Deploy Spark Cluster on Google Cloud Dataproc:

Cloud Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.

Create a bucket – The following steps need to be taken only once:

1. Login to your Google Cloud account (https://cloud.google.com) and go to Cloud Storage.
2. Create a new bucket – Please note that if you already have a bucket that would like to use for this purpose you can skip the remainder of this section.
3. Give your bucket a unique name, select **Regional** and for location select **us-**east1, and leave the rest as default and click on **Create**
4. To test everything is working fine from Cloud SDK run the following query. This should list all of the buckets in your project including the one you just made: `gsutil ls`

Create a cluster – The following steps need to be done every time we start a new cluster:

1. Login to your Google Cloud account (https://cloud.google.com) and open a Google Cloud Shell. Make sure you are in the intended project.

   Important note: Now that you have Cloud SDK installed on your laptop you can simply use that shell instead of Google Cloud Shell.

2. Run the following script on Cloud SDK shell (or Google Cloud Shell) to launch a new cluster named *mycluster*. Copy this command to an editor of your choice and make sure to replace <PROJECT-ID> with your project-id (it can be found by clicking on the project name, under ID) and <BUCKET-NAME> with the bucket you created in the previous section:

```
gcloud dataproc clusters create mycluster \
  --project <PROJECT-ID> \
  --bucket <BUCKET-NAME> \
  --region us-east1 \
  --subnet default \
  --zone us-east1-b \
  --single-node \
  --master-machine-type n1-highmem-2 \
  --master-boot-disk-size 30 \
  --image-version 1.3-deb9 \
  --initialization-actions gs://dataproc-initialization-actions/jupyter/jupyter.sh \
  --initialization-actions gs://dataproc-initialization-actions/connectors/connectors.sh \
    --metadata 'gcs-connector-version=1.7.0' \
    --metadata 'bigquery-connector-version=0.11.0'
```

   (This script is also accessible in 05-Basic-DF-Operations/ Deploy-Dataproc-Cluster-Single-Node.txt)

   - This will give you a single-node cluster with 30 GB HHD, 2 CPU cores, and 13GB Memory (n1-highmem-2)

3. From Dataproc page (located under Big Data in the Google Cloud Platform menu) confirm that your cluster is up and running.

Connecting to the Cluster

1. From Cloud SDK shell execute the following command to establish a secure SSH tunnel:

```
gcloud compute ssh --zone us-east1-b mycluster-m -- -L 2222:localhost:8123
```

2. You can now log in to Jupyter notebook on your master node from your local browser:

```
localhost:2222
```

Note: You could get disconnected time to time, if that happened you can simply repeat step 1 and refresh your browser. Your work will stay saved in your cluster and in your bucket under *notebook* folder.

Clean up

To make sure we won't be charged for any of the resources we are not using, delete the cluster after each use:

- From the Dataproc page select the cluster and click on the DELETE button.
    - Alternatively, you can also use the following command from **a new** Cloud SDK terminal (please check from the UI to make sure the cluster is being deleted):

```
gcloud dataproc clusters delete mycluster --region us-east1
```

Note1: Notice that even after deleting the cluster your notebooks will persist in the bucket and when you create a new cluster that points to the same bucket you can simply reuse those notebooks.

Note2: You can't terminate a cluster from within itself. Make sure that you open a new Cloud SDK terminal and you are not using the one that is tunneled to the cluster.

Deploying a Large Cluster

To create a large cluster (CAUTION) with multiple nodes, use the following command:

```
gcloud dataproc clusters create bigcluster \
 --project <PROJECT-ID> \
 --bucket <BUCKET-NAME> \
 --region us-east1 \
 --subnet default \
 --zone us-east1-b \
 --master-machine-type n1-highmem-2 \
 --master-boot-disk-size 50 \
 --num-workers 2 \
 --worker-machine-type n1-highmem-2 \
 --worker-boot-disk-size 50 \
 --num-preemptible-workers 2 \
 --image-version 1.3-deb9 \
 --initialization-actions gs://dataproc-initialization-actions/jupyter/jupyter.sh \
 --initialization-actions gs://dataproc-initialization-actions/connectors/connectors.sh \
  --metadata 'gcs-connector-version=1.7.0' \
  --metadata 'bigquery-connector-version=0.11.0'
```

(This script is also accessible in 05-Basic-DF-Operations/ Deploy-Dataproc-Cluster-Multi-Node.txt)

Important note: Please be advised of its cost. Cost can be calculated using Dataproc cost calculator (make sure to include GCE is selected): https://cloud.google.com/products/calculator/

The cluster above has 1 master node and 4 workers (10 CPUs & 80 GB memory). Two of these workers are preemptible workers. Preemptible instances are offered with a big discount (~70%) but will not last for more than 24 hours. This cluster will cost about $0.60/hour.

Connecting to the Cluster

1. From Cloud SDK shell execute the following command to establish a secure SSH tunnel:

```
gcloud compute ssh --zone us-east1-b mycluster-m -- -L 2222:localhost:8123
```

2. You can now log in to Jupyter notebook on your master node from your local browser:

```
localhost:2222
```

Clean up

```
gcloud dataproc clusters delete bigcluster --region us-east1
```