Homework 4
CSCI 585 – Database Systems

Due Date: Monday, July 8th, 2019 at 11:59pm PST.

In this homework, we will use google cloud platform. You have done the basic setup in HW#2.
Objectives:
- Exploiting integrated cloud platforms for variety of data analysis tasks
- Working with Big Datasets in cloud
- Using Notebooks for generating reports
- Visualization and information retrieval

"We have to do better at producing tools to support the whole research cycle—from data capture and data curation to data analysis and data visualization." - Jim Gray[1]

Part 1: Google BigQuery (2pt)

a) Introduction:

With big data, we've come to expect big responsibility. To handle the complexity of today's data, you need to spend a lot on hardware, plan ahead for scalability, and pour hours into system architecture. And this just gets you up and running. You still need someone (most probably a team) to make sure everything's running smoothly. Even after all that headache, running queries can still take anywhere from minutes to hours, and mistakes can be costly. Wouldn't you rather focus on finding insights from your data than developing the infrastructure around it?
Google BigQuery is a fully managed data warehouse that removes set up hassles and runs queries rocket fast, fast enough to analyze terabytes of the data in seconds, petabytes in minutes.
What if you have a really big query? It's simple. Google will spin up an entire data center to quickly process it for you. Google BigQuery encrypts, replicates, and deploys your data across multiple data centers for maximum durability and service up time.
With just a couple of clicks, you can control where you store your data. Sharing and collaboration are easy as well. You decide who can access your data. And because you can use standard SQL queries, anyone can get involved. Moreover, BigQuery integrates with Google Cloud Platform products and other software, so you can readily load, process, and make interactive visualizations of your data.[2]

---

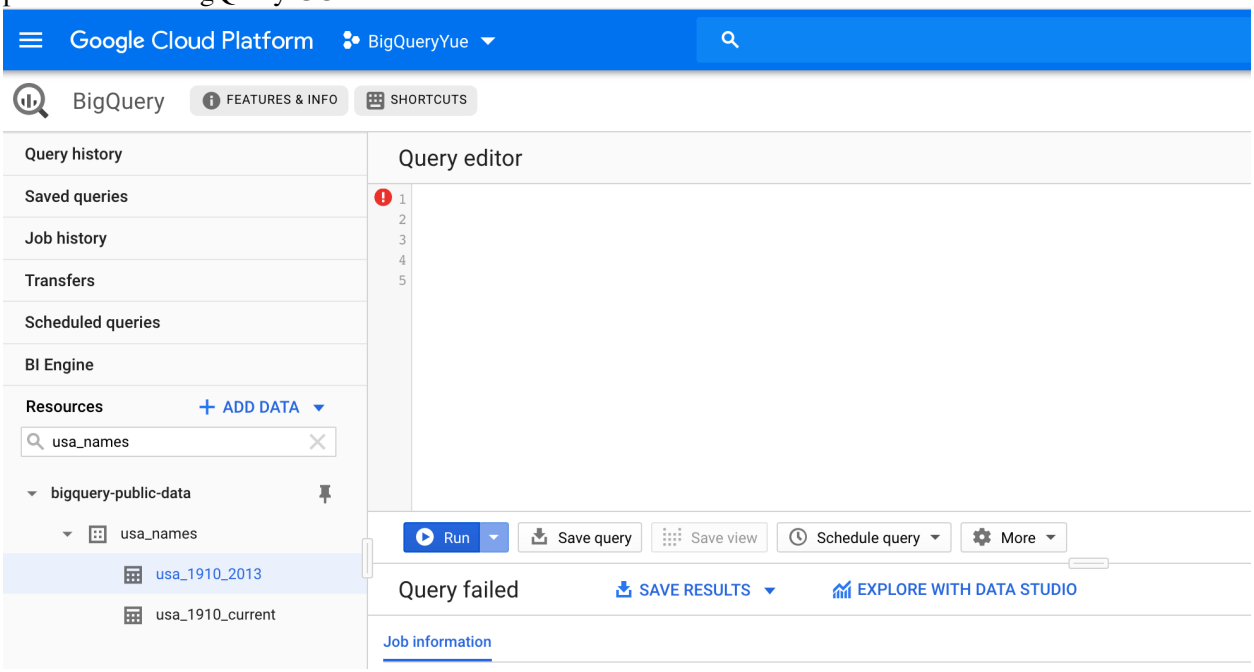
1 Jim Gray was (1944-2007) was a computer scientist who received the Turing Award in 1998 for his work on databases.
2 You can watch this video for a more complete tutorial on Google BigQuery

b) Starting with Google BigQuery:

To start with using BigQuery Google, go to https://cloud.google.com/bigquery/. You need to sign in to google cloud platform with the same user used for HW#2. Follow "**View documentation**" and then "**Quickstarts**". In order to make all the steps integrated, we will use web UI in GCP console, select "**Quickstart Using the classic web UI**", and then click on the "**BigQuery Web UI**" in the **caution message** or simply click(https://cloud.google.com/bigquery/docs/bigquery-web-ui?refresh=1).
Follow with the tutorial step by step (but do not clean up!).

Tasks:

Use functional programming (refer to: https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators) to complete the following tasks for a public available dataset "usa_names.usa_1910_2013". To check the schema and description of the table, you could search for the "usa_names" table in the bigquery-public-data in BigQuery GCP console.



Query #1

Using the "bigquery-public-data.usa_names.usa_1910_2013" table, **query names, gender and total counts of the names *which have the second letter "o" in the name (e.g. Rock) AND are from the gender 'F'*, limit the result to 10 rows.** Please note that you need differentiate between the gender. The results should be in descending order of name. You need to use the SUBSTR function in this query. **You need include the query (text) and the result (either copy and paste the text or take a screenshot) in your report (1 point).**

Hint:
1.String functions:
refer to the link: https://cloud.google.com/bigquery/docs/reference/standard-sql/string_functions
2. How to query the table:
refer to the link:
https://cloud.google.com/bigquery/docs/quickstarts/quickstart-web-ui?refresh=1

Query #2 (1 point)
Using the "bigquery-public-data.usa_names.usa_1910_2013" table, **find out the total counts of names starting with the letters "Al".** Please note that you do not need to differentiate between the gender here. You must use "START_WITH" function. The results should be in descending order of total count. **Include your query(text) and the first 5 rows of results (either copy and paste the text or take a screenshot) in your report (1 point).**

Hint:
START_WITH function: *https://cloud.google.com/bigquery/docs/reference/standard-sql/string_functions*

Part 2: DataLab and Notebooks (3pt)

a) Introduction:

Notebooks are becoming more and more favorable every day in different areas specially data-science. As their name suggests, notebooks carry the metaphor of paper books forward. They're pretty much your old lab book from high school science, but with a Harry Potter twist. Like photographs in the Daily Prophet, the code in a notebook can be executed and results displayed as part of the page.
Notebooks can be saved as files, checked into revision control just like code, and freely shared. They run anywhere, thanks to their browser-based user interface. Though the most influential notebook, Jupyter, has its origins in the Python programming language, it now supports many other programming languages, including R, Scala, and Julia. The popular Apache Spark analytics platform has its own notebook project, Apache Zeppelin, which includes Scala, Python, and SparkSQL capabilities, as well as providing visualization tools.

Notebooks are changing the way data science teams work, thanks to the combination of the rich web browser user interface, open source, and scale-out cloud big data solutions. Not only do we now spend less time in accessing, sampling and transporting data, but we gain great features for collaboration, sharing, and explanation. Given the rapid evolution and innovation in notebooks, we've only seen the start of where this will lead—it's unthinkable that future analytic platforms won't include and extend these powerful collaborative capabilities3.

b) Starting with Google DataLabs:

Start with the Google tutorial on DataLabs (https://cloud.google.com/datalab/docs/quickstart). Here you need to install the google cloud SDK in your local computer, and set the account ID, project ID properly in your local configuration. If you did not or failed to add the cloud SDK tools to your PATH environment variable, when you follow the steps to set up and open cloud datalab, you have to use the full path for the gcloud. For example, command "gcloud components update" should be "<directory-to-gcloud-in-your-local-computer>/gcloud components update".

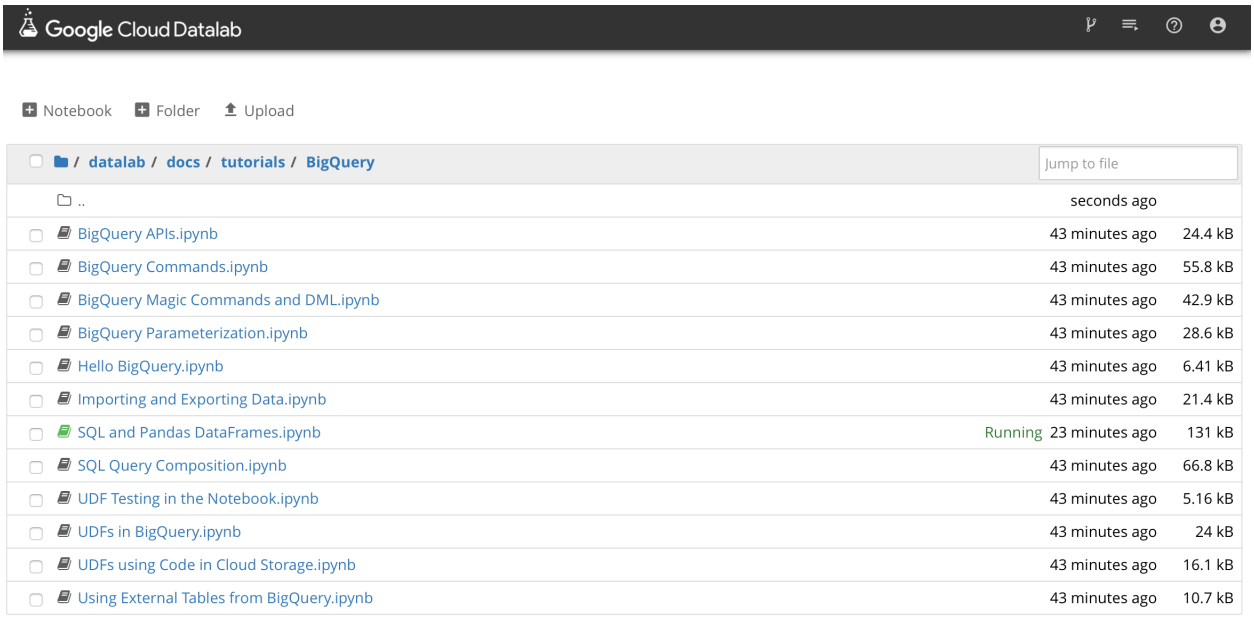
Please note that there might be some errors or warnings occurring while you follow the steps in tutorial. You just need to follow the instructions in the error/warning message to resolve the issues. Also, the instance name you create must be lowercase, and sometimes the region you choose to create the instance might do not have enough resources, you could create an instance in a different region.

Please make sure you follow the steps to clean up everything after you are done with this part to avoid charges.

Tasks:

Now that you did set up a notebook, it's time to start playing with the data.

1. Start with the BigQuery tutorial in notebook by go to
   datalab->docs->tutorials->BigQuery
   You need to run two notebooks: "BigQueryAPIs" and "BigQuery Commands.ipynb" as shown below. This step is just for you to get familiar with BigQuery in datalab.
   Simply take a screenshot of the result for the last cell you run for each notebook (For the BigQuery API notebook, take screenshot of the cell before delete resource) to prove that you have followed the tutorial**(1pt)**.



2. Create a new notebook.

   Hint: Please read the other ipynb files under the above BigQuery tutorial if you still do not know how to do the following.

a) In the first cell, write a query using the natality dataset (you could find it in BigQuery public dataset and read the table description and schema by yourself, the table is "***bigquery-public-data.samples.natality***"), to **count how many people has the birth year 1990 and birth month 7 in state 'CA'**. Simply take a screenshot of the cell with the query and result (no need to include the text of the query) **(1pt)**.

b) In the second cell, **write a query and find the number of people born on the June 30 in different years. There is no need to report the text result. You just need to visualize the data like what you did in tutorial but use a 'pie' graph.** You do not have to show all the labels for different regions in the pie graph. Please use the method you learn from the "**BigQuery Commands.ipynb**". Simply take a screenshot of the cell with the query and graph (no need to include the text of the query) **(1pt)**.

   Hint:

   1. You need use the CAST function to typecast the year type from integer to string in order to draw the pie graph (https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators)

2. Datalab commands:
   https://googledatalab.github.io/pydatalab/google.datalab%20Commands.html

Part 3: Big Public Data, Visualization and Interpretation (2pt)

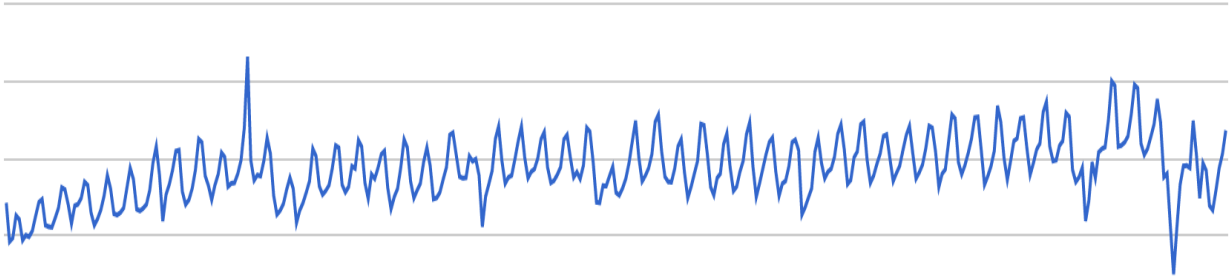
a) Introduction:

In this part, we use chicago_taxi_trips data("***bigquery-public-data.chicago_taxi_trips.taxi_trips***"). Similar to **Part 1**, you could search for the table and read the schema for the table. The size of the data is about 65GB. We want to have some understanding about this data using basic visualizations.

b) Visualization:

1).Use the method you learned from "**BigQuery Commands.ipynb".** Write a query to retrieve the sum of trips for each single day for the year 2013, 2014, and 2015 and visualize in three figures. In case of a better visualization for each year, include the first 10 days of the next year as well. This means you will visualize (2013-01-01 till 2014-01-10, include both dates) for year 2013.

Can you find a general semi-periodical pattern in the data like the figure below?

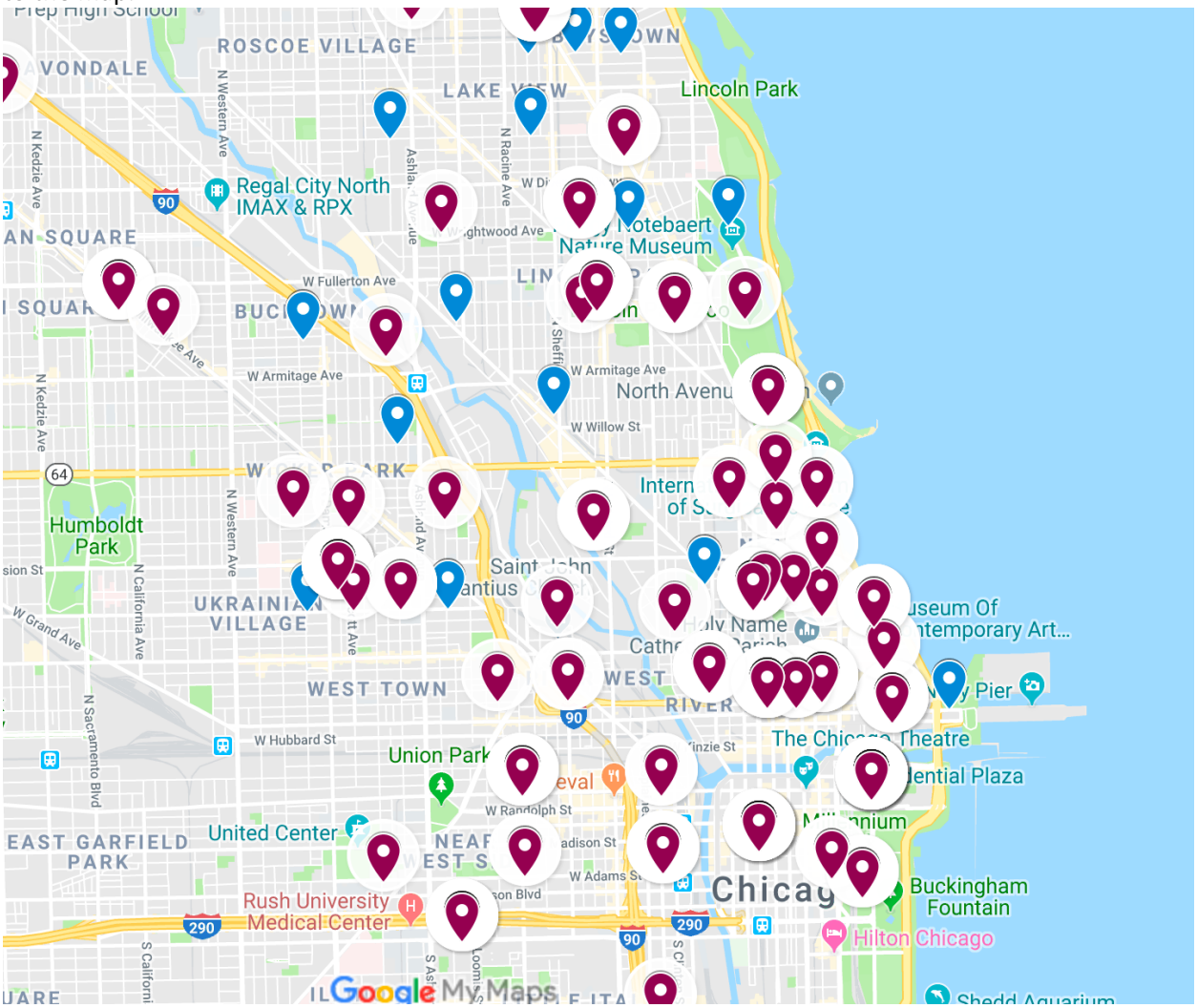


There are two unusual patterns (anomaly) being repeated in all three figures. One big decrease in numbers happens in the first few weeks(Hint: Long weekend, I have a dream https://www.youtube.com/watch?v=3vDWWy4CMhE). The other one happens at the end/beginning of each year Report your figures and an explanation for these two anomalies.

You need take three screenshots for the query and figure for the year 2013, 2014 and 2015 respectively. And answer the question above. (1pt)

Hint: Use the FORMAT_TIMESTAMP(format_string, timestamp[, time_zone]) function to get the day of the timestamp in string format.
https://cloud.google.com/bigquery/docs/reference/standard-sql/timestamp_functions

2) We want to investigate the pick-up and drop-off locations for expensive rides (trip_total between $300 and $400, include both) for future planning. You can use 2013 data (2013-01-01 till 2013-12-31). Feed in the coordinates (longitude, latitude) data into Google my maps (https://www.google.com/maps/d/). Differentiate between drop-off and pick-up locations using different marker colors. The final figure should be something like the following (the figure is just an example, you need take your own screenshot). Please note that the longitudes and latitudes values might be null, you just need to retrieve all the not null values and download as csv files, and the import to the map.



You need to include the query (text format) for pickup/dropoff latitudes and longitudes and the screenshot of the marked map (there is no need to include all the locations, you could zoom in the map and take a screenshot that includes most locations) (1pt)

Submission Guidelines:

The total points for this homework is 7. The submission **MUST** *be a pdf file named [Student First Name]_[Student Last Name]_HW4.pdf that includes all the required parts above marked in red color.*

If you have any general questions about the homework, please post your questions on HW4 discussion on USC DEN course forum. Before asking, you should have a look to see whether similar questions were asked and answered. Thank you!

Students can submit the assignment to USC DEN. Just go to the course → MY TOOLS → Assignments → Homework 4. The deadline is firm. ***You will not be able to submit your homework after the deadline****. The deadline is firm, only submissions that make it to the system will be graded. It is irrelevant if you submit your work at 11:59 PM according to your clock, the system will stop accepting submissions at 11:59 PM according to the clock on DEN (Dropbox) server. You will not be able to submit your homework after the deadline. Also, please expect the heavy network traffic around the deadline and network delay won't be treated as a valid reason for late submission. The system accepts multiple submissions and only the most recent submission will be graded. Therefore, we advise you to make the initial submission at least a day before the deadline, and overwrite it with a better version or more complete submission after you have it. Make up data and have fun with this!*