Data Platform Lab Assignment

#### (Provide your code snippets. You may answer the questions with screenshots if applicable.)

**●** - 1 per group

## Create an Azure Storage account (Data Lake) **●**

Create an Azure Storage Account with a hierarchical namespace, making it a data lake.

Graphical user interface, text, application

Description automatically generated

#### Notes:

* Name your storage account `**dsba6190storage<GROUP>**` (You may have to abbreviate the group name.)
* Make sure your Location is **East US**.
* Under the **Advanced** tab, enable the hierarchical namespace.

#### A screenshot of a computer Description automatically generated

## Create a Data Lake Container

Once you create the Storage Account, create a container called `**datalake**`.

Graphical user interface, text, application, email

Description automatically generated

## Upload data to the data lake.

Upload the [*NYCTaxiCompanies.csv*](https://github.com/colbyford/DSBA6190-CloudComputing/raw/master/2%20-%20Data%20Platform/NYCTaxiCompanies.csv)supplementary data into the data lake container created in the previous step. This can be uploaded using the web UI from the Azure Portal or the Azure Storage Explorer.

Graphical user interface, text, application, email

Description automatically generated

## Create an Azure Synapse Data Warehouse

Graphical user interface, table

Description automatically generated with medium confidence

#### Notes:

* Name the managed resource group `**dsba6190\_<GROUP>-mrg`.**
* Name the Synapse workspace`**dsba6190asa<GROUP>`**.
* Select the Storage Account you created earlier in this lab.
* For the File System, click **Create new** and name it `**datawarehouse**`.

## Create SQL Database

Create a SQL database using the Serverless Pool type from the Data tab by clicking the “**+**” button at the top and then **SQL database**. Name the database `**dsba6190db<GROUP>**`.

Graphical user interface, text, application

Description automatically generated

## Load in the NYC Taxi Data

Browse the gallery in the Data tab of Synapse and load in the “NYC Taxi & Limousine Commission - yellow taxi trip records” data.

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## Create an External Table of the NYC Taxi Data

Create an External Table of the NYC Taxi data using the SQL Pool created in the previous step.

Name the table `**nyctaxiyellow**`.

Graphical user interface

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## Create an External Table out of Supplementary Data

Create an external table out of the uploaded Supplementary Data using a CREATE EXTERNAL TABLE query.

#### Questions:

1. Provide the CREATE EXTERNAL TABLE query you use to create the external table. (Provide the SQL scrip or a screenshot if necessary.)
2. Inner join the supplementary data to the **nyctaxiyellow** external table. Which taxi vendor made the most money in tips in 2010?
3. Using the charting functionality, provide a screenshot of a bar chart that shows a monthly breakdown of City Municipal Transport’s fares in 2009 where passengers paid by credit card.

## Understand Database vs. Data Warehouse Differences

#### Questions:

1. If you were asked to create a database or data warehouse for reporting purposes, which would you choose to create and why? The database or data warehouse will have to pull in data from multiple different systems and be focused on getting large amounts of data aggregated for enterprise-level reports.
2. What is/are the difference(s) between a star schema and a snowflake schema in a data warehouse?

## Understand platform selection differences.

#### Questions:

1. What are the capability differences in using Azure DB vs. Azure Synapse?
2. What are the cost differences in using Azure DB vs. Azure Synapse?
3. What are the capability differences in using Azure Blob Storage vs. Azure Data Lake?
4. What are the cost differences in using Azure Blob Storage vs. Azure Data Lake?

## Industry Use Case

*University Supplies Corporation* needs to create a data backend for their ordering system. Their system will need to be fast and return data back to the user while placing orders through their website. *University Supplies Corporation* has locations all across the United States, but their current website and ordering system is hosted on Azure in the West Central US region.

#### Questions:

1. Would you recommend *University Supplies Corporation* create a database or a data warehouse? Why?
2. Would you recommend *University Supplies Corporation* use Azure SQL DB or Azure Synapse? Why?
3. Which region would you provision the Azure SQL DB/Synapse? (Bonus: Are there any considerations or capability limitations for choosing this region?)
4. If *University Supplies Corporation* is expecting to house ~500GB of data in the database/data warehouse, how much would you expect for that service to cost them per month? List all your assumptions like the number of vCores, Billing Option, Backup Option, etc.