

About

This module includes all prerequisites for running the Serverless Spark lab-

1. [Declare variables](#)
2. [Create a Bucket](#)
3. [Create a Spark Persistent History Server](#)

0. Prerequisites

1. GCP Project Details

Note the project number and project ID.
We will need this for the rest of the lab

2. IAM Roles needed to create Persistent History Server

Grant the following permissions - Viewer - Dataproc Editor - Storage Admin

3. Attach cloud shell to your project.

Open Cloud shell or navigate to shell.cloud.google.com

Run the below command to set the project in the cloud shell terminal:

```
gcloud config set project $PROJECT_ID
```

1. Declare variables

We will use these throughout the lab.

Run the below in cloud shells copied to the project you selected-

```
PROJECT_ID= #Project ID
REGION= #Region to be used
BUCKET_PHS= #Bucket name for Persistent History Server
PHS_NAME = # Spark Persistent History Server name
```

2. Create a bucket

A bucket is created which will be attached to history server for storing of application logs.

```
gsutil mb -p $PROJECT_ID -c STANDARD -l $REGION -b on gs://$BUCKET_PHS
```

3. Create a Spark Persistent History Server

A single node dataproc cluster will be created with component gateways enabled.

```
gcloud dataproc clusters create $PHS_NAME \
  --project=${PROJECT_ID} \
  --region=${REGION} \
  --single-node \
  --image-version=2.0 \
  --enable-component-gateway \
  --properties=spark:spark.history.fs.logDirectory=gs://${BUCKET_PHS}/phs/*/spark-job-history
```