This deliverable consists of 4 parts:

Section 1: Theory (10 points) - 5 questions about topics discussed in class. Please limit each answer to 100 words.

Section 2: Database Design (10 points) - 3 database design questions requiring ER diagrams.

Section 3: Data Analysis with SQL (20 points) - Query a dataset and submit a 1 page report diagnosing problems and providing resolutions and recommendations.

Section 4: Data Visualization on top of SQL (15 points) - Build an operational dashboard using Google Data Studio.

Please read the directions for each section carefully before starting work.

Section 1: Theory (10 points)

In this section, you will answer 5 questions about topics we discussed in class. Each question is worth 2 points. Please limit each answer to 100 words.

1. What is a Data Moat? Why is it important to have one?
2. What is the difference between OLAP and OLTP Databases? Why would you choose one over the other?
3. What are the 3 different roles in a modern data team? Which problems do each of them solve? How do they compare with each other?
4. What is the difference between the WHERE and HAVING clauses?
5. How would you define the relationship between employees and offices in the Entity Relationship (ER) model? Please provide an explanation why using real world examples.

Section 2: Database Design (10 points)

In this section, you will answer 3 database design questions. The first two questions are worth 3 points each and the last question is worth 4 points. Please do not forget to provide information about **entities**, **relationships**, and **attributes** for each question to get full marks.

1. You are asked to model the many to many relationship between students and classes in a relational database.
    - What changes do you need to make to support this relationship?
    - Please create an ER diagram to show how these entities will relate to each after your changes.

STUDENTS M:M CLASSES

2. You are asked to model the many to many relationship between customers and products in a relational database.
    - What changes do you need to make to support this relationship?
    - Please create an ER diagram to show how these entities will relate to each after your changes.

CUSTOMERS M:M PRODUCTS

3. Design an ER diagram for a library reservation system for a family of libraries based on the given characteristics.
    - This system is for multiple libraries
    - This system is for multiple borrowers
    - There are multiple types of content that can be borrowed
    - Borrowers can borrow multiple items at the same time
    - Borrowers can borrow multiple types of content

Be sure to list all necessary entities, relationships, and attributes to model this system in a relational database

Section 3: Data Analysis with SQL (20 points)

In this section, you're going to use the data set **new_york_citibike** (under **bigquery-public-data)** in Google BigQuery to answer some business questions using SQL. Take some time to familiarize yourself with the data set before answering your questions.

Your output will be a 1 page report, which diagnoses the problems you see, provides a few potential resolutions, and recommends one solution with a justification of why. The report must fit on one page.

You will also submit an appendix, which includes all the SQL you ran to get to your answer and any tables, maps, or charts you think are helpful to make your point. Please add a comment on top of each figure in your appendix to explain what insight it is providing.

Connecting to BigQuery

1. Navigate to the BigQuery UI (Links to an external site.)
2. If necessary, start a New Project
3. Click “ADD DATA”
    -
4. Click “Pin a project”
    -
5. Paste “bigquery-public-data" in the “enter a project name” text box and click PIN
    -
6. Search for the new_york_citibike dataset
    -

Business Questions

You’ve been told by customer support that customers frequently complain about bike stations being empty. You need to analyze the data in your data set to understand this problem and make suggestions about how to address it. Some items to consider are below. **Please note that the questions below are just a guiding point for your analysis. You don’t need to explicitly answer them all:**

- Can you find any traces of empty stations?
    - If yes, how big is this problem?
- What are the most popular stations in the network?
    - When does their usage peak?
- What are the most popular trips in the network?
- Are there differences in the types of rides that people take?
- Is there a pattern in the types of stations that are empty?

Your output will be a 1 page report, which diagnoses the problems you see, provides a few potential resolutions, and recommends one solution with a justification of why. The report must fit on one page.

Potentially useful resources

https://cloud.google.com/bigquery/docs/reference/standard-sql/date_functions#date_trunc (Links to an external site.)
https://cloud.google.com/bigquery/docs/reference/standard-sql/timestamp_functions#extract (Links to an external site.)
https://cloud.google.com/bigquery/docs/gis-getting-started (Links to an external site.)
https://bigquerygeoviz.appspot.com/ (Links to an external site.)

Section 4: Data Visualization on Top of SQL (15 Points)

In this section, you're going to build an operational dashboard, using Google Data Studio (Links to an external site.), to track the health of your bike system. Use the same data set as in section 3. You will paste a screenshot of your response in your assignment submission and share your report with so we can take a direct look at it.

Build an operational dashboard to answer the following business questions:

Station Health

- How many stations are at capacity, empty, or out of service?
- What is the fill rate(bikes available/capacity) for each station?
- What is the most popular station to start rides for all time?
- What is the most popular station to end rides for all time?
- What are the top 3 most popular trips (start and end station combination) for all time?
- Which hours of the day does usage peak on weekdays?
- Which hours of day does usage peak on weekends?

System Health

- How many trips are there per day?
- What is the average trip duration?
- What was the shortest trip?
- What was the longest trip?

- How many total hours of usage does each bike have?

Potentially useful resources

https://webflow-blog.periscopedata.com/blog/periscope-datas-visualization-flow-chart (Links to an external site.)

Outputs

1. Share a screenshot of each of your dashboard pages

Final Assignment Checklist

Please ensure you have all of the following items in your assignment submission. The assignment can be submitted in word or PDF format.

Section 1

- **1-5:** Answers to each of the 5 questions. No more than 100 words each

Section 2

- 1-2: Text answers to the initial question and ER diagrams for the updated relationship
- 3: An ER diagram for the library reservation system

Section 3

- A one page report
- All SQL code in an appendix
- Any relevant tables, maps, or charts for the analysis in an appendix

Section 4

- Screenshots of your dashboard (the dashboard can be on multiple pages)