# Virginia Tech

## Bradley Department of Electrical and Computer Engineering

## ECE 5984 Data Engineering Project
## Fall 2023

## Assignment 4
## Visualization and the Business Intelligence Pipeline

Please note the following:

- Solutions must be clear and presented in the order assigned. Solutions must show the work needed, as appropriate, to derive your answers.
- Data being processed in both the lab and homework should be in the correct format and location and any other deliverables should also be completed as mentioned in the modules.
- For the Homework, the submission process requires that all code files should be uploaded to Canvas before the deadline.
- Submit your Homework using the respective area of the class website by 11:55 p.m. on the due date.
- When a PDF file must be submitted, include at the top of the first page: your name (as recorded by the university), email address, and the assignment name (e.g., "**ECE 5984, Homework/Project 1**"). Submit a single file unless an additional file is explicitly requested.

# Lab 4

## Push data stored in S3 bucket (data warehouse) to a MySQL database on AWS RDS and access it through Tableau to perform Data visualization

In this lab we will be pushing our stored data in our S3 bucket (Data Warehouse) to a MySQL database hosted on AWS RDS. Afterwards we will be using Tableau to access the data that we pushed to make meaningful data visualizations.

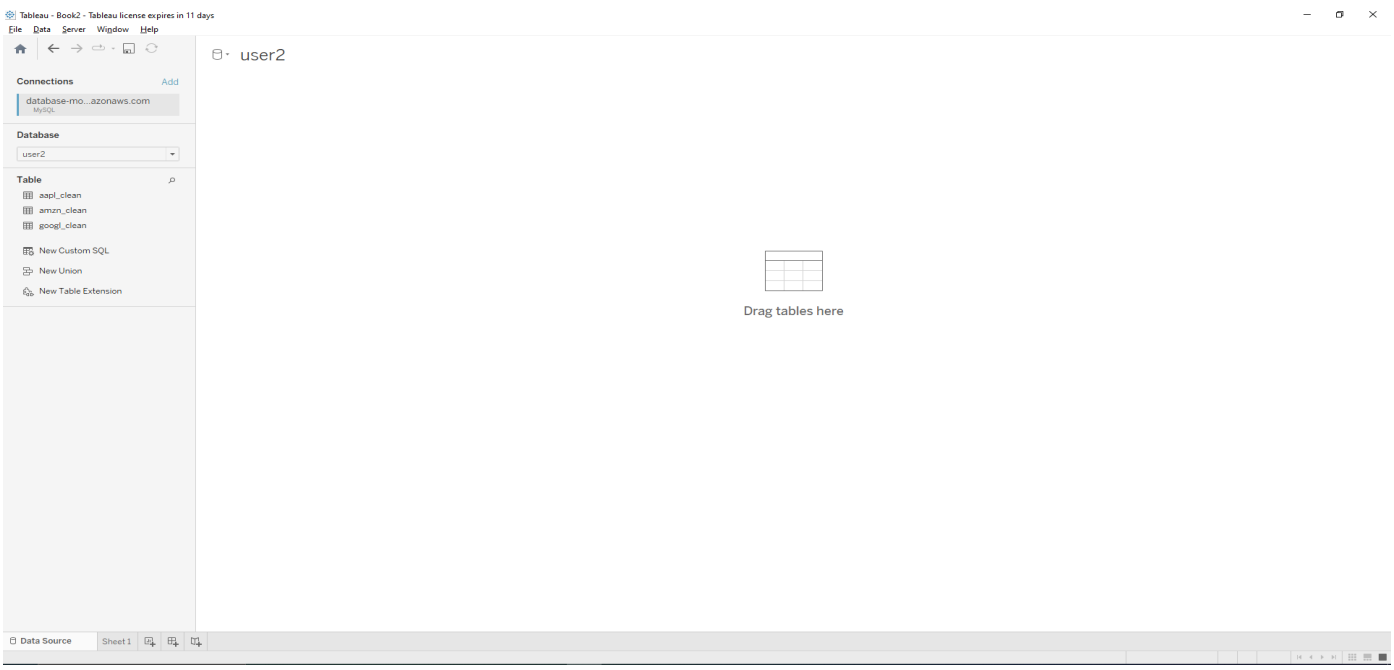
### Step to perform Lab 4

1. Download both dag_db.py and load_db.py on your local machine. Open the load_db.py file and insert your pid as the variable name for db. This is done in order to make a database with the name of your pid and going forward this is the database that should be used to push all your tables onto.
2. Next fill in the S3 bucket location where you have the clean_aapl.pkl, clean_amzn.pkl and clean_googl.pkl from lab 2.2 saved. Finish by saving both files.
3. Spin up a docker container as shown in Step 2.2: Spin up and exit out of a docker container (lab 0)
4. Install the needed packages for the dataset running the following commands
    a. `pip install pymysql`
5. Navigate to the airflow/dags folder by typing the following command:
    a. `cd airflow/dags`
6. Remove any other .py files already there using the following command (Note: make sure you have a backup of your previous lab code and homework code on your local machine)
    a. `rm file1 file2` (Any number of files can be added with a space in between)
7. Create a new dag_db.py and copy all the code from your local machine dag_db.py onto it. To do so enter the command:
    a. `sudo nano dag_db.py`
8. This will open a command line based text editor. Copy the contents of your local dag_db.py file to the newly created one inside your container (Tip: use ctrl+shift+V to paste content directly if using the web browser)
9. Save by pressing ctrl+x , then press y to confirm changes and then hit Enter
10. Similarly create the load_db.py file and copy code from local machine load_db.py onto it using the same steps
11. Access your airflow GUI (Step 2.4: Launching and access the airflow GUI (lab 0))
12. You should see the same **dag_db**. Click it and go to graph. On the right side click the play button and press trigger DAG
13. This triggers the dag we just uploaded which intern runs the load_data() function from load_db.py.
14. After the DAG finishes running you should have created a new MySQL database by your name and stored the data inside the clean_aapl.pkl, clean_amzn.pkl and clean_googl.pkl files onto 3 separate tables namely aapl_clean, amazn_clean and googl_clean respectively.
15. Next we need to launch Tableau and access the data in the tables we just created

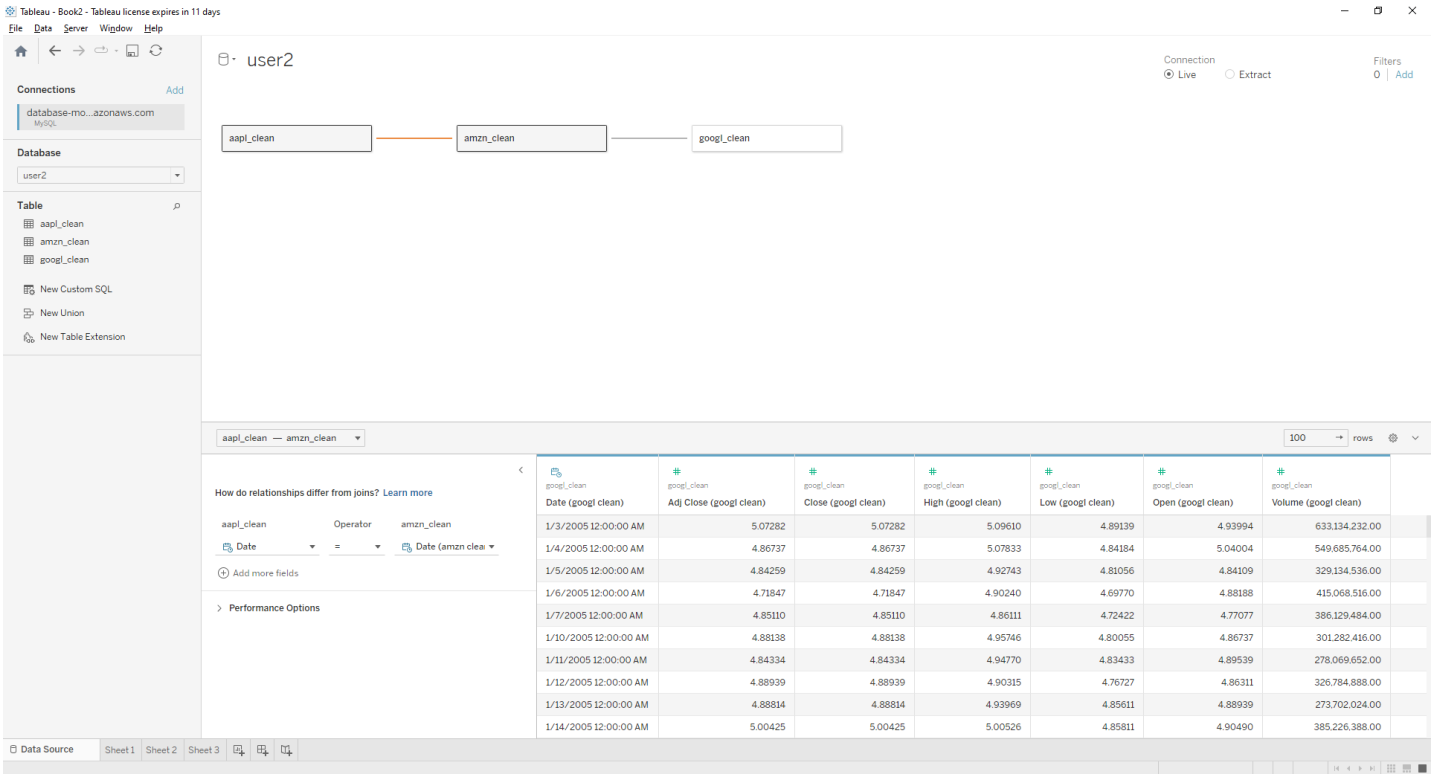
## Using Tableau

16. Download tableau desktop using the following link and activate your free trial:
    https://www.tableau.com/products/desktop/download
    OR
    Sign up as a student and get Tableau Desktop for free using the following link:
    https://www.tableau.com/blog/tableau-students-free-access-tableau-desktop
17. Install Tableau on your local machine
18. Launch Tableau and click more under Connect -> To a Server
19. Click MySQL
20. Fill in the following details
    i. Server: database-eng2.cwgvgleixj0c.us-east-1.rds.amazonaws.com/
    ii. Port: 3306
    iii. Database: pid (same one used in load_db.py)
    iv. Username: admin
    v. Password: wY59A93oZFz2Vff7sjpd

21. If your airflow code worked correctly you should see 3 tables, each having the stock data of each company namely Apple, Amazon, Google




22. If you hover over any of the tables you can click ‘view data’ and you should see the data in the tables
23. Now drag the 3 tables towards the middle of the screen. After each table is dragged you will be prompted to make relationships between the tables. For our example you can choose the Date column for each company table data to be equal. The end result should look something like this:




24. Now navigate to sheet 1 towards the bottom. Here you should be able to see all the columns of each of the tables available on the left-hand pane.
25. Start by dragging the 'Date' metric from the left-hand pane to the columns field towards the top. This sets the Date as your x-axis
26. Drag the 'close' metric to the rows field. This sets your close values as the y-axis and you should have a graph showing the 'close' value of the AAPL stock over time
27. Now drag the 'close' metric to the rows field for the Amazon and Google stock. Now you should have 3 graphs each showing the close value of a particular company over time

After confirming your data arrived is accessible using Tableau, do not forget to close airflow by pressing ctrl-C on the command line where airflow is running on the EC2 instance and also remember to exit out of the container you have been working from using the command `exit` and double check if the container actually stopped by using the command `docker ps`. To manually stop the container if it had not done so automatically run the command `docker stop <pid>`.

## Deliverables

1. The dag_db.py file and the load_db.py for the lab should be uploaded to Canvas.
2. Submit a pdf with the screenshot of the final 3 graphs of the lab