Dataproc staging and temp buckets

When you create a cluster, by default, Dataproc creates a Cloud Storage staging and a Cloud Storage temp bucket (/storage/docs/key-terms#buckets) in your project or reuses existing Dataproc-created staging and temp buckets from previous cluster creation requests.

- Staging bucket: Used to stage cluster job dependencies, job driver output (/dataproc/docs/guides/driver-output#accessing_job_driver_output), and cluster config files. Also receives output from the Cloud SDK gcloud dataproc clusters diagnose (/dataproc/docs/support/diagnose-command) command.
- Temp bucket: Used to store ephemeral cluster and jobs data, such as Spark and MapReduce history files.

If you do not specify a staging ot temp bucket, Dataproc sets a Cloud Storage location in US, ASIA, or EU (/storage/docs/locations#location-mr) for your cluster's staging and temp buckets according to the Compute Engine zone where your cluster is deployed, and then creates and manages these project-level, per-location buckets. Dataproc-created staging and temp buckets are shared among clusters in the same region. By default, temp bucket has a TTL of 90 days.

Instead of relying on the creation of a default staging and temp bucket, you can specify existing Cloud Storage buckets that Dataproc will use as your cluster's staging and temp bucket.

gcloud commandREST API (#rest-api)Console (#console)

Run the `gcloud dataproc clusters create` command with the `--bucket` (/sdk/gcloud/reference/dataproc/clusters/create#--bucket) and/or `--temp-bucket` (/sdk/gcloud/reference/dataproc/clusters/create#--temp-bucket) flags locally in a terminal window or in Cloud Shell (https://console.cloud.google.com/?cloudshell=true) to specify your cluster's staging and/or temp bucket.

```
$ gcloud dataproc clusters create cluster-name \
    --region=region \
    --bucket=bucket-name \
    --temp-bucket=bucket-name \
    other args ...
```

Dataproc uses a defined folder structure for Cloud Storage buckets attached to clusters. Dataproc also supports attaching more than one cluster to a Cloud Storage bucket. The folder structure used for saving job driver output in Cloud Storage is:

```
-storage-bucket-name
oogle-cloud-dataproc-metainfo
 list of cluster IDs
   - list of job IDs
     - list of output logs for a job
```

You can use the `gcloud` command line tool, Dataproc API, or Google Cloud Console to list the name of a cluster's staging and temp buckets.

gcloud commandREST API (#rest-api)Console (#console)

Run the `gcloud dataproc clusters describe` (/sdk/gcloud/reference/dataproc/clusters/describe) command locally in a terminal window or in Cloud Shell (https://console.cloud.google.com/?cloudshell=true). The staging and temp buckets associated with your cluster are listed in the output.

```
$ gcloud dataproc clusters describe cluster-name \
    --region=region \
...
clusterName: cluster-name
clusterUuid: daa40b3f-5ff5-4e89-9bf1-bcbfec ...
config:
    configBucket: dataproc-...
    ...
    tempBucket: dataproc-temp...
```

Last updated 2020-08-14 UTC.