FA22: MGMT ACCESS USE BIG DATA: 11107

Final Project

Julia Nguyen

Introduction

My final assignment will demonstrate the creation of virtual machines, both with J2 and Google Cloud Platform; and the implementation of a data pipeline including – loading dataset, transforming, summarizing and visualizing information.

I will create new VMs from beginning, set configuration to fit my purpose of project here, implement data pipeline and finally, shelve the instance. In the VM on J2, Jupyter Notebook will be used in the VM to analyze data about social media influencers on Instagram and YouTube[1]. I will also discuss my finding and learning after demonstrating the project results.

Background

This is the first time I've learned about virtualization and cloud computing, therefore I found myself particularly interested in working with virtual machines. I chose 2 platform – J2 and Google Cloud Platform as they were introduced in this course. Also, doing that will provide me with a better understanding of the similarities and differences between these two.

I selected to work on data about social media influencers because my job in the past was heavily involved in assessing data to inform marketing strategies. Comparing to traditional marketing strategies such as TV/radios commercials, newspapers, printed materials etc., social media and influencers have proven to be an effective trend to reach consumers in this digital era. In this paper, the business problem I have at hand is to determine which platforms can penetrate better in which geography areas, in order to maximize the benefits of running influencer ads on said platforms, since the business wish to expand their revenues in both domestic and international markets. Therefore, I'm interested in learning how Instagram and YouTube – the two major platforms that the business has the most content investment in – have performed in some large countries from March 2022 through November 2022.

[1] Data source: https://www.kaggle.com/datasets/ramjasmaurya/top-1000-social-media-channels

Methodology

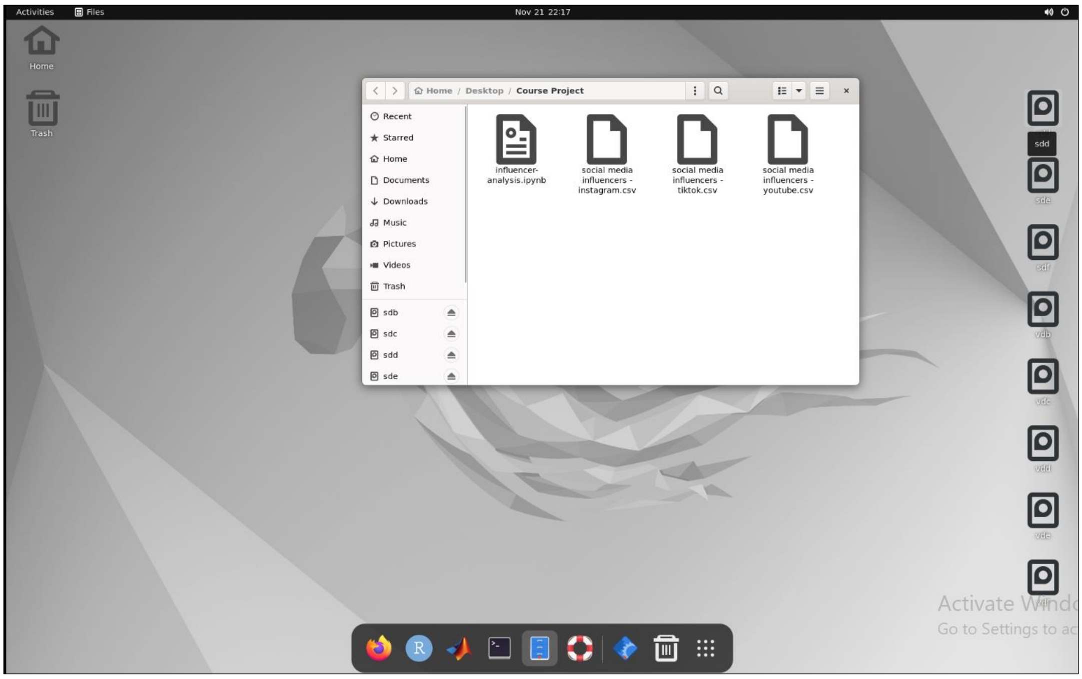
Virtual Machine Instance with J2

I launched a new virtual machine instance on a cloud environment J2 by following these steps:

1. Access Exosphere interface, which is a user-friendly interface that can be used with most cloud infrastructure including J2.
2. Access allocation CIS220079 that was provided by IU during the course.
3. Click 'create' >> Instance to create a new instance for this final project.
4. Select 'By type' tab and then the newest Ubuntu version, which should contain the latest official image for the operating system.
5. I configure this instance as follows:
    - Name: 'final_project'
    - Flavor: m3.small
    - Root Disk Size: 250
    - Quantity: 1
    - Enable Web Desktop: Yes

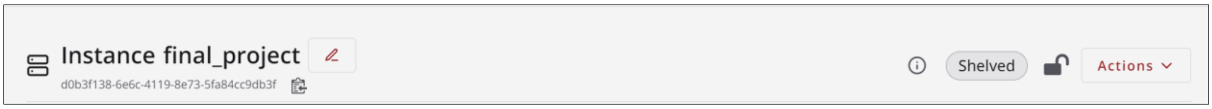
Data transfer and analysis in J2

1. Launch the VM web enviroment by navigating to this newly created instance >> 'Interactions' >> 'Web Desktop'.
2. Open a terminal and type the following commands
    - sudo snap install jupyter *# to install jupyter notebook*
    - jupyter notebook *# to launch jupyter notebook*
3. Create a folder named 'Course Project' on VM desktop and transfer the 3 datasets from local computer to this VM.
4. Create a Python 3 notebook in this folder and perform the data analysis in Jupyter notebook (see attached file influencer-analysis.py)


Activities
Files
Nov 21 22:17
Home
Trash
Home / Desktop / Course Project
Recent
Starred
Home
Documents
Downloads
Music
Pictures
Videos
Trash
sdb
sdc
sdd
sde
influencer-analysis.ipynb
social media influencers - instagram.csv
social media influencers - tiktok.csv
social media influencers - youtube.csv
sdd
Activate Windo
Go to Settings to ac

Screen capture of the VM

5. Finally, go back to the instance in J2 >> click 'Actions' >> 'Shelve' to shut down this instance and offload it from my computer.

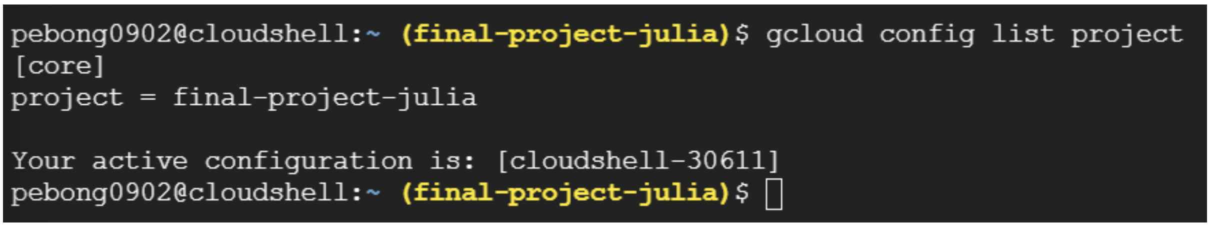


Set up Google Cloud Platform

1. To start, I activate cloud shell environment by clicking Activate Cloud Shell button at the top right corner of the Google Cloud console.
2. I named my project 'final-project-julia' by typing in the following command:

   gcloud config set project final-project-julia

   and then click 'Authorize' to proceed with the project name change.



```
pebong0902@cloudshell:~ (final-project-julia)$ gcloud config list project
[core]
project = final-project-julia

Your active configuration is: [cloudshell-30611]
pebong0902@cloudshell:~ (final-project-julia)$ 
```

****Note to instructors:*** While I registered to receive Google credits coupon with my IU email address, I did not sign out of my personal Google account when I redeemed the coupon and therefore please note that my personal account is displayed here.

Create Virtual Machine Instance

3. Click the 'Navigation menu' icon on the top left corner of the console >> Compute Engine >> VM Instances.

4. It took approximately 2 minutes for initialization.

5. Click 'Create Instance' at the bottom of the page.

6. I configured the new instance with the following parameters:

   - **Name:** project-vm-11212022
   - **Region:** us-west4 (Las Vegas) for ease of keeping track as this is where I'm located
   - **Zone:** us-west4-a
   - **Series:** E2
   - **Machine type:** e2-medium (2 vCPU, 4GB memory), which provides 2-CPU and 4GB RAM. I selected this option to preserve resource as I believe this is sufficient for the scope of this project.
   - **Enable display device** is checked
   - **Boot disk:** New 10 GB balanced persistent disk – Operating System Image: Debian GNU/Linux 11 (bullseye)
   - **Access:** Allow full access to all Cloud APIs
   - **Firewall:** Allow HTTP traffic to allow access a web server

7. Click 'SSH' to connect to the virtual machine

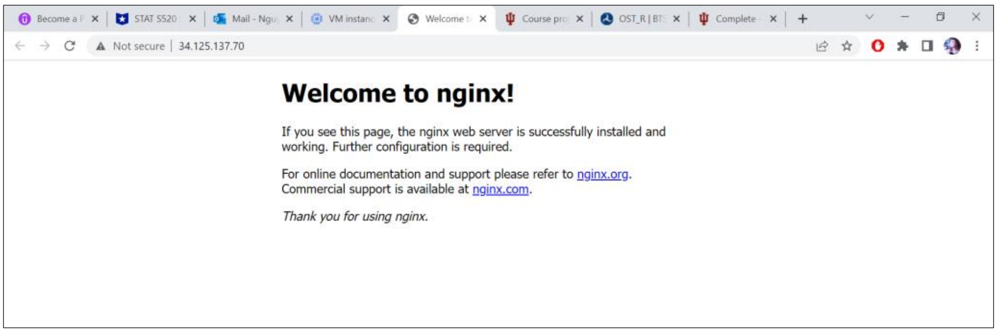
8. Install NGINX web server to which I could connect my VM later on by typing in this command:

   ```
   sudo apt-get install –y nginx
   ```

9. Confirm that the NGINX is active and running properly:

   ```
   Ps auwx | grep nginx
   ```

10. Click Exteral IP link in the row of my machine and the web page is displayed as below:



Welcome to nginx!

If you see this page, the nginx web server is successfully installed and working. Further configuration is required.

For online documentation and support please refer to nginx.org.
Commercial support is available at nginx.com.

Thank you for using nginx.

Results

There are 2 VMs that have been created and are fully functioning, one via J2 and one via Google Cloud Platform. In J2, the instance is named 'final-project' and in GCP, the instance is named 'project-vm-11212022'. The J2 VM has been shelved properly.

Social media influencers dataset was first downloaded from Kaggle and was available to access in J2 VM. Jupyter Notebook was installed and used to perform the exploratory data analysis. Insights into Instagram and YouTube influencers' performance by countries are visualized in bar charts.

Discussion

There are not many significant differences between the two clouding infrastructure in term of creating new VM. The total amount of time for an instance to be built in J2 is 3 minutes. And it took less than a minute for a new instance to be created in GCP. However, it was quite slow when I tried to deploy Python Jupyter Notebook in J2 VM. I've learned that the processing speed could have been improved by initial configuration during setup.

I had some difficulties when trying to download Python packages in virtual environment in order to perform data analysis with Python Jupyter. Lack of knowledge in computer-related background and familiarities with some commands also hindered my work. Yet I learned to overcome this hurdle by doing more researches online, thanks to the Internet! Another learning point that I wish I could have done differently is the inability to source real-time data. I was struggle identifying some real-time data and ingest it through GCP, similar to what was taught in the Qwiklab but I've failed to do so. Instead, I downloaded the dataset locally to the VM.

The VM speed also had an impact on the model I chose for my data analysis. While the data analysis I did in Python was simple, I believe the main focus of this project is to demonstrate my understanding of how big data works in relation to the whole scheme of data science, and not necessarily the coding. This assignment has provided me with a deeper dive into data implementation.

Conclusion

Working on Jetstream Virutal Machines feel just like working on my own home computer, only with cloud-based and more computing powerful. GCP is also a great great advance as this platform integrates many other functions and features such as multi-format databases, storages, data analytic tools, etc… Data Pipeline is extremely important as the 'volume', 'variety' and 'velocity' of data has grown so much throughout the years, as known as 'big data'. Data Pipeline allows different format of big data, such as structure or un/semi-structure, to be processed on a large scale.

References

Data source: https://www.kaggle.com/datasets/ramjasmaurya/top-1000-social-media-channels

Reading and lecture materials provided throughout the course FA22-BL-INFO-I535-11107, as well as Qwiklabs in some modules.