Homework 4
CSCI 585 – Database Systems

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

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

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 BigQuery Docs" and then "Quickstarts". In order to make all the steps integrated, we will use web UI, select "Quickstart Using the Web UI" or simply click here.

Follow with the tutorial step by step (but do not clean up!).

Try it yourself:

Query #1 (1 point)
Using the names_2014 table, query names and their count which have the same second letter as your own name AND are from the same gender group as you. For example, for Mr. Tommy Trojan, the result would be 1319 names, such as: Noah, Logan, Joseph, etc. The results should be in descending order of count. Print your query in your report. Save the result of your query in JSON format and name it "HW4_Q1.json".

Query #2 (1 point)
Using the names_2014 table, find out the sum of count of names starting with your name. For example, Alex is the starting part of 57 names such as Alexx, Alexander, Alexxa, etc. The result for Alex would be 36176. Print your query and the result (one single integer value) in your report.

Part 2: DataLab and Notebooks

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 capabilities[3].

Starting with Google DataLabs:

Start with the Google tutorial on DataLabs here. Like previous parts, we will use Cloud shell, there is no installation required.

If you have close your notebook, you can list your Notebooks and run one of them following these commands:

```
tommyTrojan@cs585hw4:~$ datalab list
tommyTrojan@cs585hw4:~$ datalab connect hw4instance
```

Visualizing Google BigQuery data in Google DataLabs:

Now that you did set up a notebook, its time to start playing with the data. Start with the Google instructions here. Follow the instructions step by step. Create required cells and at the end save the notebook as HW4_dataLab1. You need to submit this file (HW4_dataLab1.ipynb).

Try it yourself:

Create a new notebook and name it HW4_datalab2. Make the first cell a Markdown with these characteristics:

Context	Format
CS585 - Databases Systems	Alt-Header 1
Homework 4	Alt-Header 2
Name	Header 3
USC email	Header 3
A brief description	At least make one word **bold** and one other word *italic*.

You DO NOT need to know any detail about Markdown, just take a look into the first few lines of the cheat-sheet given in this link.

In the 2nd cell, write a query using natality dataset and find the weekday of your own birthdate. Report the query.

[3] Source: https://svds.com/why-notebooks-are-super-charging-data-science/

In another cell (3rd cell), write a query and find the number of people born on the same day as you in different years, store the data in a dataframe. Visualize the data just like what you did in tutorial but use a 'pie' graph[4]. In addition to the query, take a snapshot and add it to your report.

What you need to submit for part 2 (each 1 point):

- HW4_dataLab1.ipynb
- HW4_dataLab2.ipynb
- Query in your 2nd cell
- Query in your 3rd cell
- Snapshot of your visualization

Part 3: Big Public Data, Visualization and Interpretation

Introduction:

In this part, we want to use a public data. This dataset includes trip records from all trips completed in yellow taxis in NYC since 2009. The size of the data is about 130GBs (Big data?!). We want to have some understanding about this data using basic visualizations. You may want to take a look into the documentation of this data set in its providers website.

We will start with BigQuery (you may read this link if you need further help to start working with dataset in BigQuery - but information given in part 1 of this tutorial is enough).

BigQuery & DataLab – not quite the same:

Write a query that retrieves the sum of number of passengers for each single day before 2015. Sort the data by the date. It should be noted that we consider pickup time as the main timestamp for a trip (Hint to validate your answer: the total number of passengers for the first date of dataset (2009-01-01) is 602881 - Wow!

Now create a new datalab and name it HW4_nyc_taxi. Use exactly the same query format bq.Query('YourQuery') introduced in previous part. Run your cell. Does it work?! Create a new Markdown cell and report in your error, explain why this query worked totally fine in Google BigQuery but not in DataLab (Hint: Someone almost had almost the same issue in this post).

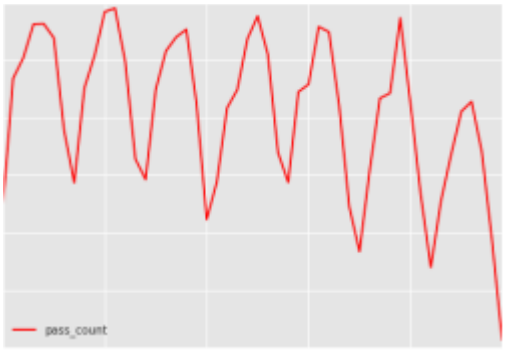
You need to correct your query. Checkout Google documents about Migrating to Standard SQL, click here. Store the results in panda dataframe. Report your codes (query, pandas).

[4] Not necessary but you may want to take a look into this link: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.plot.html

Visualization:

Start with a new cell. Visualize the total count of passengers for 2009, 2010, and 2014 in different 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 (2009-01-01 till 2010-01-10) for year 2009.

Can you find a general semi-periodical pattern in the data like the figure below? Explain the pattern. Without writing it, suggest a query that proves your hypothesis and report it (1 point).

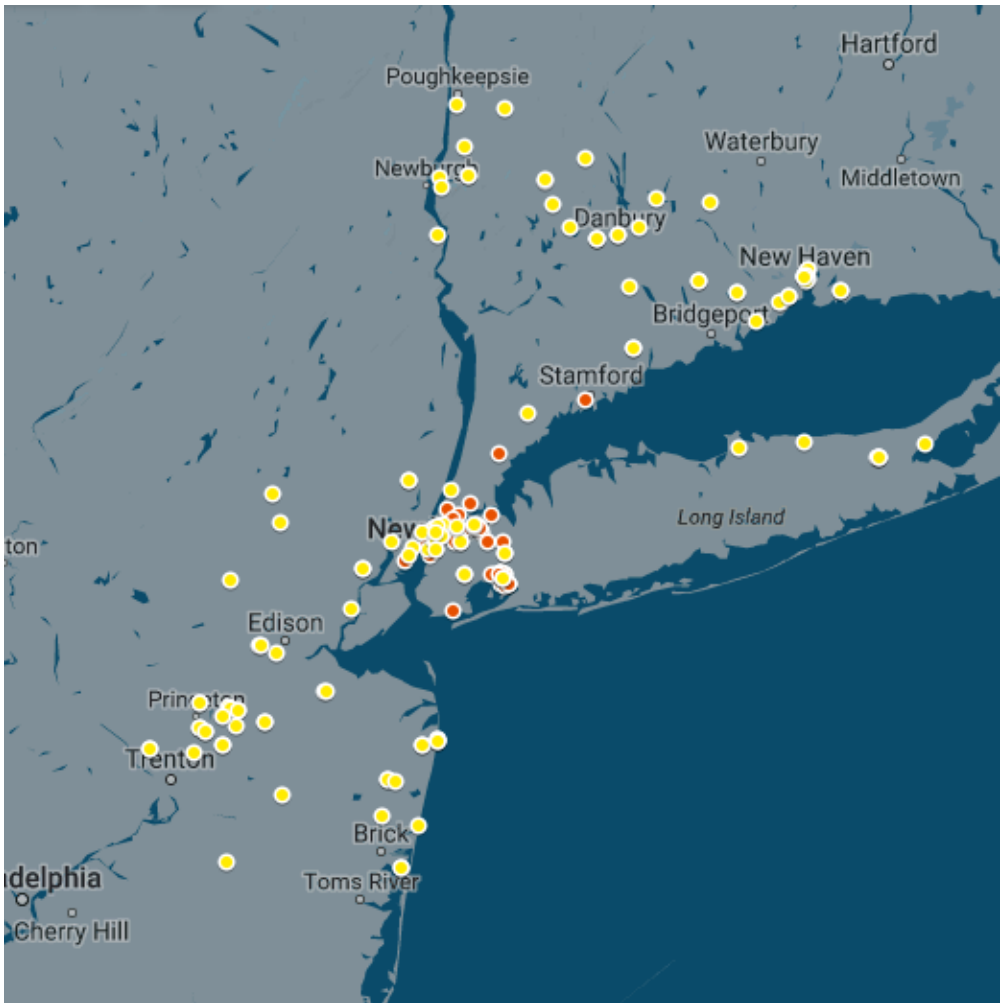


There are two unusual patterns (anomaly) being repeated in all three figures. One big decrease in numbers happens in the first few week (Hint: long weekend – I have a dream). The other one happens at the end/beginning of each year (Hint). Report your figures and an explanation for these two anomalies (1 point).

Visualize the complete data for year 2011, 2012, and 2013. Find the minimum point (you can query this or just find it manually). Simply search the date and find out what caused this. For the first two years you may find natural disasters. However, for 2013, the decrease lasted for a few days, you will find meaningful information here on how new regularizations affected the business (1 point).

Bonus Part – 1 extra point:

We want to investigate the pick-up and drop-off locations for expensive rides (between $300 and $400) at night (after 6PM) for future planning. You can use 2013 data. Feed in the coordinates (longitude, latitude) data into Google my maps. Differentiate between drop-off and pick-up locations using different marker colors. The final figure should be something like this:



How to turn in?

- Report your work including required queries, snapshots and explanations in a pdf file named [Student First Name]_[Student Last Name]_HW4.pdf.
- Create a zipped folder including your pdf report, Json files and your notebooks and name it [Student First Name]_[Student Last Name]_HW4.zip, upload this in DEN website.
- Don't forget to shut down your cloud systems to avoid further charging.
- The deadline is Tuesday, June 27, 2017 11:59 PM. No submissions will be accepted past the deadline.