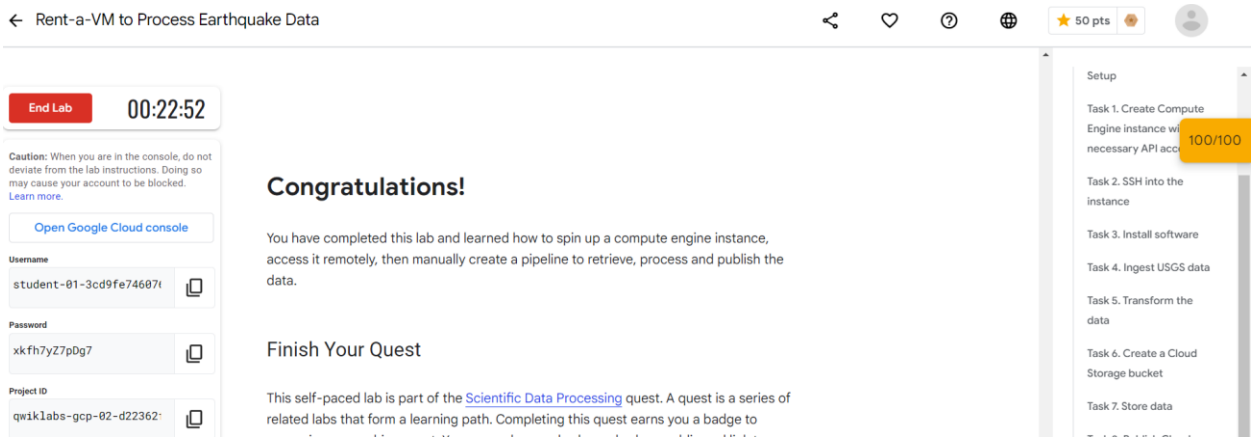
Fall 2024- I535
Mgmt. Access Use Big Data

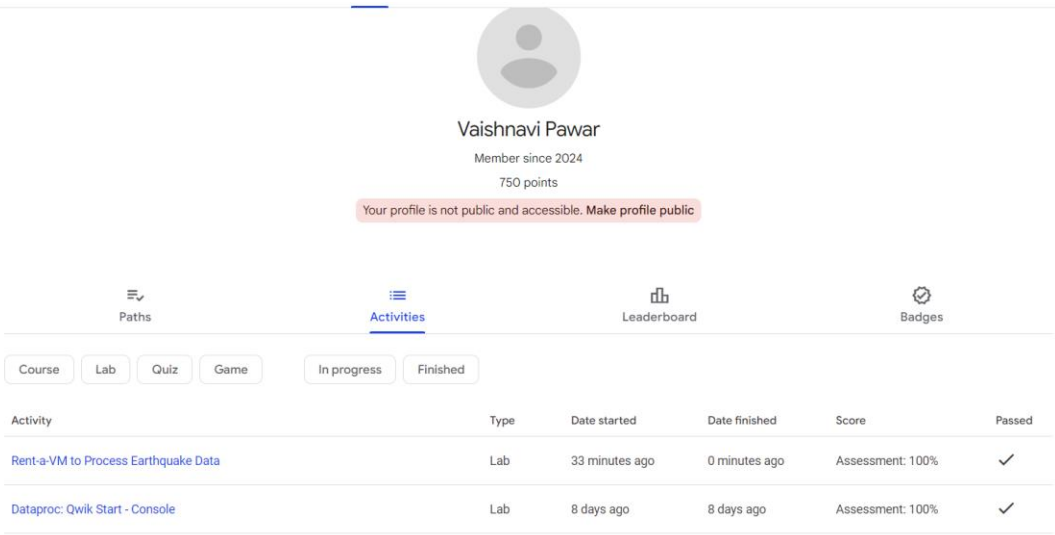
Vaishnavi Vishwas Pawar
vpawar@iu.edu

Lifecycle and pipelines Assignment

As part of my assignment, I completed a hands-on lab to gain real-world experience with Google Cloud services. The "Rent-a-VM to Process Earthquake Data" Qwiklab is the focus of this report, as it demonstrates how to set up a virtual machine to process real-time earthquake data from the USGS. These labs demonstrate how cloud computing can simplify the management of large datasets, allowing researchers to focus on data analysis and visualization. This report describes the data processing pipeline used, examines data lifecycle management, and makes recommendations for improvements.

Rent-a-VM to Process Earthquake Data





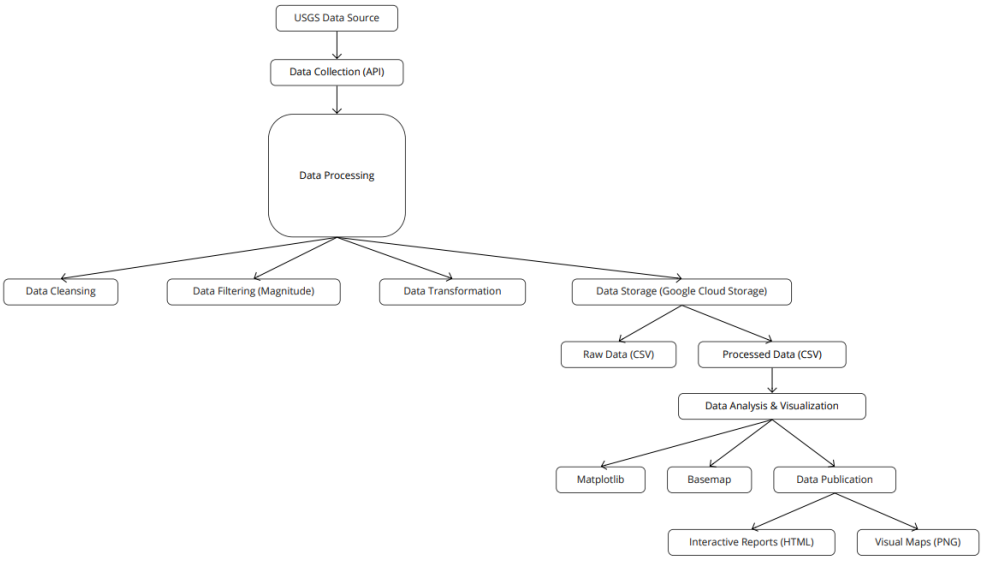
High level architecture of this pipeline

Data Sources and Types: The primary data source for this project is the United States Geological Survey (USGS), which provides real-time earthquake data from around the world. This data is accessed through a public API, which returns information in CSV format. The CSV files contain detailed records of each earthquake, such as timestamps, geographic coordinates (latitude, longitude), depth, magnitude, and other pertinent seismic parameters.

Data Storage Technologies: All data within the earthquake processing pipeline is managed using Google Cloud Storage (GCS), renowned for its scalability and security. Raw and processed earthquake data are stored as CSV files, while final outputs include CSV for records, HTML for interactive reports, and PNG for visualizations, ensuring both consistency and accessibility.

Data Transformation Tools and Techniques: Python is used to transform earthquake data, with libraries such as Requests fetching data from the USGS API and CSV for data handling. NumPy and Pandas are used to perform advanced processing tasks such as filtering and statistical analysis. Matplotlib and the Basemap toolkit handle visualization, plotting earthquake events on a global map and indicating magnitude and depth with color and size variations.

Other Data Components: The visualization process produces PNG images that summarize earthquake activity, which aids in public understanding and scientific analysis. Additionally, the pipeline generates HTML reports, which improves data accessibility and interactivity. These reports include interactive elements such as maps and graphs, making the insights from the earthquake data more engaging and accessible via a web browser.



Flow chart of the Pipeline

Aspects of a data lifecycle are implemented in this pipeline

The "Rent-a-VM to Process Earthquake Data" pipeline effectively implements several aspects of the data lifecycle, ensuring comprehensive data management and utilization:

1. Data Collection: The lifecycle begins with the acquisition of real-time earthquake data from the US Geological Survey (USGS) via an API. This ensures a consistent and up-to-date stream of seismic data entering the pipeline.
2. Data Processing: Once collected, the data is processed, which includes cleansing, filtering by parameters such as magnitude, and transformation. This stage is critical for converting raw data into a structured format that can be analyzed and visualized. Python libraries like Pandas and NumPy are used to manipulate data efficiently.
3. Data Storage: After processing, the data is saved to Google Cloud Storage (GCS). This stage employs scalable and secure storage solutions to manage both raw and processed data. The raw data is stored in CSV format, making it easy to access and process as needed.
4. Data Analysis and Visualization: Visual representations are created by analyzing the data using tools such as Matplotlib and the Basemap toolkit. These visualizations, which include maps and graphical displays, aid in understanding the data's geographical distribution and magnitude characteristics.
5. Data Publication: The final stage entails publishing the results in a variety of formats. Processed data is made available via interactive HTML reports and visual PNG images, allowing users to interact with and explore earthquake data in greater depth.

Each stage of this lifecycle complements the next, ensuring that the data is accurate, accessible, and actionable from collection to publication. This structured approach not only maximizes data value, but also improves pipeline efficiency.

Additions to the pipeline that can be done:

Several enhancements to the existing "Rent-a-VM to Process Earthquake Data" pipeline could be implemented to improve data handling and analytical capabilities. First, using real-time data streaming tools such as Google Pub/Sub and Google Dataflow could allow for immediate processing of incoming seismic data, lowering latency and allowing for faster response times during seismic events. Second, the use of machine learning models may provide predictive insights, such as forecasting potential aftershock locations and magnitudes, adding a proactive component to the pipeline. Finally, advanced data quality management tools can improve data accuracy and reliability by automatically detecting and correcting anomalies and inconsistencies in real-time earthquake data. These improvements would not only speed up data processing, but also increase the pipeline's utility in seismic research and emergency response planning.