

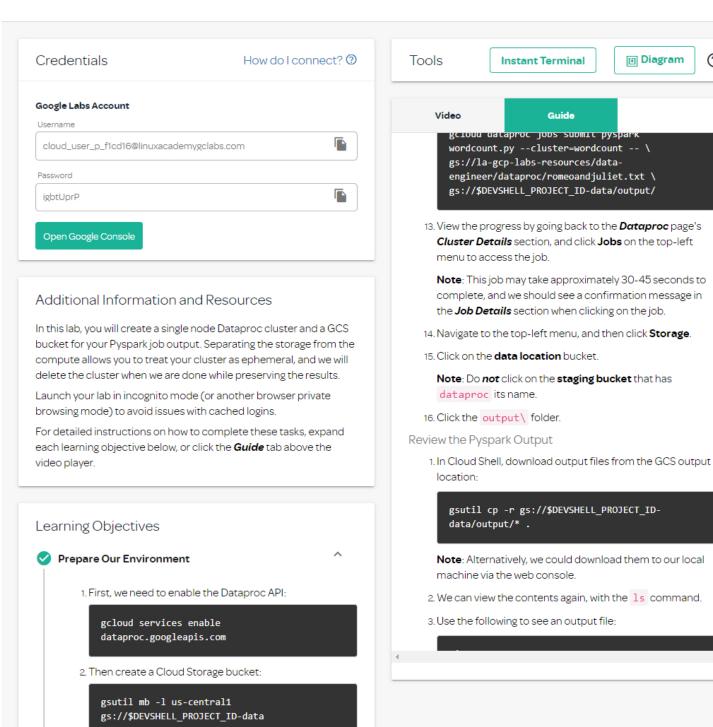
As we work to finalize work on our Course Progress system, some variations may be seen. Please know that your progress is being tracked properly, even if it is not shown properly on the display. If you have an immediate need, please contact Support so we can investigate your account individually. Thank you for your continued patience!

## Running a Pyspark Job on Cloud Dataproc Using Google Cloud Storage





(?)



3. Now create the dataproc cluster:

- gcloud dataproc clusters create wordcount -zone=us-central1-f --single-node --mastermachine-type=n1-standard-2
- 5. And finally, download the wordcount.py file that will be used for the pyspark job:

```
gsutil cp -r gs://la-gcp-labs-
resources/data-engineer/dataproc/* .
```

Submit the Pyspark Job to the Dataproc Cluster

```
gcloud dataproc jobs submit pyspark wordcount.py -
-cluster=wordcount -- \
gs://la-gcp-labs-resources/data-
engineer/dataproc/romeoandjuliet.txt \
gs://$DEVSHELL_PROJECT_ID-data/output/
```

Review the Pyspark Output

In Cloud Shell, type:

 In Cloud Shell, download output files from the GCS output location:

```
gsutil cp -r gs://$DEVSHELL_PROJECT_ID-
data/output/* .
```

**Note**: Alternatively, we could download them to our local machine via the web console.

Delete the Dataproc Cluster

- We don't need our cluster any longer, so let's delete it. In the web console, go to the top-left menu and into BIGDATA > Dataproc.
- Select the wordcount cluster, then click **DELETE**, and **OK** to confirm.

Our job output still remains in Cloud Storage, allowing us to delete Dataproc clusters when no longer in use to save costs, while preserving input and output resources.