Databricks (GCP)

Last updated: December 8, 2023

Migration guide for init scripts on cloud storage (GCS)

For Databricks Runtime 11.2 and below, move init scripts from DBFS to Google Cloud Storage (GCS).

Note: In cases where your init scripts are "self-contained," i.e., DO NOT reference other files such as libraries, configuration files, or shell scripts. We highly recommend storing init scripts as workspace files.

In general, you must do the following:

1) Configure the clusters to authenticate to the GCS.
2) Copy all your init scripts and files referenced by the init scripts from DBFS to GCS.
3) [If applicable] Update the init scripts to reference files on GCS.
4) Update cluster configuration and cluster policies to reference the init scripts on GCS.

1) Configure the clusters to authenticate to the GCS using Google Cloud service accounts

Follow the instructions in our documentation to configure your clusters with a Google Cloud service account to access GCS.

2) Copy all your init scripts from DBFS to GCS

Identify all clusters, jobs, DLT pipelines, etc. that use init scripts on DBFS. Use the script detection notebook to prepare a list of individual init scripts. For init scripts that reference files on DBFS (e.g. libraries, configuration files, or other files), prepare the list of referenced files.

1. Create a new GCS bucket (or use an existing one) to move all the init scripts and referenced files from DBFS to GCS.
2. Check the bucket permissions (IAM) and confirm that the service account specified in the cluster's configuration can access this new GCS bucket.
3. In your Databricks Notebook, use the dbutils.fs.cp command to copy files from DBFS to GCS:

```
dbutils.fs.cp("dbfs:/<path>", "gs://<bucket>/<path>")
```

3) Update the init scripts to point to GCS locations for files referenced in the init script

You need to modify your init scripts that reference files such as libraries, other scripts, or configuration files, to use GCS paths. Change references inside the init script from `/dbfs/path` to `gs://bucket/path`.

With the service account attached to the cluster, you can directly use the gcloud CLI tool to work with files on GCS. The gcloud CLI can be installed using the `apt-get` command. Use `gcloud storage cp` command (see documentation) to copy files from GCS to the local disk as follows:

```
#!/bin/bash

# Install gcloud cli:
https://cloud.google.com/sdk/gcloud#download_and_install_the
curl https://packages.cloud.google.com/apt/doc/apt-key.gpg | gpg --dearmor
-o /usr/share/keyrings/cloud.google.gpg \
  && echo "deb [signed-by=/usr/share/keyrings/cloud.google.gpg]
https://packages.cloud.google.com/apt cloud-sdk main" | tee -a
/etc/apt/sources.list.d/google-cloud-sdk.list \
  && apt-get update && apt-get install google-cloud-cli

# This command will use authentication of attached service account
gcloud storage cp "gs://<bucket>/<path>/file-to-read.txt local-file.txt
```

4) Update cluster configuration and cluster policies to reference the init scripts on GCS

Change the init script paths in affected clusters, jobs, Delta Live Tables pipelines, and cluster policies to point to GCS instead of DBFS. If you use the init scripts detection notebook, click links in the generated HTML tables:

- **Clusters**: edit your existing clusters' configuration: update init scripts from source "DBFS" to source "GCS", providing the path to the file on GCS, e.g. `gs://<bucket>/path/to/init-script` (more information).
  Note: If you are using cluster policies, you must update both the cluster policy and the configurations of any other clusters using the policy. Cluster policy changes do not propagate to clusters using that policy.
- **Jobs**: Edit the cluster configuration for each task that uses a dedicated job cluster and each shared job cluster. Update the init script location from DBFS to GCS, as described above.

- **If using Cluster policies**: Open and edit the cluster policy. Search for blocks like the following, where N is the item number

```
{
  "init_scripts.N.dbfs.destination": {
    "type": "fixed",
    "value": "dbfs:/FileStore/init-scripts/empty_init_script.sh"
  }
}
```

and replace the value, adjusting the file path from DBFS to GCS:

```
{
  "init_scripts.N.gcs.destination": {
      "type": "fixed",
      "value": "gs://<bucket>/empty_init_script.sh"
  }
}
```

- **Delta Live Tables pipelines**: Open the pipeline settings, select the "JSON" tab, and in the cluster definition(s), change entries in the `init_scripts` array from DBFS to GCS;

```
{
 "dbfs": {
   "destination": "dbfs:/FileStore/init-scripts/empty_init_script.sh"
 }
}
```

```
{
 "gcs": {
   "destination": "gs://<bucket>/empty_init_script.sh"
 }
}
```