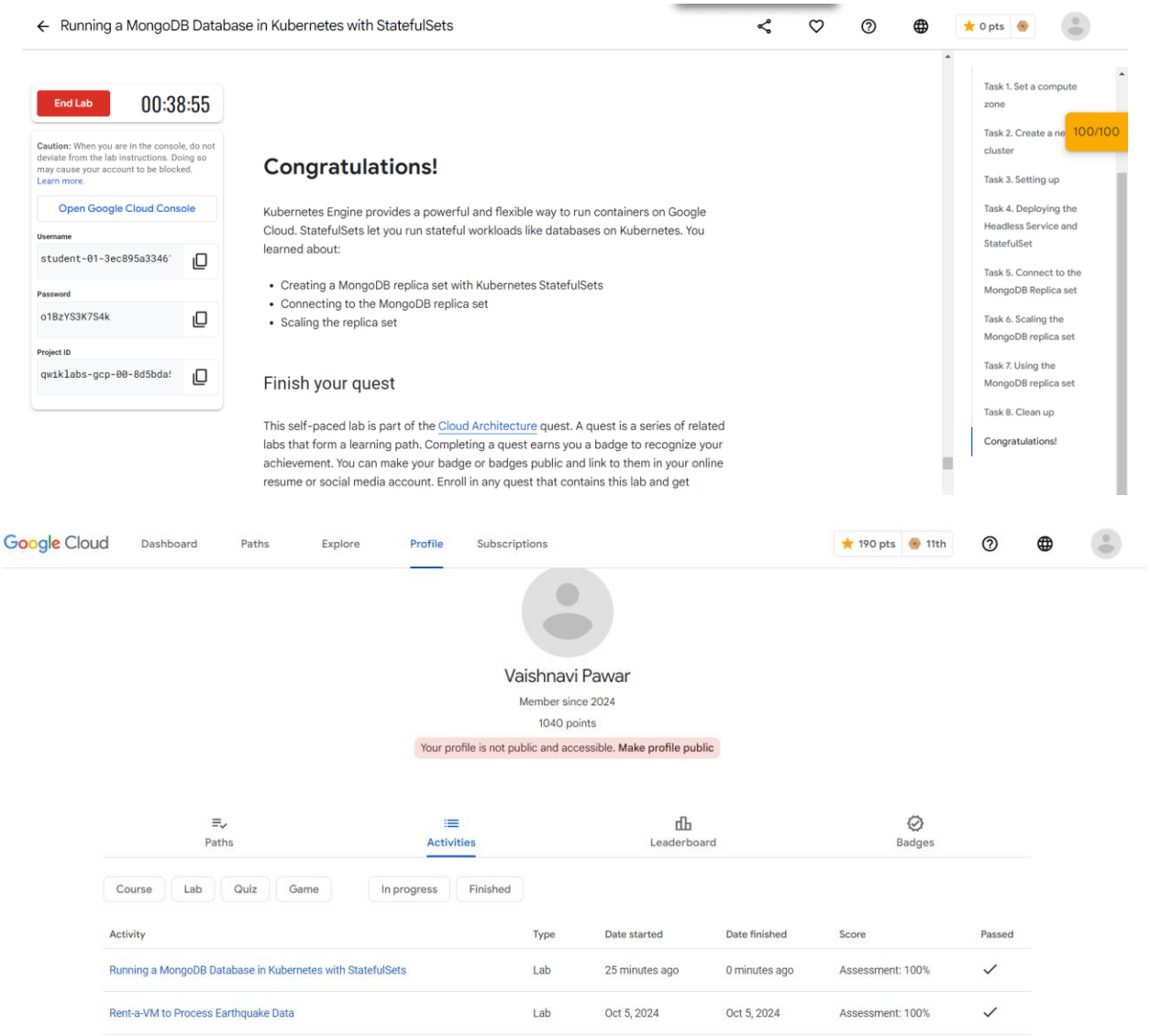
Fall 2024- I535
Mgmt. Access Use Big Data

Vaishnavi Vishwas Pawar
vpawar@iu.edu

MongoDB and BigQuery Assignment

As part of my assignment, I completed hands-on labs to gain practical experience with Google Cloud services. This report focuses on two labs: "Running a MongoDB Database in Kubernetes with StatefulSets" and "Ingesting New Datasets into BigQuery." These labs show how cloud technologies make it easier to deploy and manage databases and large datasets, allowing organizations to focus on effective data analysis. The report provides an overview of the key concepts involved, investigates the underlying infrastructure, and highlights the advantages and disadvantages of using Kubernetes and BigQuery in practical scenarios.

Running a MongoDB Database in Kubernetes with StatefulSets



Vaishnavi Vishwas Pawar
vpawar@iu.edu

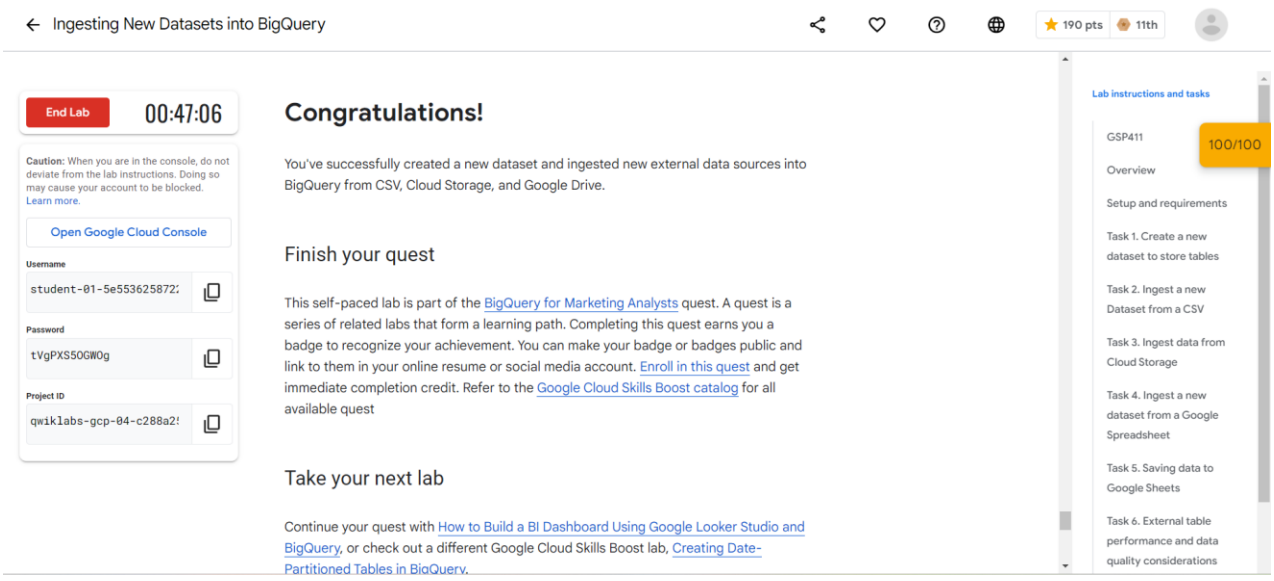
Experience and Key Concepts:

Headless Service: In this lab, I learned about a headless service, which provides direct DNS access to individual pods without load balancing. This is critical for managing MongoDB replica sets because each pod (or instance) must communicate directly with the others to ensure replication and consistency.

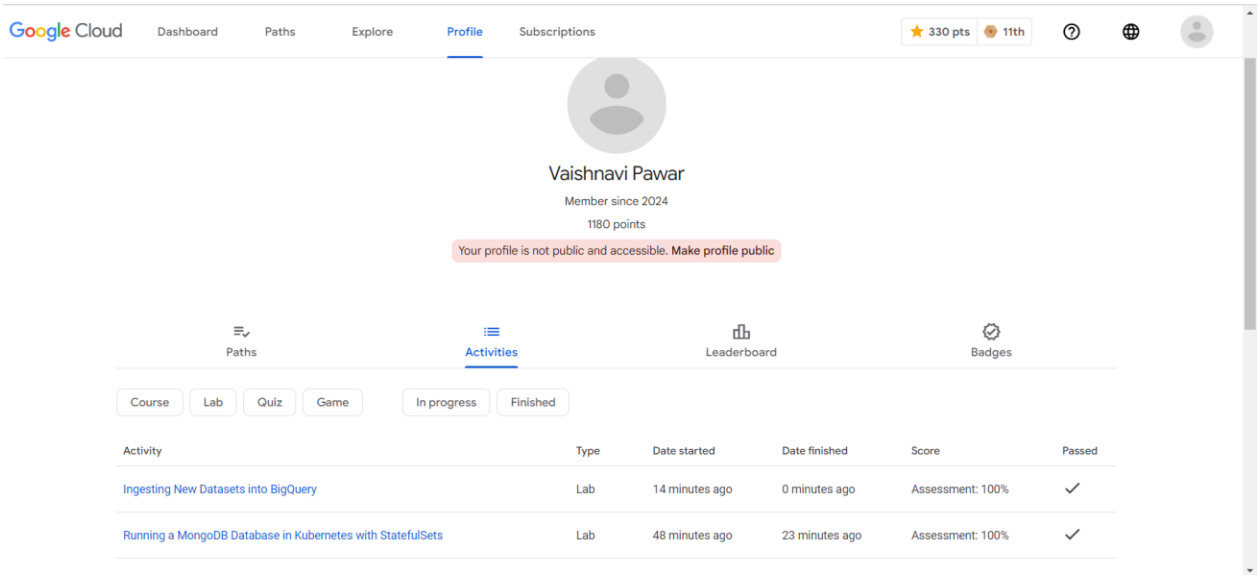
StatefulSet: The StatefulSet resource maintains each MongoDB pod's identity and data across restarts. This setup prevents data loss and ensures that each replica operates independently while remaining in sync with the rest of the cluster.

Running a Database on Containers: Kubernetes uses StatefulSets and persistent volumes to preserve state, making it possible to run databases such as MongoDB in containers. This combination enables scalability, high availability, and durability, demonstrating that containerized databases are an effective solution for production environments.

Ingesting New Datasets into BigQuery



Vaishnavi Vishwas Pawar
vpawar@iu.edu



Experience and Key Concepts:

Dataset Ingestion Methods: I loaded data into BigQuery from various sources, including a CSV file, Google Cloud Storage, and Google Sheets. These ingestion methods demonstrate Big Query's versatility in handling various data sources.

Using SQL for data analysis: To extract meaningful insights, I ran SQL queries on the ingested data. Based on inventory turnover, I identified products with the highest stock levels and those that needed to be restocked right away.

Create an External Table from Google Sheets: The lab also covered adding Google Sheets as an external data source to BigQuery. While this approach is convenient, it has some drawbacks, including a lack of data consistency and reduced query optimization.

Pros and Cons of BigQuery:

Pros:

- Fully Managed Service: Eliminates the need for infrastructure management, freeing users to concentrate on data analysis.
- Scalable and fast: Capable of efficiently handling queries on terabytes of data.
- Flexible Integration: Supports a variety of data sources and works well with other Google Cloud services.
- Cost-Effective: Pay-as-you-go pricing ensures cost-effectiveness based on actual use.

Cons:

- Limited Optimization for External Tables: When working with data stored outside of BigQuery, query performance may degrade.
- Data Consistency Issues: When external sources, such as Google Sheets, are modified during queries, the resulting data can be inconsistent.
- Learning Curve: Users who are unfamiliar with SQL or cloud platforms may require more time to become comfortable.

These labs not only taught technical skills, but also demonstrated the significance of aligning technology with real-world applications. Deploying MongoDB on Kubernetes demonstrated how cloud infrastructure can meet the needs of stateful applications while maintaining seamless scalability and availability. Similarly, working with BigQuery demonstrated how businesses can extract meaningful insights from complex datasets without managing backend infrastructure. This hands-on experience demonstrated that cloud platforms are about more than just technology; they enable faster, more agile responses to business challenges, allowing users to focus on innovation and decision-making.

These labs provided valuable hands-on experience managing containerized databases and using cloud-based analytics tools. The MongoDB lab demonstrated Kubernetes' ability to effectively manage stateful workloads, ensuring data persistence and scalability. The BigQuery lab demonstrated how cloud analytics platforms can streamline data ingestion and analysis while accommodating a variety of data sources. Overall, these labs emphasized the importance of leveraging cloud technologies to simplify complex operations and focus on generating actionable insights.