1. **Business Problem:**

I have been approached by New York State and some data analysts to build a data warehouse that combines and stores data from the New York State Department of Buildings (DOB). The requirements involve obtaining data via NYC Open Data, the data concerns NYC area. The data gets updated daily on the state website, so it needs to be pulled every 24 hours and then added to a data warehouse that can be accessed by different users with attributed credentials and roles.

1. **Business Impact:**

Having a warehouse set up will ease the process of data mining, as well as a concrete form of reporting. It can show different pictures of the approval process, number of permits per type of jobs, or per borough and other data.

The costs of this data warehouse will be initially low, as I will use free trials available, later we will discuss the budget depending on the client’s requests and needs. Snowflake offers a beginner plan for an extra-small sized warehouse that is a good fit for this project, so I will use Snowflake - the computing-based data cloud, to store and analyze the data. Eventually, as the DOB will add more in-state data, costs could and will increase.

1. **Business Persona:**

The data warehouse will be primarily used by government agencies and data analysts/scientists. It will allow the users to make faster, educated decisions on how the approval/ application process for permits goes in NYC, the distribution of job types throughout the 5 boroughs of NYC, the speed of executed work etc.

1. **Data:**

Data comes from the NYS Open Data Source.

Through the link below it can be accessed the initial data set:

<https://data.cityofnewyork.us/Housing-Development/DOB-NOW-Build-Approved-Permits/rbx6-tga4>

Size: ROWS: 450K, COLUMNS: 35.

*I chose to work, later on, with a smaller set, over 1000+ rows and fewer columns, since the warehouse didn’t allow me to insert such a big volume. The data set will be provided in the scripts.*

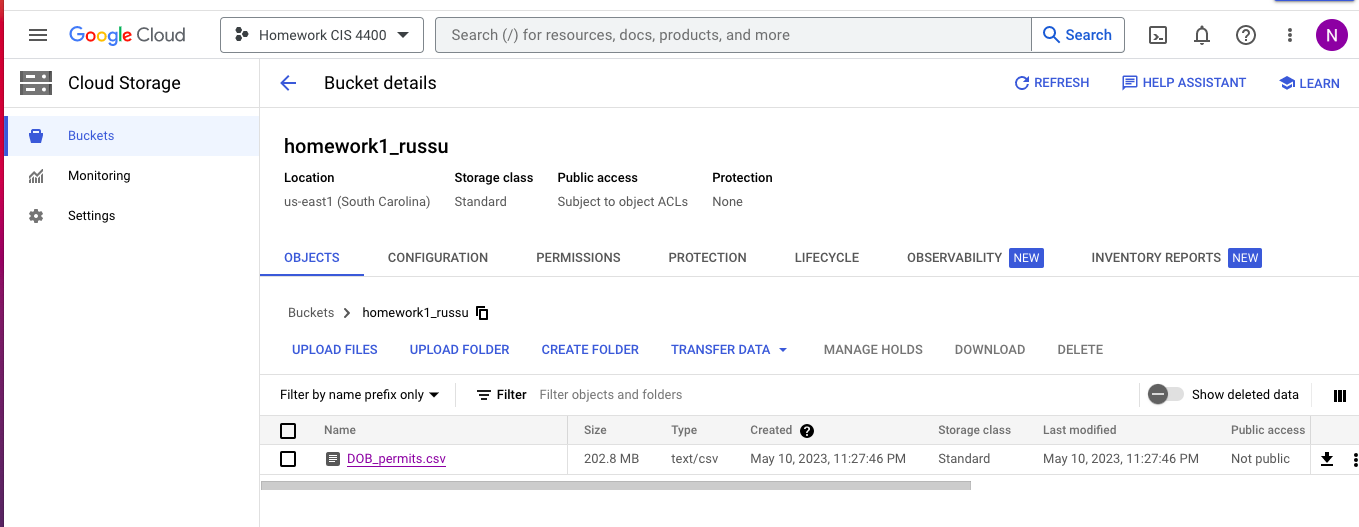
1. **Methods:**

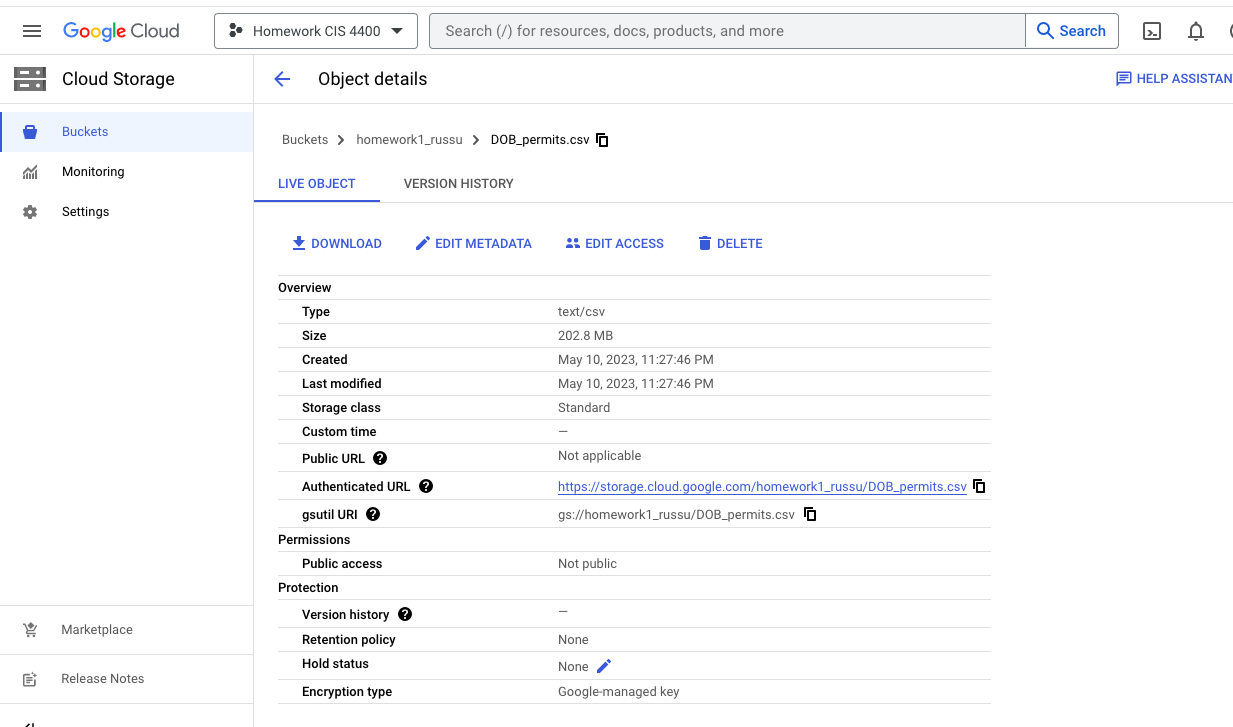
I wrote a Python script to extract/analyze data from the above CSV, with an API, and also did EDA. I created a bucket to upload the data set in the Google Cloud storage.

The bucket (this contains the data set):

<https://storage.cloud.google.com/homework1_russu/DOB_permits.csv>

Screen shots:





*Data Dictionary* (I downloaded from the state’s website, see the 2nd sheet of the file):

<https://github.com/NataliaRussuu/CIS-4400-Homework-1-2/blame/009a7210eb5a016cf768dbb8fd7ab721351120f3/DD_DOB_Permit_Issuance_2019-07-29.xlsx>

*Dimensional Modeling*: see below the model for the project, with permit item being the granularity of the dimensional model. DIM modeling, created with Lucid Chart, having one fact table that used permit

<https://lucid.app/documents/view/afe9cb41-ba63-499a-b706-e346a4838e61>

A snapshot of dim. model:

Diagram

Description automatically generated

1. **Data Tools:**

Data Storage:

To store: google cloud storage to hold the raw csv file of the project:

<https://storage.cloud.google.com/homework1_russu/DOB_permits.csv>

Data Processing: used Python.

Initially I used Fivetran as my ELT platform for extracting the data from the storage every 24 hours, so I connected to Google Cloud and tried loading it directly into my data warehouse for transformation later. Then I used SQL code in Snowflake and imported the tables into the warehouse.

Graphical user interface, application, Teams

Description automatically generated

Data Orchestration: I Used Snowflake, Fivetran, Tableau.

1. **Interface:**

Data can be viewed in Snowflake, I created visuals with Tableau. Snowflake is my data warehouse that also allows for dashboards to be built directly in the UI.

HERE is the snowflake access of the tables, there can be also viewed the codes in :

<https://app.snowflake.com/us-east4.gcp/he04105/#/data/databases/DOB_NYC_PERMITS/schemas/PUBLIC/table/NYC_PERMITS/columns>

Quick preview of the database:

Graphical user interface, application

Description automatically generated

For visuals I used Python, matplotlib and Tableau. Chart, pie chart

Description automatically generated

NUMBER of permits, 2016-2023:

A picture containing chart

Description automatically generated

Links to Tableau Public for more visuals:

<https://public.tableau.com/app/profile/natalia4559/viz/AtimelineofNYCissuedpermitsinNYC/Dashboard4>

Chart, line chart

Description automatically generated

<https://public.tableau.com/app/profile/natalia4559/viz/BarchartcomparisonofDOBpermitsissuedvspermitsexpired/Dashboard2>

Chart, bar chart

Description automatically generated

<https://public.tableau.com/app/profile/natalia4559/viz/TreemapofissuedpermitsinNYC2016-2023/Dashboard3>

Chart, treemap chart

Description automatically generated

Most permits were issued in 2022, General Construction leads, as opposed to the demolition where there were only 20 permits issued since 2016.