Hult International Business School

Data Management and SQL

Prof. Andrew Peynado

Mauricio Marcon Teles
Divisadero – Cohort 4
MSBA – 2019/2020

Section 1: Theory (10 points)

1. What is Data Moat? Why is it important to have one?

   Data Moat is a metaphor to represent the importance of an organized and enriched data structure. By setting correctly the data strategy, where to store the data, how it's going to be accessed, what they represent, and which insights it will generate, the company creates protection against its competitor through its differentiated assets. The amount of data makes a difference, too, once it takes time to collect and space to store. Therefore, companies that are collecting data for a more extended period are highly evaluated because they can build better AI models based on their data sample.

2. What is the difference between OLAP and OLTP Databases? Why would you choose one over the other?

   OLTP represents transactional processing, represented by fast and numerous operations with the database. They are usually performing inserts, updates, deletes, and browsing data.

   OLAP is related to analytical processing, which mainly uses batches of stored data to process massive amounts of registers to consolidate, report, and analyze to extract insightful information.

   For operations that will represent a user journey (purchase, withdraw, or texting), I would use an OLTP, once the database operations must happen in a fastly and uniquely. On the other hand, to understand how the user behavior and where it's possible to improve the process, I would use an OLAP to identify patterns and generate informative models from the data.

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?

   Data Engineer: sets up and prepares the infrastructure to make data accessible for users. Customize the databases and provide support in using its tools

   Data Analyst: connects business and data. Accesses SQL and non-SQL databases, code programs, or scripts (Python or R) to clean and complete the data to proceed with analysis and conclusions.

   Data Scientist: builds and refine prediction models and algorithms to support decisions and implement machine learning and artificial intelligence structures.

   All of them are related to data, but the engineer provides the base that will store the data, the analyst access, and uses the raw data to connect with the business. The scientist manages to improve the business supported by his models.

4. What is the difference between the WHERE and HAVING clauses?

   Both are used to filter the output of a query. The difference is in the level that they work.

   WHERE is used to filter the records from a table considering its values and can always be used.

   HAVING is used to filter grouped and aggregated values and can only be used with grouped outputs.

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.

   Nowadays, with the growth of remote work and the constant need for travel, I see this relationship as a many-to-many. However, for administrative purposes, employees have a “home” office in which he is connected, and in this case, it is a one-to-many.

   One employee belongs to one office, and one office hosts many employees.

   I visited the Google office in Cambridge, and that office hired most people who were working there, but the office allows Google visitors employees to go there to work.

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 relationships between students and classes in a relational database.
   - What changes do you need to make to support this relationship?

   I will create one entity named Program and other named Program_Classes.

   Student will be assigned to a Program and Program_Classes will relate Classes and Programs

   One student attends one program, and one program contains N students.

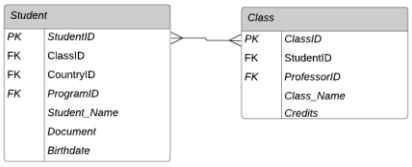
   One program contains N Classes, and one Class is taught in N programs, which demands the creation of the relationship Program_Classes

   The main attributes are:

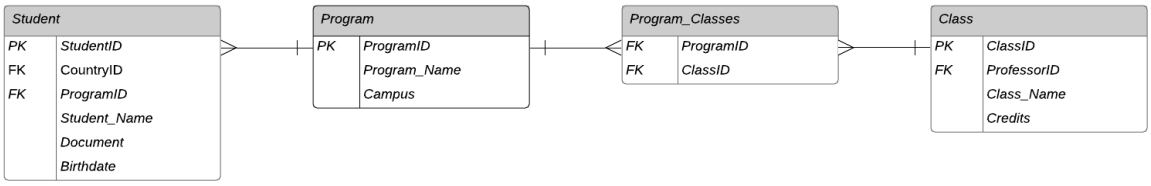
   - StudentID: primary key for Student
   - Program ID: primary key for Program and used as the foreign key in Student and Program_Classes
   - ClassID: primary key for Class and foreign key in Program_Classes

   - Please create an ER diagram to show how these entities will relate to each after your changes.

Before



After



2. You are asked to model the many to many relationships between customers and products in a relational database.
   - What changes do you need to make to support this relationship?

   The solution is similar to the previous question. I will create one entity named Order and other named Order_Items.

   Client purchase an Order and Order_Items will relate Orders and Products

   One client purchases N orders, and one order belongs to one client.

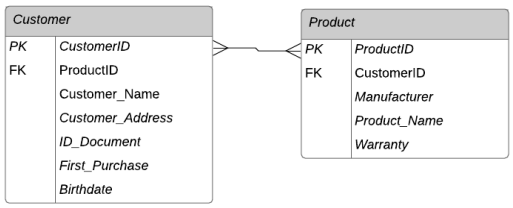
   One Order contains N Products, and one Product is bought in N Orders, which demands the creation of the relationship Order_Items

   The important attributes are:

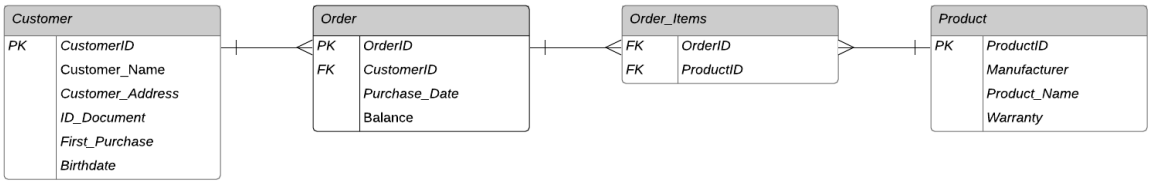
   - ClientID: primary key for Client and foreign key for Order
   - OrderID: primary key for Order and used as the foreign key in Order_Items
   - ProductID: primary key for Product and foreign key in Order_Items

   - Please create an ER diagram to show how these entities will relate to each after your changes.

Before

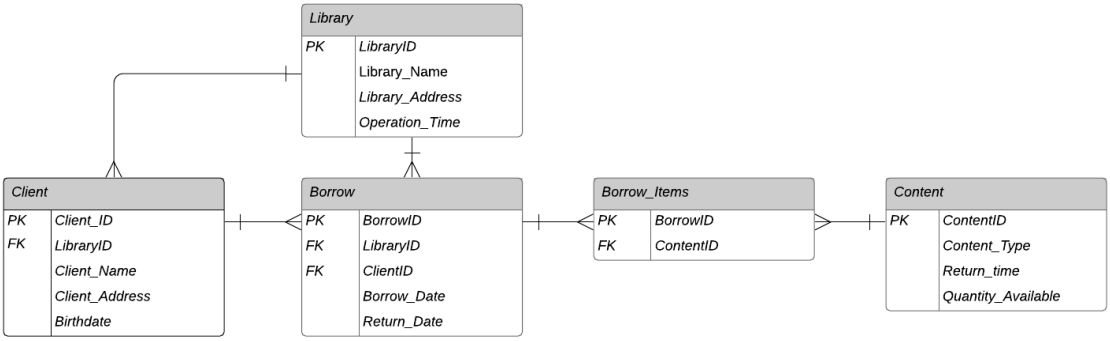


After



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



The client is connected to a library, and the library may have many clients.

The client can borrow in any library, so Borrow has an ID and Library and Client ID as well.

A Borrow belongs just to one client and to one library. Through the relationship Borrow_Itens, the client can borrow many Contents from many types, identified by the field Content_Type

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.

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?

New York City Bikes Analysis

Based on received complaints of bicycle availability in the city, I investigated to identify locations and reasons why people are not able to find bikes to rent and ride.

Firstly, my approach was to understand the locations and conditions of bicycles. What surprised me is that, among the stations that I considered able to hold bicycles (Appendix 1), only 55 had had a low number of bikes (Appendix 2). Since sometimes people don't ride bikes alone, I set a threshold of 2 bicycles per station.

Secondly, I analyzed the trips to search where most of the trips start and which ride (start – end) is most prevalent. The most demanding station is Pershing Square North (Appendix 3), which makes sense once this station is close to the Grand Central Terminal, and people may use the bike to commute. Pershing Square North is not the start of the most frequent ride, but it appears three times in the top 15 rides (Appendix 5). However, this station wasn't in the list of stations with low offerings.

Lastly, I joined both results. The most popular rides that started in stations with a low number of bikes. This search led me to West St & Chambers St and Pier 40 – Hudson River Park (Appendix 6). These stations are the start (and often, the end) of most popular rides in the list of the stations with few bikes.

By looking at their location (Appendix 7), it's possible to see that these stations are close to the Hudson River what makes this ride attractive to tourists and groups.

Based on this information, I looked for stations that are close to these spots. There are a few that do not show the same low availability of bikes (Appendix 8).

Based on that, I suggest the company to:

- Increase the awareness of users of the stations close to West St & Chambers St and Pier 40 – Hudson River Park. It will help the customers find options and skip complaining.
- Increase the capacity in these stations. The number of trips is starting at these stations is high and justifies the investment.
- Incentivize users to return their bikes on these stations by offering them a discount on the ride fee.

Appendix

Appendix 1. 6 stations with Capacity = 0 seems to be that are not operating– need to remove them.

```
SELECT *
FROM `bigquery-public-data.new_york_citibike.citibike_stations`
where capacity = 0
LIMIT 1000
```

name	latitude	longitude	capacity	num_bikes_available	num_bikes_disabled	num_docks_available	num_docks_disabled	is_installed	is_renting	is_returning
Myrtle Ave & Marcy Ave	40.69539817	-73.94954908	0	0	0	0	0	TRUE	TRUE	TRUE
E 106 St & Lexington Ave	40.791976	-73.945993	0	0	0	0	0	TRUE	TRUE	TRUE
W Broadway & Spring St	40.72494672	-74.00165856	0	0	0	0	0	TRUE	TRUE	TRUE
31 Ave & Steinway St	40.7611488	-73.9170071	0	0	0	0	0	TRUE	TRUE	TRUE
Grand St	40.71517768	-74.03768331	0	0	0	0	0	TRUE	TRUE	TRUE
Pierrepont St & Monroe Pl	40.69535693	-73.99344027	0	0	0	0	0	TRUE	TRUE	TRUE

Appendix 2. 10 stations were empty – few disabled bikes. 22 with just one bike and 23 with 2 bikes

```
SELECT *
FROM `bigquery-public-data.new_york_citibike.citibike_stations`
where capacity <> 0
and num_bikes_available <= 2
order by num_bikes_available
LIMIT 1000
```

Name	latitude	longitude	num_bikes_available	num_bikes_disabled
Norman St & Wyckoff Ave	40.69517	-73.90311	0	0
Withers St & Kingsland Ave	40.71773	-73.94051	0	0
Madison St & Evergreen Ave	40.69122	-73.91693	0	0
Linden St & Knickerbocker Ave	40.69714	-73.91566	0	0
Hancock St & Wyckoff Ave	40.6972	-73.90674	0	0
Harman St & Seneca Ave	40.70577	-73.91292	0	2
Bushwick Ave & McKibbin St	40.705517	-73.93936	0	1
31 Ave & 34 St	40.763154	-73.920827	0	0
14 St & 7 Ave	40.663779	-73.98396846	0	0
5 Ave & E 126 St	40.80698	-73.941747	0	0

Appendix 3. The most recent trips show Pershing Square North, W 21 St & 6 Ave, and Broadway & E 22 St as most popular stations

```
SELECT start_station_name, count(*) as count_trips, extract (year from starttime) as year
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
group by start_station_name, year
order by 3 desc, 2 desc
LIMIT 1000
```

start_station_name	count_trips	year
Pershing Square North	55990	2018
W 21 St & 6 Ave	37751	2018
Broadway & E 22 St	36717	2018
E 17 St & Broadway	35856	2018
W 41 St & 8 Ave	31420	2018
Broadway & E 14 St	31379	2018
W 33 St & 7 Ave	28543	2018
West St & Chambers St	27972	2018
Broadway & W 60 St	27956	2018
Lafayette St & E 8 St	26820	2018
W 31 St & 7 Ave	26576	2018
E 47 St & Park Ave	26133	2018
8 Ave & W 33 St	26072	2018
8 Ave & W 31 St	25584	2018
Christopher St & Greenwich St	25449	2018

Appendix 4. Pershing Square North was always the most demanded station. 8 Ave & W 31 St was in the past years, but not too much lately. Broadway & E 22 St and W 21 St & 6 Ave, spiked since last year.

```
SELECT start_station_name, count(*) as count_trips, extract (year from starttime) as year
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
group by start_station_name, year
order by 2 desc
LIMIT 1000
```

start_station_name	count_trips	year
Pershing Square North	135906	2017
Pershing Square North	111892	2016
Pershing Square North	104813	2015
8 Ave & W 31 St	100796	2015
8 Ave & W 31 St	100498	2014
Lafayette St & E 8 St	95890	2015
E 17 St & Broadway	95863	2017
West St & Chambers St	95233	2017
Broadway & E 22 St	93083	2017
W 21 St & 6 Ave	90696	2017
W 21 St & 6 Ave	87149	2015
Lafayette St & E 8 St	86692	2014
E 17 St & Broadway	86108	2015

Appendix 5. Even though the most popular trip doen't start in Pershing Square, the station appears three times in the top 15 trips, with short duration trips (less than 500 seconds)

```
SELECT start_station_name, end_station_name, count(*) as count_trips, avg(tripduration) as duration, extract (year from starttime) as year
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
where extract (year from starttime) = 2018
group by start_station_name, end_station_name, year
order by 3 desc
LIMIT 1000
```

start_station_name	end_station_name	count_trips	duration	year
E 7 St & Avenue A	Cooper Square & Astor Pl	3116	264	2018
Central Park S & 6 Ave	Central Park S & 6 Ave	2071	2812	2018
Grand Army Plaza & Central Park S	Grand Army Plaza & Central Park S	1874	4038	2018
S 4 St & Wythe Ave	N 6 St & Bedford Ave	1835	283	2018
Central Park S & 6 Ave	5 Ave & E 88 St	1833	1471	2018
W 21 St & 6 Ave	9 Ave & W 22 St	1827	305	2018
W 63 St & Broadway	Broadway & W 60 St	1814	185	2018
Pershing Square North	Broadway & W 32 St	1789	419	2018
Pershing Square North	W 33 St & 7 Ave	1758	493	2018
N 6 St & Bedford Ave	S 4 St & Wythe Ave	1705	320	2018
Willoughby St & Fleet St	Adelphi St & Myrtle Ave	1618	310	2018
DeKalb Ave & Vanderbilt Ave	DeKalb Ave & Hudson Ave	1593	260	2018
Greenwich Ave & Charles St	Greenwich Ave & Charles St	1579	1262	2018
Richardson St & N Henry St	Graham Ave & Conselyea St	1553	201	2018
S 3 St & Bedford Ave	N 6 St & Bedford Ave	1545	200	2018
Pershing Square North	E 24 St & Park Ave S	1508	415	2018

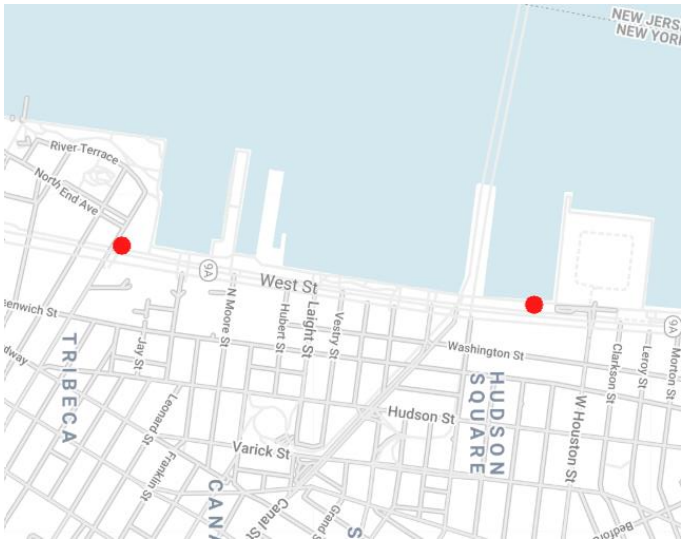
Appendix 6. Crossing the empty stations with the most demanded trips, West St & Chambers St and Pier 40 – Hudson River Park seems to be the bottleneck as start stations.

```
SELECT start_station_name, end_station_name, count(*) as count_trips, avg(tripduration) as duration, extract (year from starttime) as year
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
left join `bigquery-public-data.new_york_citibike.citibike_stations`
on start_station_id = station_id
where extract (year from starttime) = 2018
and  capacity <> 0
and num_bikes_available <= 2
group by start_station_name, end_station_name, year
order by 3 desc
LIMIT 1000
```

start_station_name	end_station_name	count_trips	duration	year
Pier 40 - Hudson River Park	West St & Chambers St	1140	513.6578947	2018
West St & Chambers St	Pier 40 - Hudson River Park	978	538.4458078	2018
11 Ave & W 27 St	W 26 St & 8 Ave	966	326.9699793	2018
West St & Chambers St	Washington St & Gansevoort St	815	993.3300613	2018
West St & Chambers St	12 Ave & W 40 St	809	1419.930779	2018
West St & Chambers St	Little West St & 1 Pl	806	646.8883375	2018
West St & Chambers St	Centre St & Chambers St	800	450.91	2018
Pier 40 - Hudson River Park	South End Ave & Liberty St	786	869.6412214	2018
Driggs Ave & N Henry St	Graham Ave & Conselyea St	767	362.8174707	2018
West St & Chambers St	Christopher St & Greenwich St	751	627.4567244	2018
West St & Chambers St	West St & Chambers St	706	1801.865439	2018
Pier 40 - Hudson River Park	Little West St & 1 Pl	699	1272.597997	2018
West St & Chambers St	Greenwich St & W Houston St	679	523.0191458	2018
Greenwich St & W Houston St	West St & Chambers St	672	587.5654762	2018
Pier 40 - Hudson River Park	12 Ave & W 40 St	632	1256.992089	2018
11 Ave & W 27 St	8 Ave & W 31 St	625	361.976	2018
West St & Chambers St	West Thames St	610	478.0147541	2018

Appendix 6. It's Hudson River margins. probably a touristic point and ride

```
SELECT name, num_bikes_available, ST_GeogPoint(longitude, latitude)  AS WKT
FROM `bigquery-public-data.new_york_citibike.citibike_stations`
where capacity <> 0
and num_bikes_available <= 2
and station_id in (426, 3256)
order by num_bikes_available
LIMIT 1000
```



Appendix 7. The region offers some options to find bikes around.

```
SELECT name, num_bikes_available, ST_GeogPoint(longitude, latitude)  AS WKT
FROM `bigquery-public-data.new_york_citibike.citibike_stations`
where capacity <> 0
order by num_bikes_available
LIMIT 1000
```

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?

Outputs

1. Share a screenshot of each of your dashboard pages
2. Share your report with datamanagementandsql2020@gmail.com
3. Share all data sources with datamanagementandsql2020@gmail.com

Link: https://datastudio.google.com/s/pUnqeln2VlU