I535 Spring 2023 - MGMT - Distributed Computing and File System

Submit a short report and the completion screenshots that compare two labs and describe your experience with them. The report should include the following:

- A short description of how you understand Dataproc capabilities (feel free to consult GCP documentation). 1 Pt.
- A summary of what you learned in each lab. 1 Pt.
- A brief discussion of how changing the number of worker nodes would affect the performance in your cluster. 1 Pt.

Dataproc Capabilities

Google Cloud Dataproc is a powerful big data processing service that enables users to run popular frameworks such as Apache Hadoop and Apache Spark on the cloud. With Dataproc, users can easily create and manage clusters, scale resources based on demand, and integrate with other Google Cloud services. One of the key benefits of Dataproc is its ease of use. Users can create clusters with just a few clicks, and Dataproc takes care of the underlying infrastructure and configuration. This makes it easy for organizations to focus on their data processing tasks rather than the underlying infrastructure.

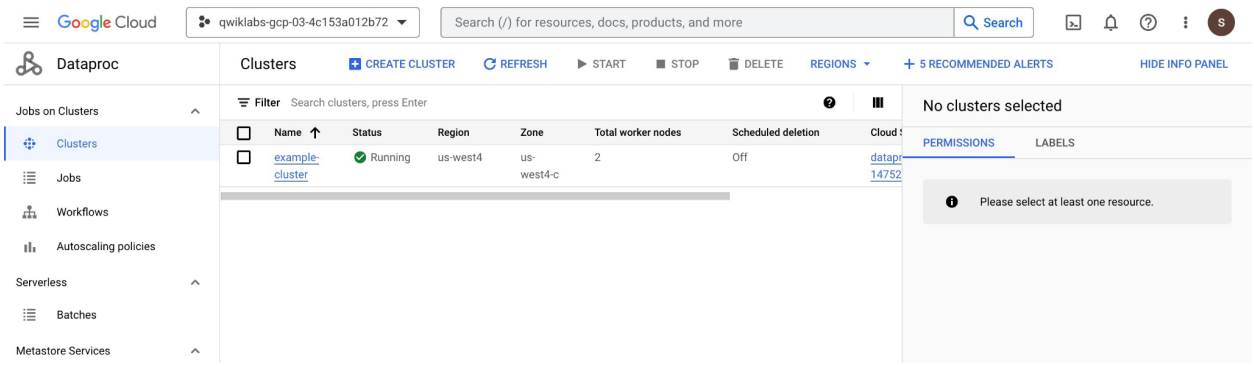
Additionally, Dataproc offers flexibility and customization options, allowing users to specify the versions of frameworks they want to use and automate setup tasks. The service also supports data encryption and integrates with Cloud IAM for access control. Overall, Google Cloud Dataproc is a robust and versatile big data processing service that offers many benefits to organizations looking to process and analyze large amounts of data on the cloud.

Dataproc: Qwik Start - Console

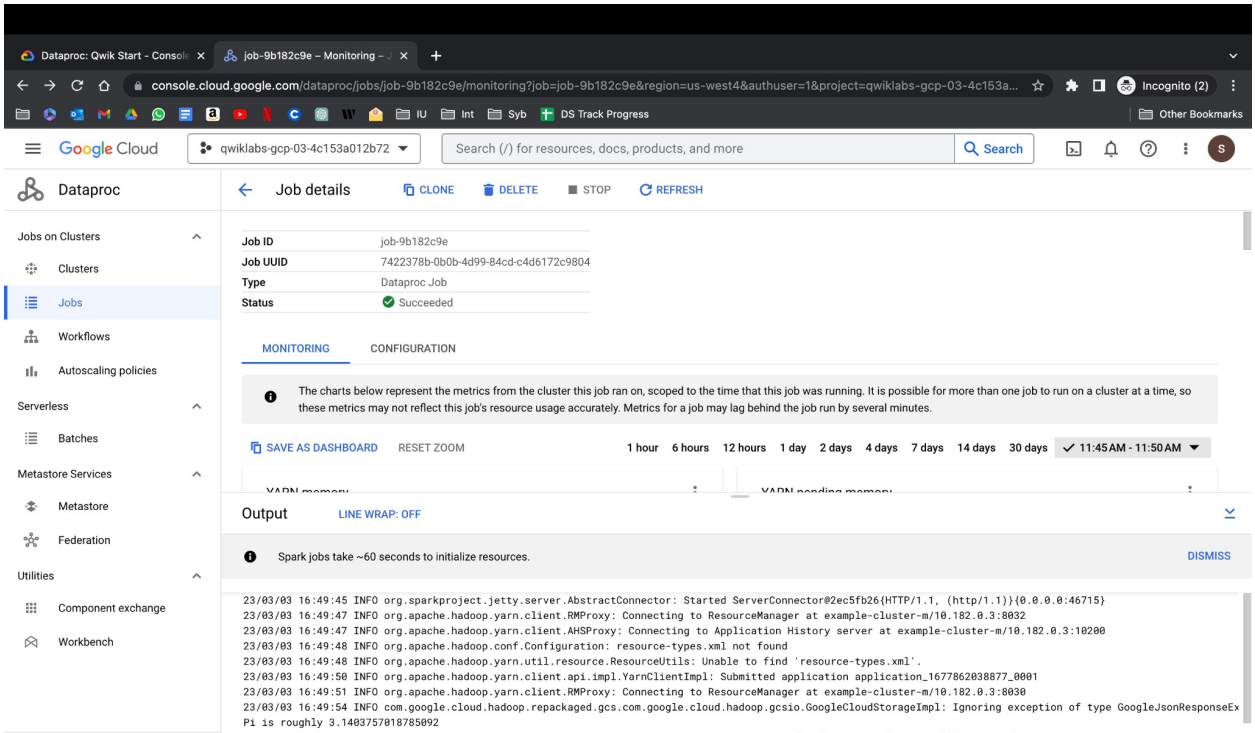
Learning of Lab 1:

In the Google Cloud Dataproc Qwik Start - Console lab, users learn the basics of creating and managing Dataproc clusters using the Google Cloud Console. This lab provides a hands-on introduction to the key features of Dataproc, such as cluster creation, job submission, and monitoring. Users learn how to create a cluster using the Cloud Console, configure cluster settings such as cluster size and machine type, and submit a job to the cluster using the Cloud Shell command line. The lab also covers how to monitor job progress and view job logs using the Dataproc web interface.

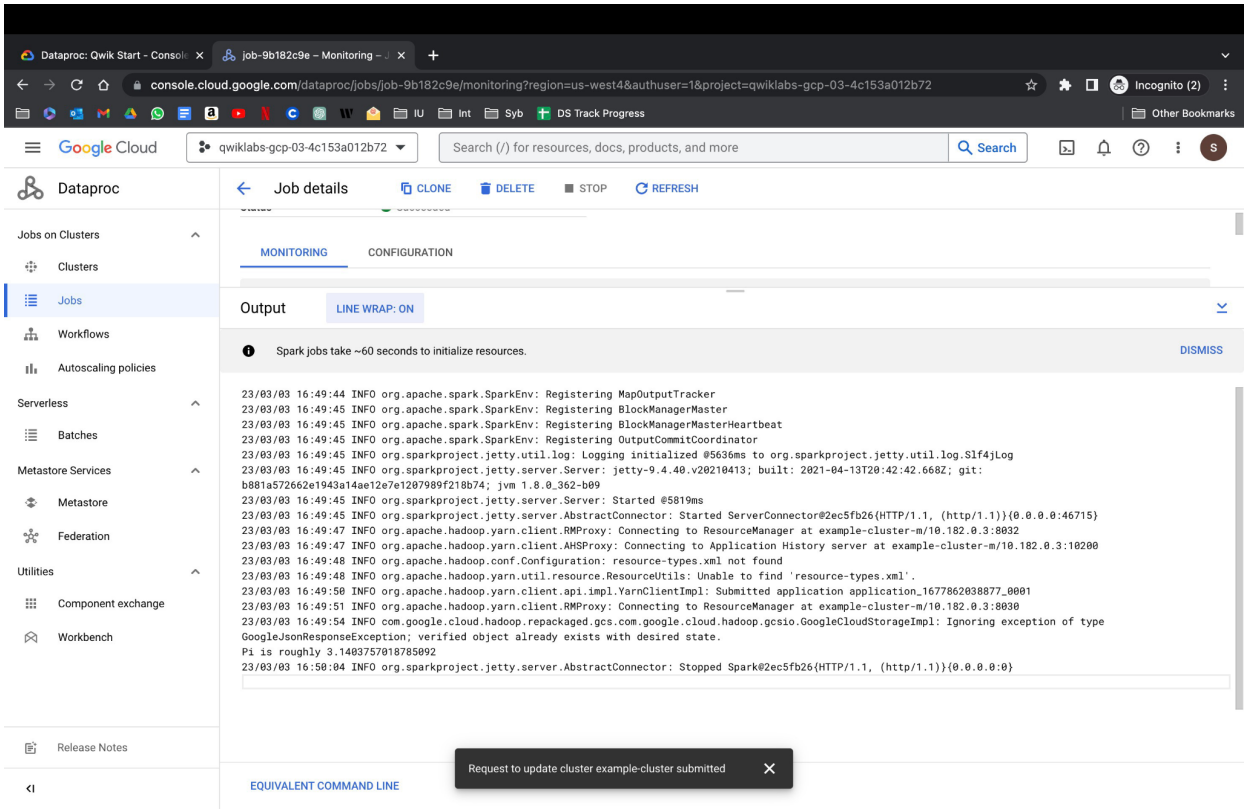
Task 1: Create Cluster



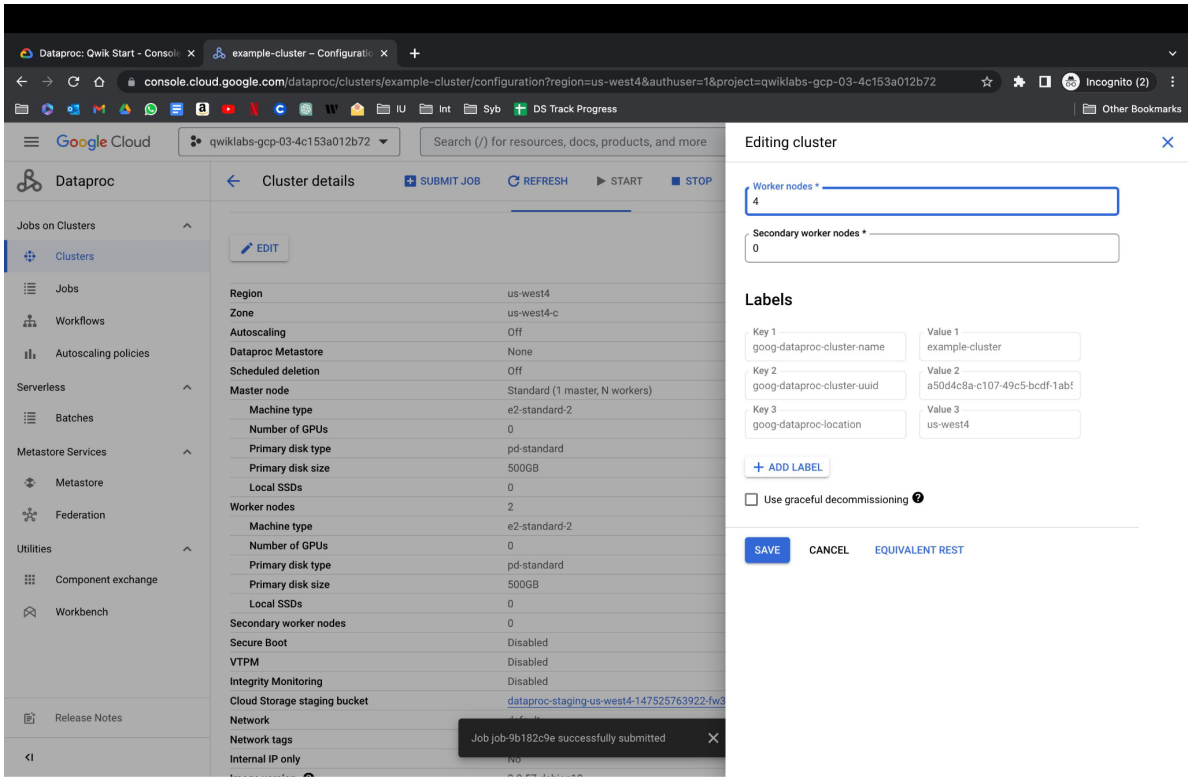
Task 2:



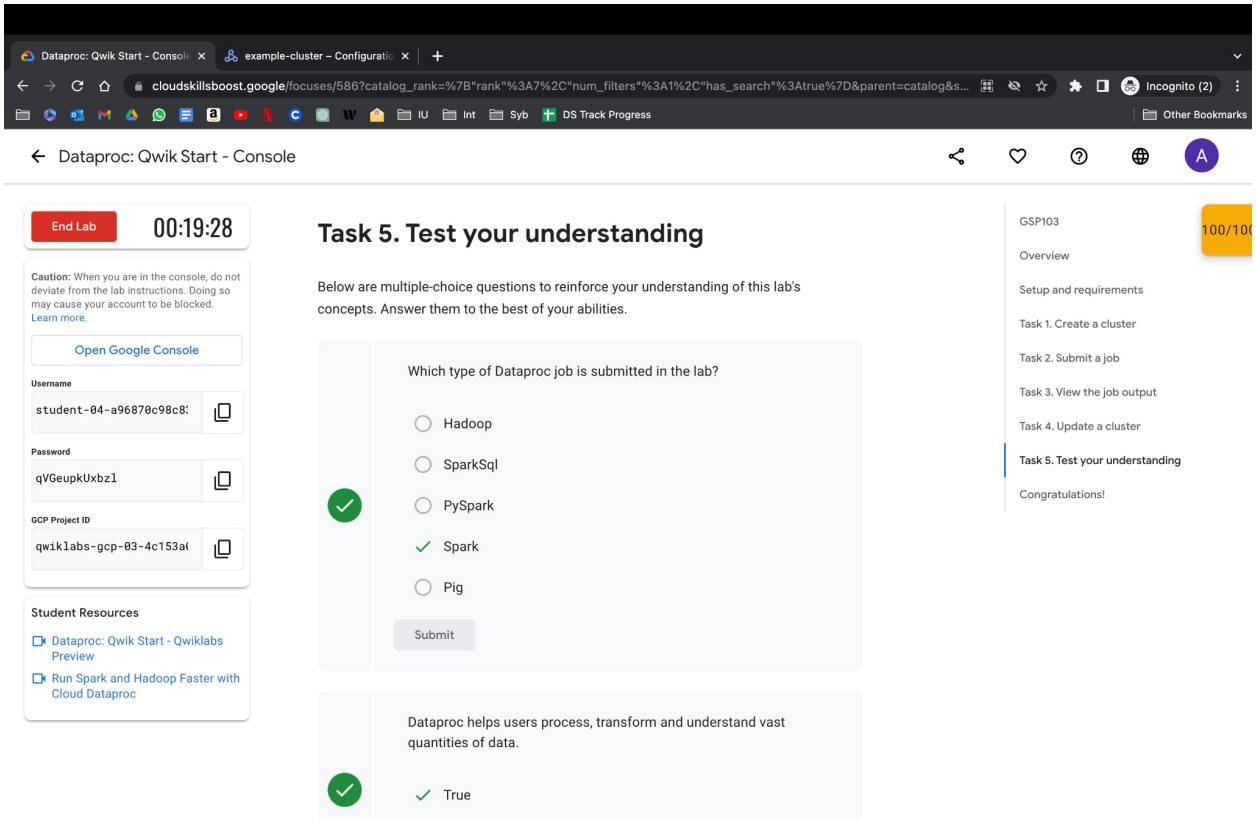
Task 3: Task 3. View the job output



Task 4: Update Clusters



Task 5: Completion



Q. A brief discussion of how changing the number of worker nodes would affect the performance in your cluster.

The number of worker nodes in a Dataproc cluster can significantly affect cluster performance. By dispersing the burden across more machines, increasing the number of worker nodes can enhance performance and allow for faster data processing. Adding extra nodes, however, raises the cost of running the cluster.

Reducing the number of worker nodes, on the other hand, can reduce performance by limiting the cluster's available processing capacity. With large datasets, this can result in slower job completion times and longer processing durations.

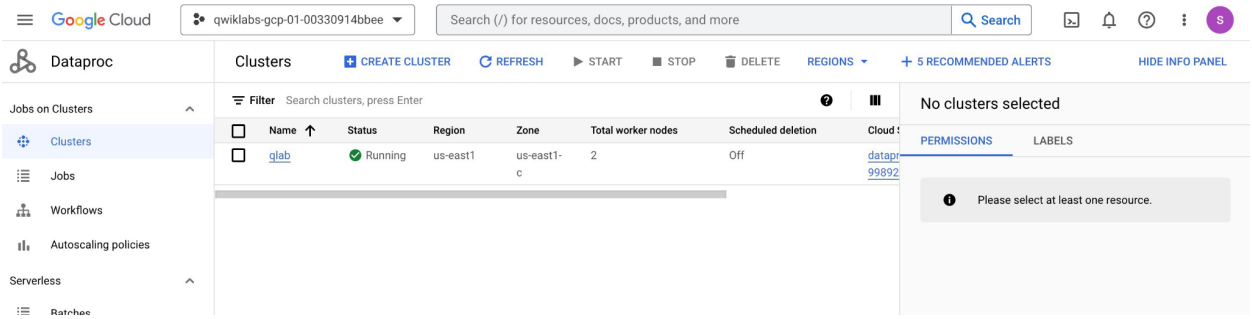
When altering the number of worker nodes in a Dataproc cluster, it is critical to carefully assess the workload and processing requirements. For example, adding more worker nodes may yield considerable performance increases for workloads with high parallelism, whereas adding more worker nodes may not enhance performance as much for tasks with low parallelism.

Introduction to Cloud Dataproc: Hadoop and Spark on Google Cloud

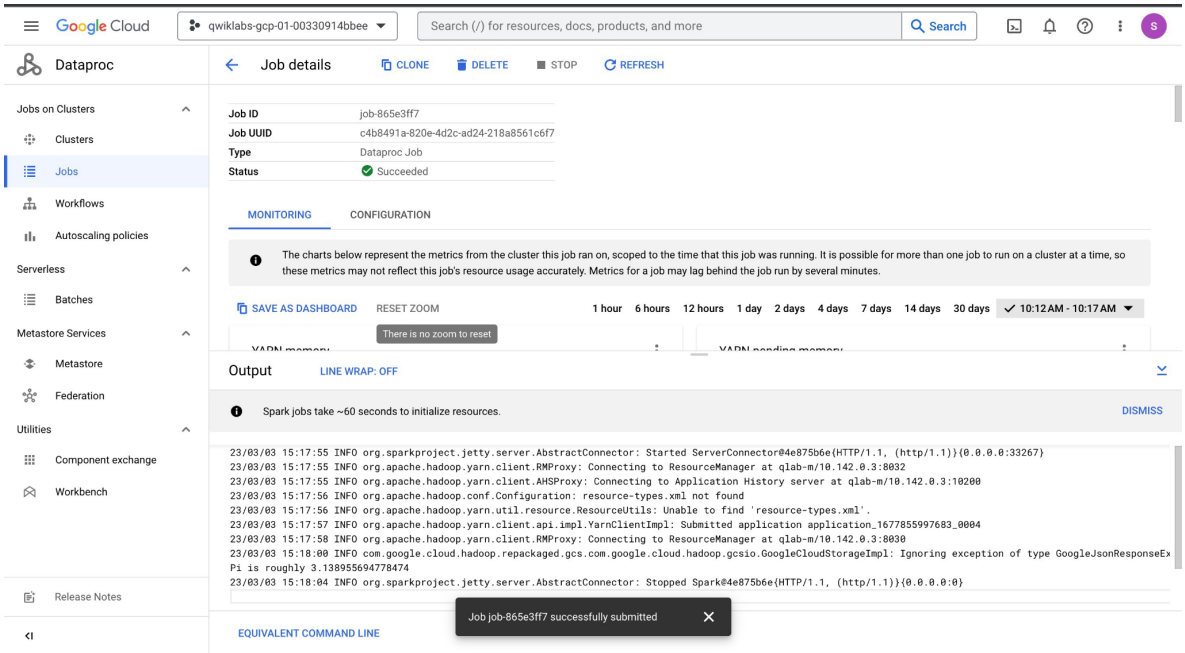
Learning of Lab 2:

In the Introduction to Cloud Dataproc: Hadoop and Spark on Google Cloud lab, users build on their knowledge of Dataproc by learning how to create a cluster using the Cloud SDK command line. This lab also covers how to submit Hadoop and Spark jobs to the cluster and monitor their progress using the Dataproc web interface. Additionally, users learn how to use the Hadoop Distributed File System (HDFS) to store and manage large amounts of data, and how to run a simple Spark job using the PySpark library. This lab provides a more in-depth look at Dataproc's capabilities and shows users how to use its features to process and analyze large datasets.

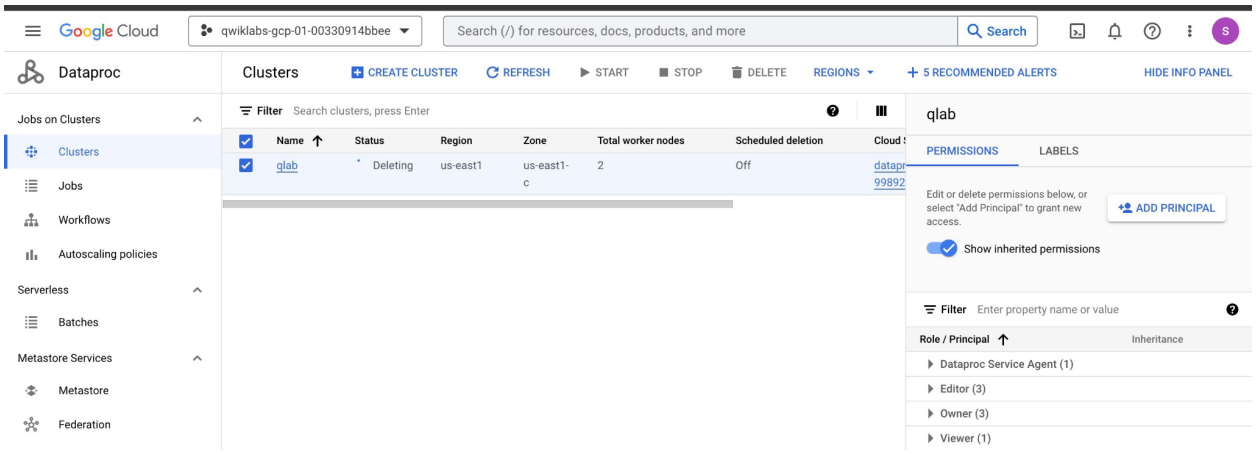
Task 1: Create a Cloud Dataproc cluster



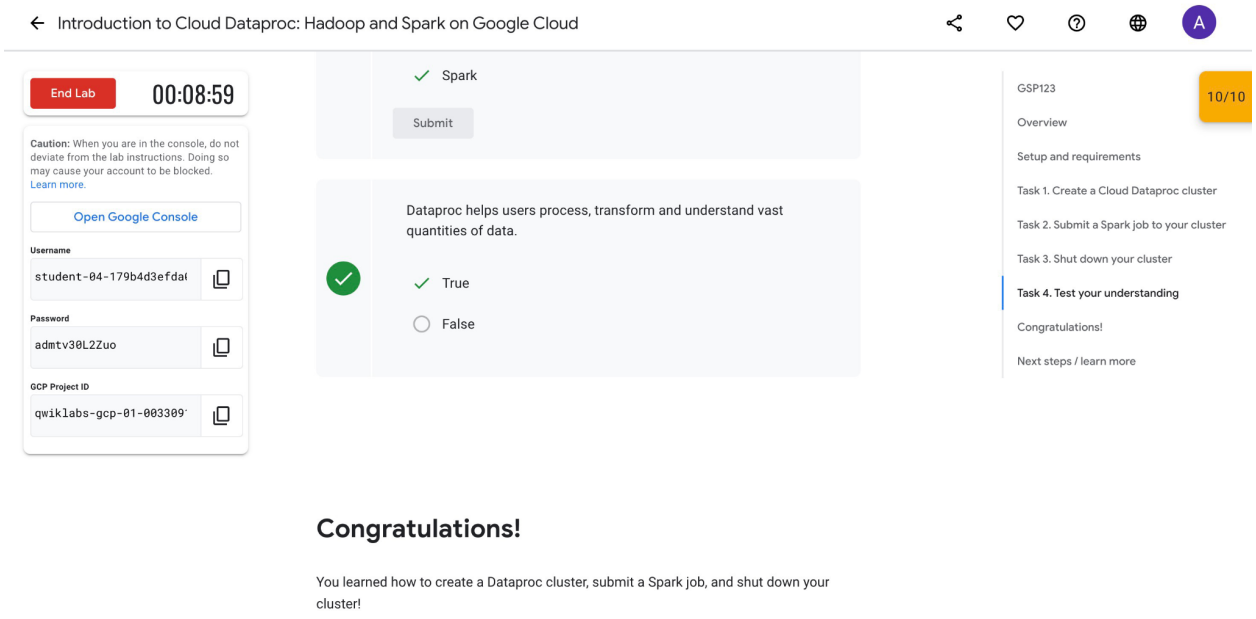
Task 2: Submit a Spark job to your cluster



Task 3: Shut down your cluster



Task 4: Completion



Overall, these labs provide users with a solid foundation for working with Google Cloud Dataproc. They cover the basics of cluster creation and management, job submission, and monitoring, as well as more advanced topics such as HDFS and PySpark. By completing these labs, users will be well-equipped to use Dataproc to process and analyze big data on the cloud.