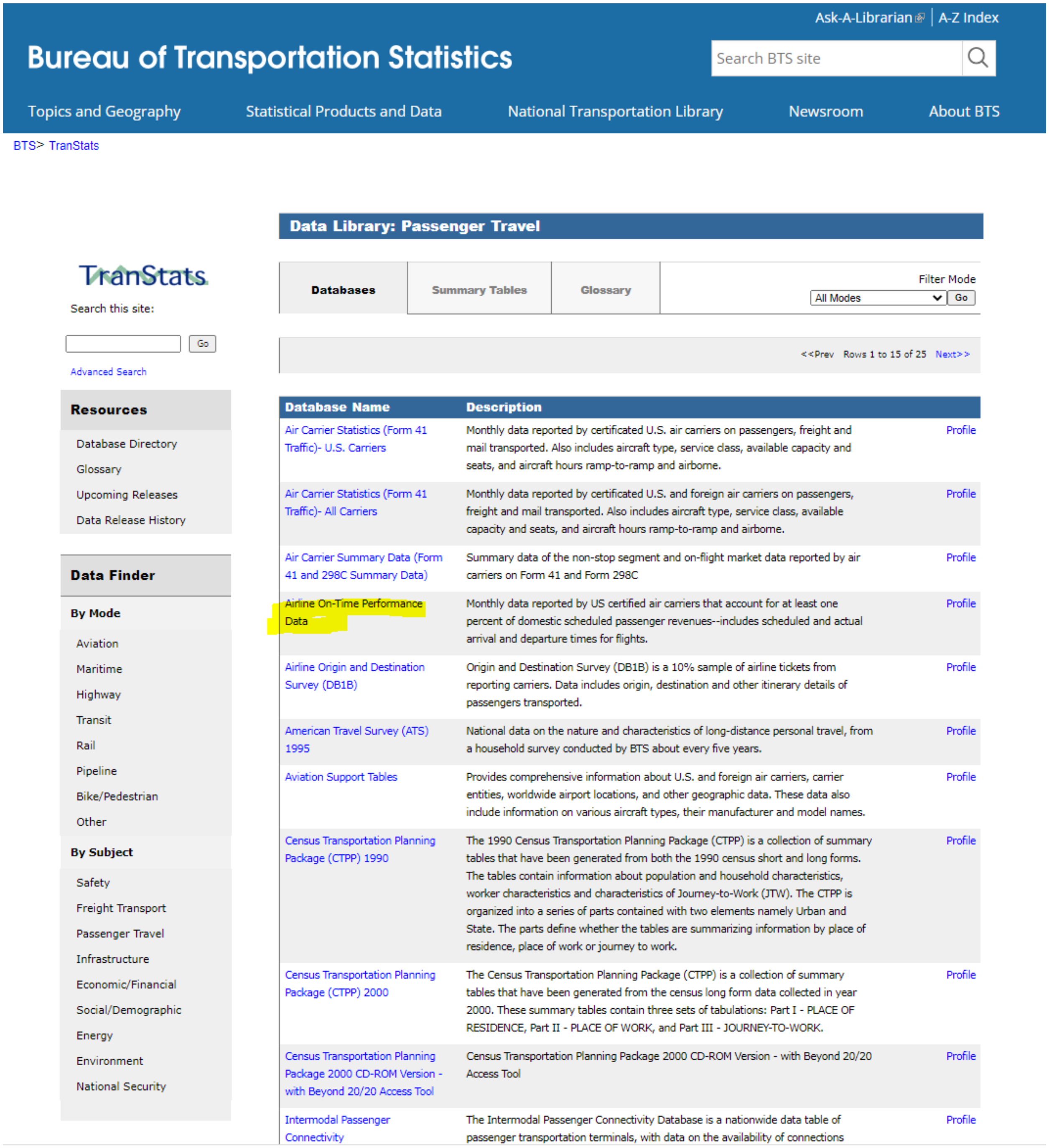
Percent of U.S. Flights Marketing Carrier Network (2021-2022)
100
50
0
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug

Click a bar for details. Mouseover it for percentage.

Average Air Fares more...

Click a bar for details. Mouseover it for total.

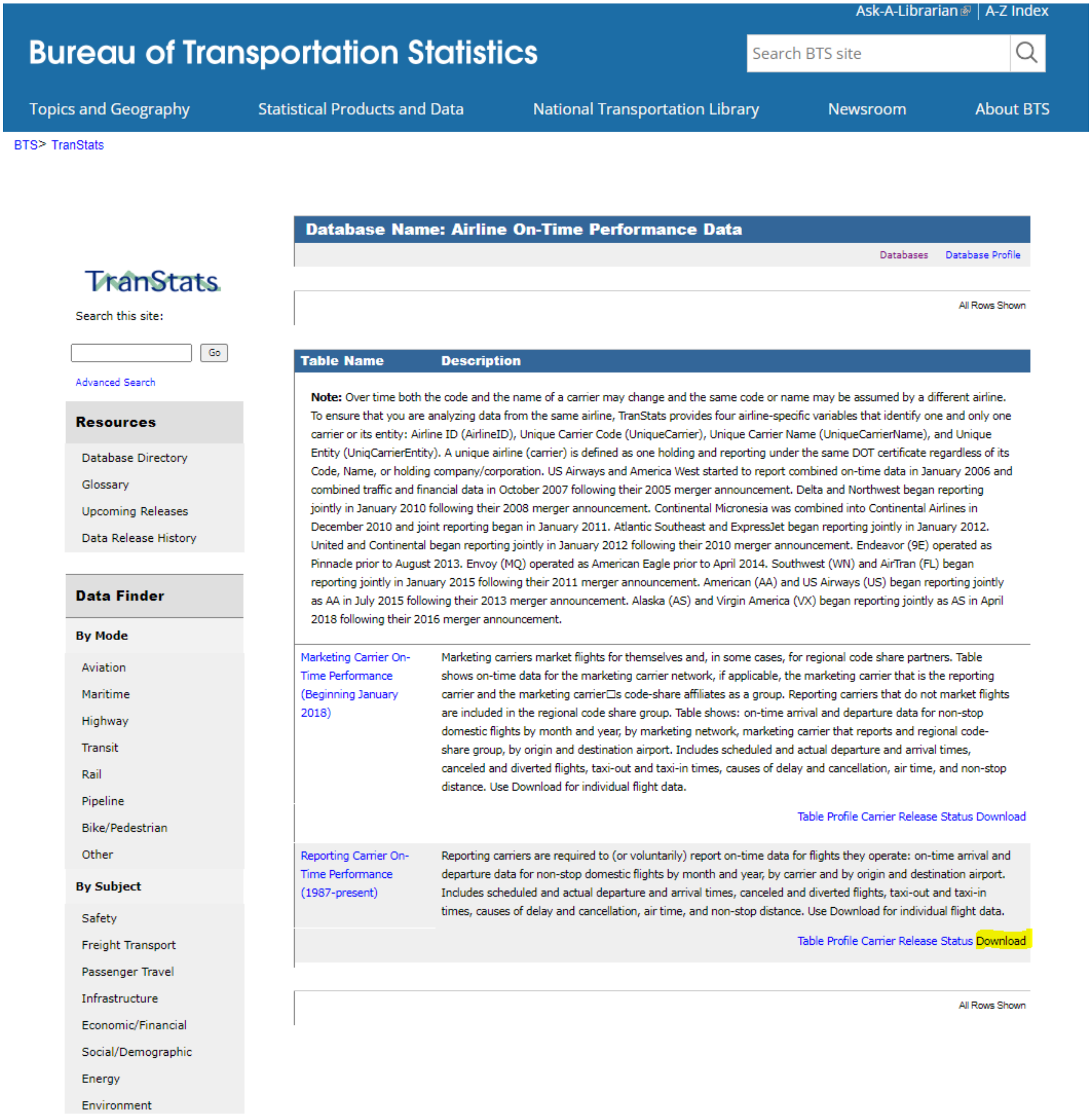
- Then click on “Airline On-Time Performance Data”.



Data Library: Passenger Travel

Database Name	Description	
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
Census Transportation Planning Package 2000 CD-ROM Version - with Beyond 20/20 Access Tool	Census Transportation Planning Package 2000 CD-ROM Version - with Beyond 20/20 Access Tool	Profile
Intermodal Passenger Connectivity	The Intermodal Passenger Connectivity Database is a nationwide data table of passenger transportation terminals, with data on the availability of connections	Profile

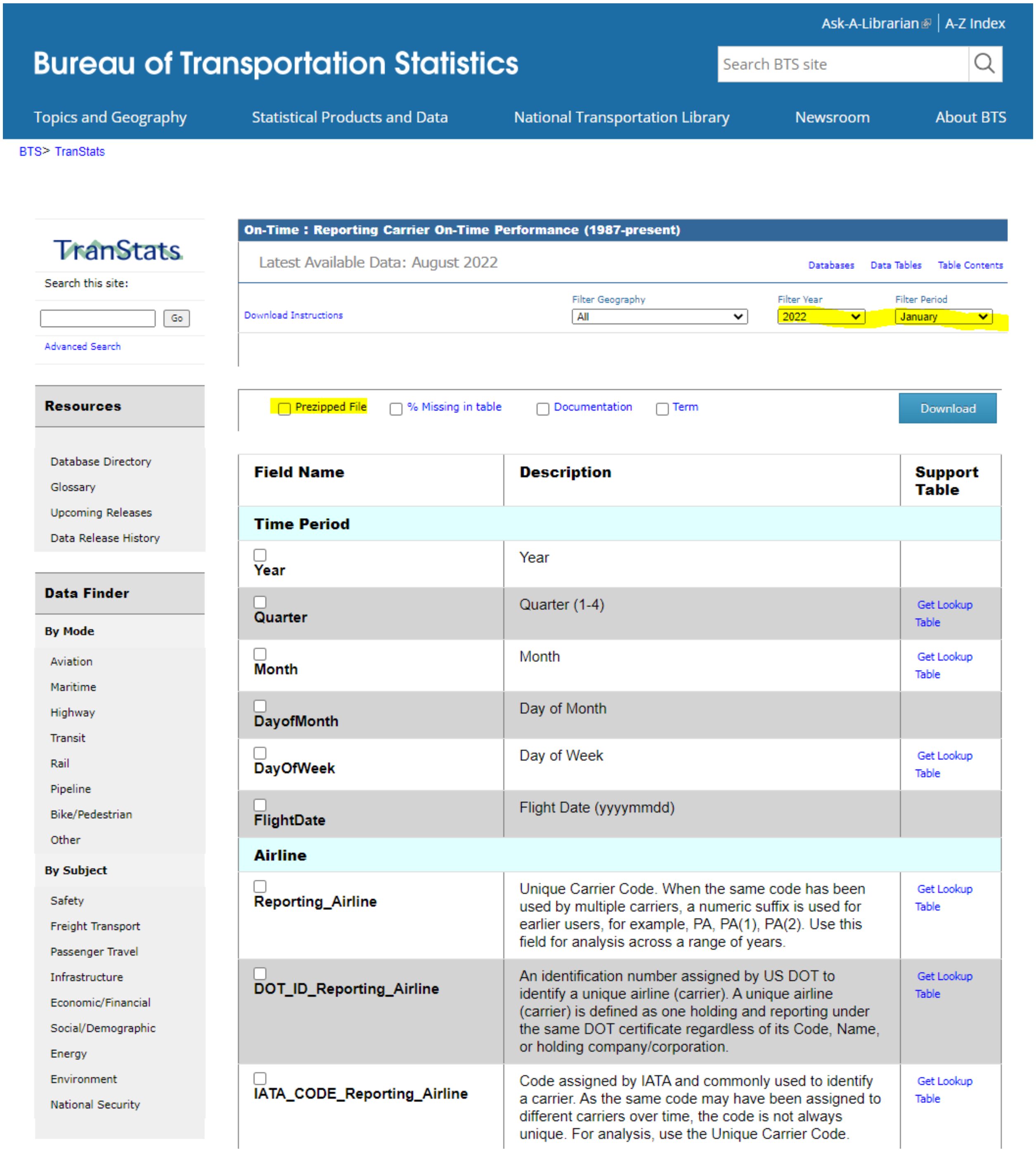
- Finally, in the section for "Reporting Carrier On-Time Performance (1987-present)", click on Download, as highlighted below:



Database Name: Airline On-Time Performance Data

Table Name	Description
	Note: Over time both the code and the name of a carrier may change and the same code or name may be assumed by a different airline. To ensure that you are analyzing data from the same airline, TranStats provides four airline-specific variables that identify one and only one carrier or its entity: Airline ID (AirlineID), Unique Carrier Code (UniqueCarrier), Unique Carrier Name (UniqueCarrierName), and Unique Entity (UniqCarrierEntity). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation. US Airways and America West started to report combined on-time data in January 2006 and combined traffic and financial data in October 2007 following their 2005 merger announcement. Delta and Northwest began reporting jointly in January 2010 following their 2008 merger announcement. Continental Micronesia was combined into Continental Airlines in December 2010 and joint reporting began in January 2011. Atlantic Southeast and ExpressJet began reporting jointly in January 2012. United and Continental began reporting jointly in January 2012 following their 2010 merger announcement. Endeavor (9E) operated as Pinnacle prior to August 2013. Envoy (MQ) operated as American Eagle prior to April 2014. Southwest (WN) and AirTran (FL) began reporting jointly in January 2015 following their 2011 merger announcement. American (AA) and US Airways (US) began reporting jointly as AA in July 2015 following their 2013 merger announcement. Alaska (AS) and Virgin America (VX) began reporting jointly as AS in April 2018 following their 2016 merger announcement.
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. Table Profile Carrier Release Status Download
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

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



On-Time : Reporting Carrier On-Time Performance (1987-present)

Latest Available Data: August 2022

Filter Geography: All

Filter Year: 2022

Filter Period: January

□ Prezipped File □ % Missing in table □ Documentation □ Term

Field Name	Description	Support Table
Time Period		
□ Year	Year	
□ Quarter	Quarter (1-4)	Get Lookup Table
□ Month	Month	Get Lookup Table
□ DayofMonth	Day of Month	
□ DayOfWeek	Day of Week	Get Lookup Table
□ FlightDate	Flight Date (yyyymmdd)	
Airline		
□ Reporting_Airline	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.	Get Lookup Table
□ DOT_ID_Reporting_Airline	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.	Get Lookup Table
□ IATA_CODE_Reporting_Airline	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.	Get Lookup Table

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.

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'@'%';
```

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 the 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.

b) Create EC2 Instance on AWS.

The Document "Create_EC2_Instance_on_AWS_instructions.docx" available in Canvas, folder Mini_project, contains the instructions.

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.

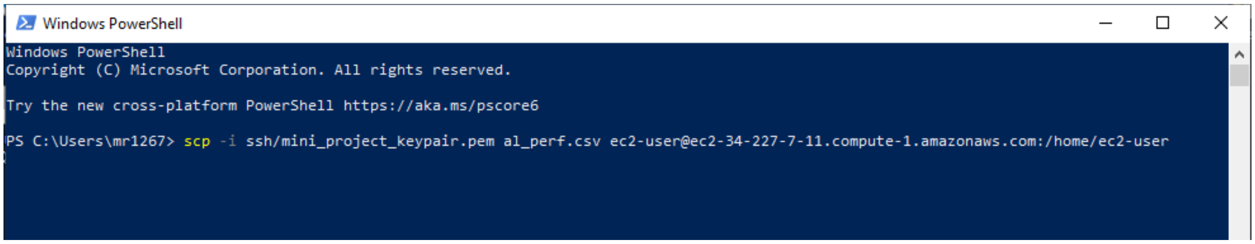
ec2-user@<your_EC_instance_public_IPv4>:/home/ec2-user : is broken down in three parts:

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 Publicc IPv4 address of you 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).

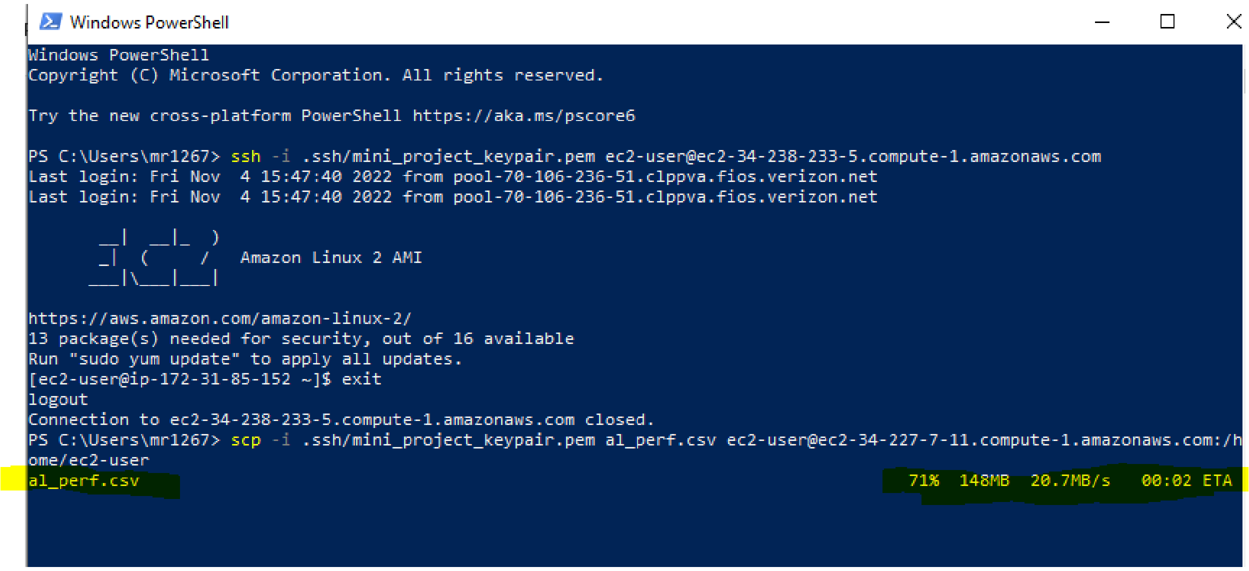


```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

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

PS C:\Users\mr1267> scp -i ssh/mini_project_keypair.pem al_perf.csv ec2-user@ec2-34-227-7-11.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/mini_project_keypair.pem ec2-user@ec2-34-238-233-5.compute-1.amazonaws.com
Last login: Fri Nov  4 15:47:40 2022 from pool-70-106-236-51.clppva.fios.verizon.net
Last login: Fri Nov  4 15:47:40 2022 from pool-70-106-236-51.clppva.fios.verizon.net

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
13 package(s) needed for security, out of 16 available
Run "sudo yum update" to apply all updates.
[ec2-user@ip-172-31-85-152 ~]$ exit
logout
Connection to ec2-34-238-233-5.compute-1.amazonaws.com closed.
PS C:\Users\mr1267> scp -i .ssh/mini_project_keypair.pem al_perf.csv ec2-user@ec2-34-227-7-11.compute-1.amazonaws.com:/h
ome/ec2-user
al_perf.csv                                                                    71%  148MB  20.7MB/s   00:02 ETA
```

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

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

Note: detailed instructions are in the previously mentioned document.

e) Run the following command to install MySQL on your EC2 instance:

```
 $sudo yum install mysql
```

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
```

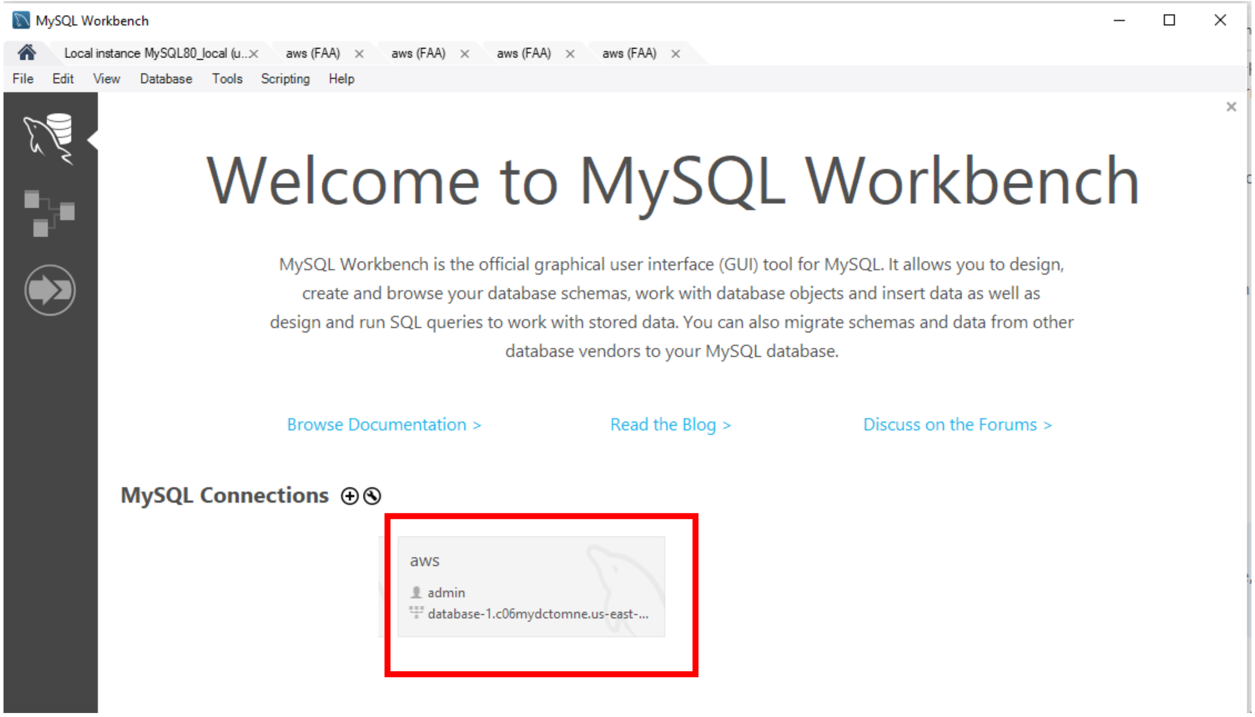
Important:

The host here **IS NOT** the EC2 IPv4 address. It is the database 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 your 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.



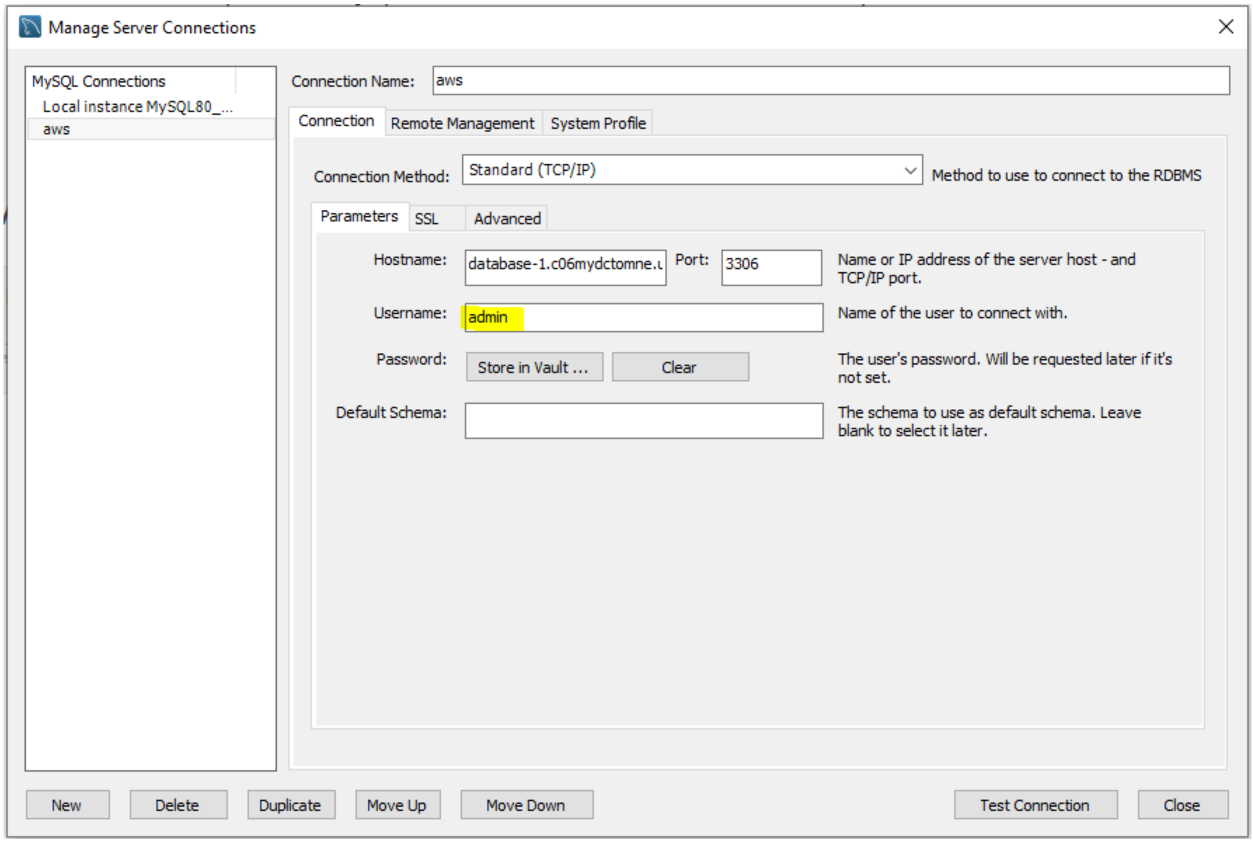
Welcome to MySQL Workbench

MySQL Workbench is the official graphical user interface (GUI) tool for MySQL. It allows you to design, create and browse your database schemas, work with database objects and insert data as well as design and run SQL queries to work with stored data. You can also migrate schemas and data from other database vendors to your MySQL database.

MySQL Connections

aws
admin
database-1.c06mydctomne.us-east-...

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



Connection Name: aws

Connection Method: Standard (TCP/IP) — Method to use to connect to the RDBMS

Hostname: database-1.c06mydctomne.u Port: 3306 — Name or IP address of the server host - and TCP/IP port.

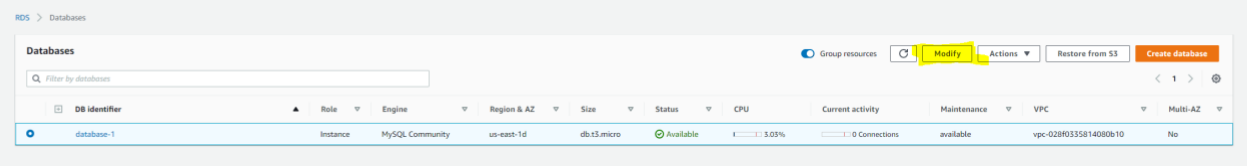
Username: admin — Name of the user to connect with.

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.

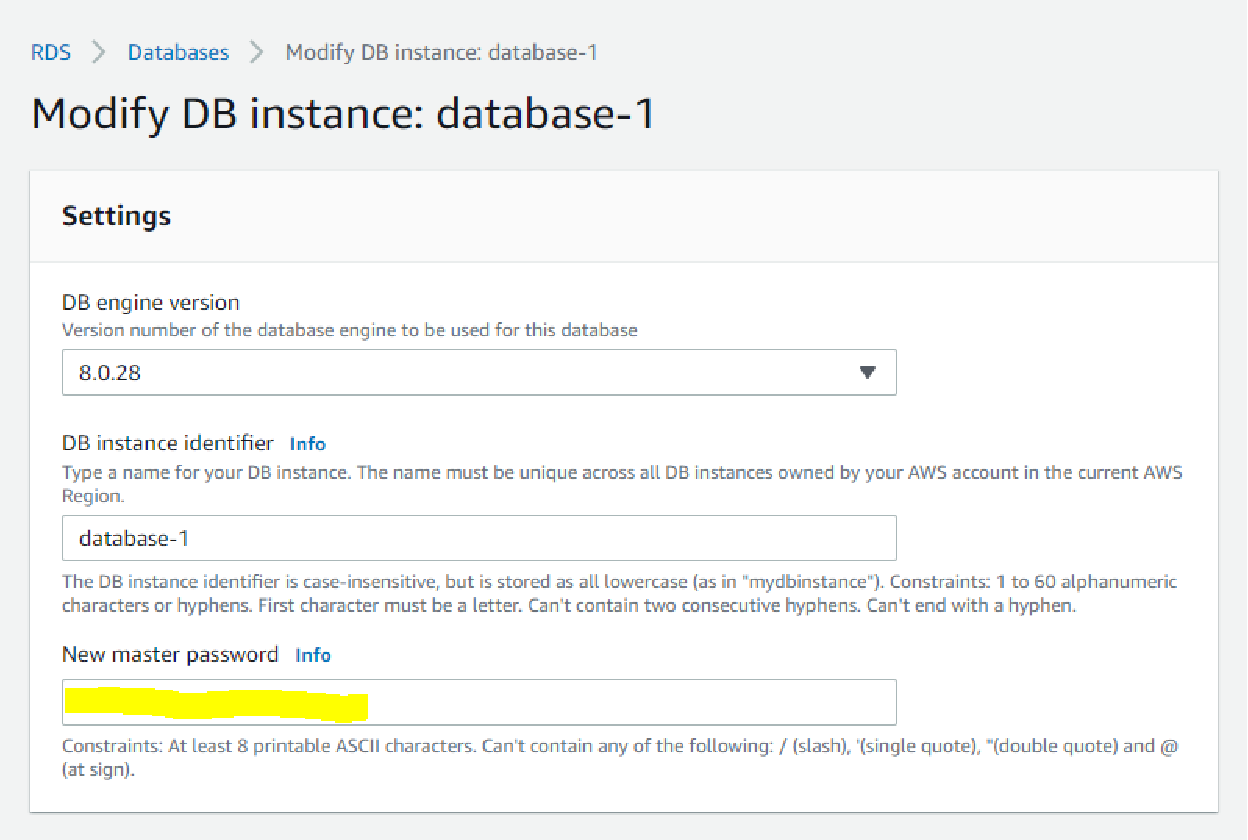
Password – This is the password that you used before when setting up your MySQL server.

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.



	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	3.03%	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
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
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 mysqlinport. In that case you will need to create the dimension tables first. You will only need these 6 tables.

L_AIRLINE_ID.csv

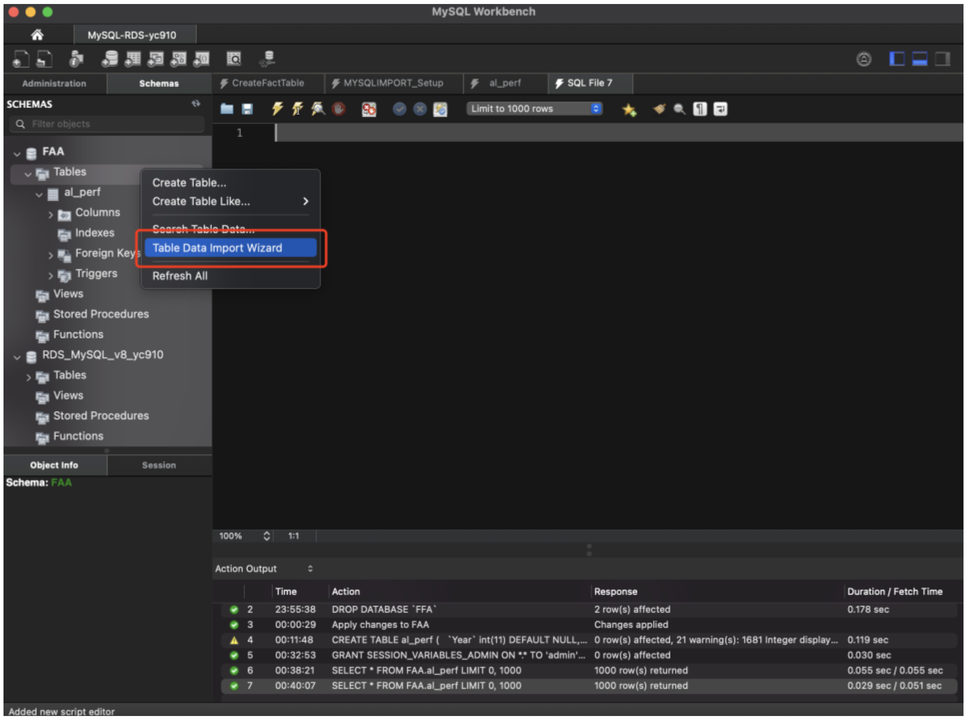
L_AIRPORT.csv

L_AIRPORT_ID.csv

L_DISTANCE_GROUP_250.csv

L_WEEKDAYS.csv

L_CANCELATION



		Time	Action	Response	Duration / Fetch Time
✓	2	23:55:38	DROP DATABASE `FFA`	2 row(s) affected	0.178 sec
✓	3	00:00:29	Apply changes to FAA	Changes applied	
⚠	4	00:11:48	CREATE TABLE al_perf (`Year` int(11) DEFAULT NULL,...	0 row(s) affected, 21 warning(s): 1681 Integer display...	0.119 sec
✓	5	00:32:53	GRANT SESSION_VARIABLES_ADMIN ON *.* TO 'admin'...	0 row(s) affected	0.030 sec
✓	6	00:38:21	SELECT * FROM FAA.al_perf LIMIT 0, 1000	1000 row(s) returned	0.055 sec / 0.055 sec
✓	7	00:40:07	SELECT * FROM FAA.al_perf LIMIT 0, 1000	1000 row(s) returned	0.029 sec / 0.051 sec

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.

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 the average number of flights over the preceding 3 days.

6) Submission

Submit one SQL file with the question number, your query and the number of rows returned for each question, exactly as you submitted all assignments before.

Submit on CSV file for the output of each question.

Please do not submit a zip file.