BigQuery Health Check Metadata Upload Guide

Product Version	4.8.2
Document Type	Health Check Preparation Guide
Authors	BigQuery Data source Team
Reviewer	Red Team & Architects
Approver	CTO
Total Pages	6
Document Status	Draft

Table Of Contents

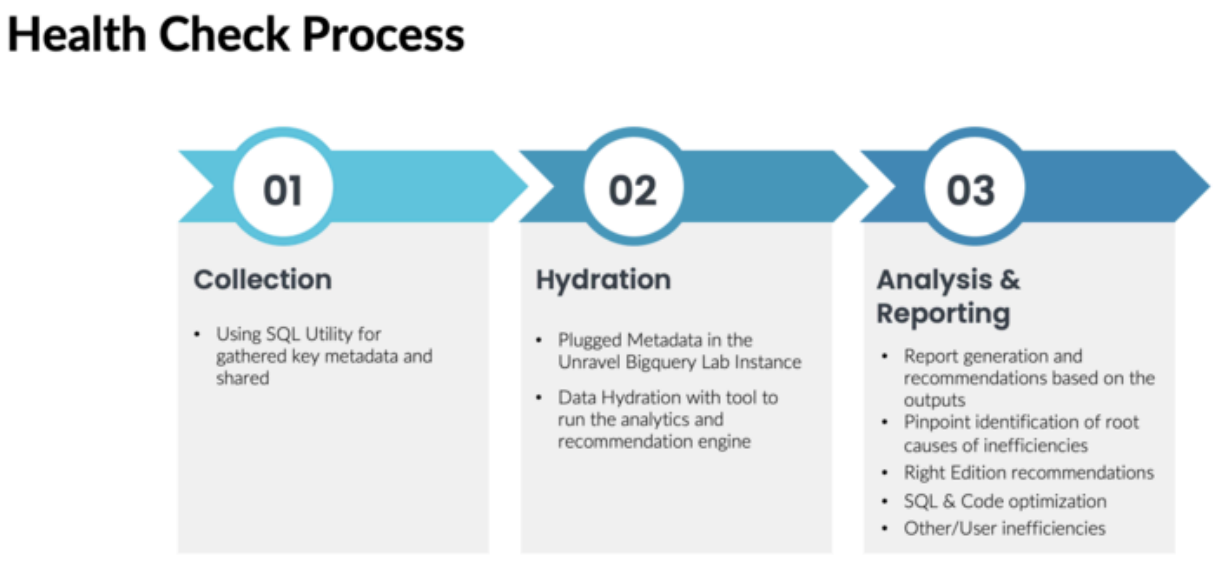
Document Version Record

Date	Version #	Author	Remarks / Reason
05-Jan-23	1.0	Dev Team	New Document

1.1 Objectives

Health check upload for bigquery unravel product.

1.2 Architecture



1.3 Prerequisite

A. **BigQuery Permissions/roles required to upload metadata :**

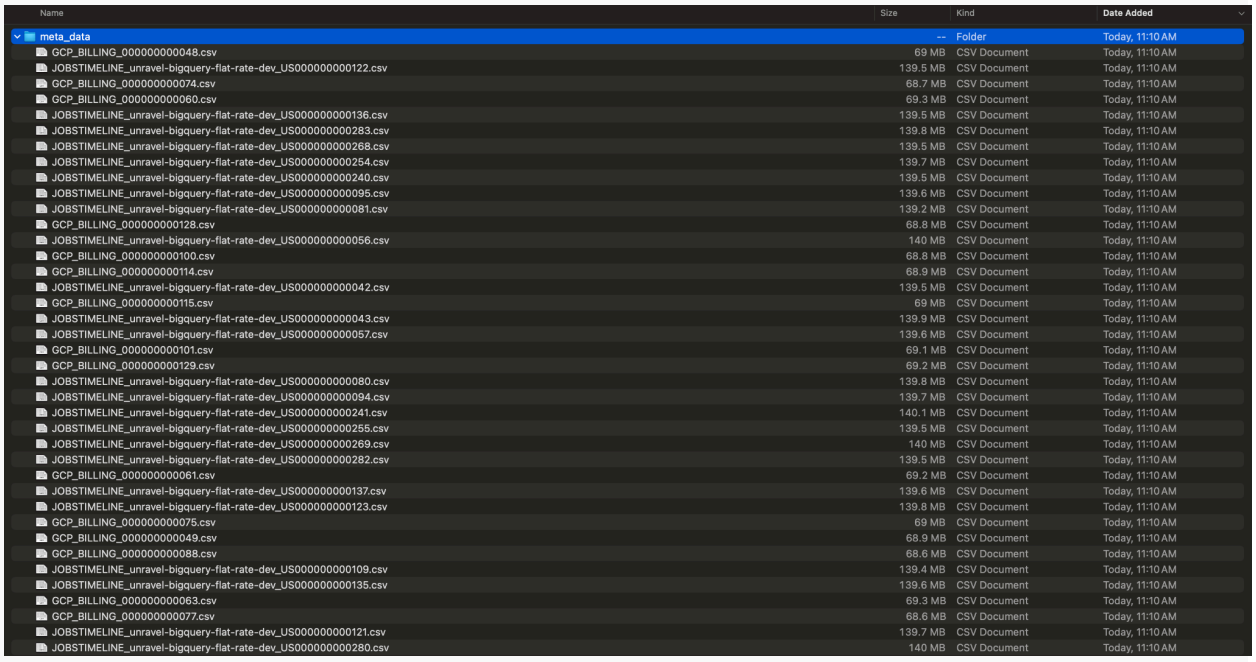
a. Permissions required to upload the JOBS, JOBS_TIMELINE and BILLING metadata to bigquery dataset :

i. bigquery.datasets.create
ii. bigquery.datasets.get
iii. bigquery.jobs.create
iv. bigquery.tables.create
v. bigquery.tables.updateData
vi. storage.managedFolders.get
vii. storage.managedFolders.list
viii. storage.objects.get
ix. storage.objects.list

1.4 Upload BigQuery Metadata for health check.

Do the following to upload BigQuery metadata :

1. Upload these csv files to a gcs bucket(s) (Steps for creating a gcs bucket) :



Name	Size	Kind	Date Added
meta_data	--	Folder	Today, 11:10 AM
GCP_BILLING_000000000048.csv	69 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000122.csv	139.5 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000074.csv	68.7 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000060.csv	69.3 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000136.csv	139.5 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000283.csv	139.8 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000268.csv	139.5 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000254.csv	139.7 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000240.csv	139.5 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000095.csv	139.6 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000081.csv	139.2 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000128.csv	68.8 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000056.csv	140 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000100.csv	68.8 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000114.csv	68.9 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000042.csv	139.5 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000115.csv	69 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000043.csv	139.9 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000057.csv	139.6 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000101.csv	69.1 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000129.csv	69.2 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000080.csv	139.8 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000094.csv	139.7 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000241.csv	140.1 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000255.csv	139.5 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000269.csv	140 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000282.csv	139.5 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000061.csv	69.2 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000137.csv	139.6 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000123.csv	139.8 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000075.csv	69 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000049.csv	68.9 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000088.csv	68.6 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000109.csv	139.4 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000135.csv	139.6 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000063.csv	69.3 MB	CSV Document	Today, 11:10 AM
GCP_BILLING_000000000077.csv	68.6 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000121.csv	139.7 MB	CSV Document	Today, 11:10 AM
JOBSTIMELINE_unravel-bigquery-flat-rate-dev_US000000000280.csv	140 MB	CSV Document	Today, 11:10 AM

2. Clone this repo: https://github.com/unraveldata-org/BigQuery-data-loader.git
3. Follow below Steps to create a Service Account credential key that authenticates to upload the metadata to a bigquery dataset :

a. cd terraform/healthcheck
b. cp input.tfvars.example input.tfvars

c. open input.tfvars
d. Under monitoring_project_ids add the project id where you want to upload data to bigquery and connect with unravel to perform health check
e. Under svc_account_project_id add the project on which you want the service account to be created to execute the upload script
f. Now, open local.tf
g. Under monitoring_project_role_permission add the below permissions, comma separated (ignore if the same permissions are already present) :
    i. “bigquery.datasets.create”
    ii. “bigquery.datasets.get”
    iii. “bigquery.jobs.create”
    iv. “bigquery.tables.create”
    v. “bigquery.tables.updateData”
    vi. “storage.managedFolders.get”
    vii. “storage.managedFolders.list”
    viii. “storage.objects.get”
    ix. “storage.objects.list”
h. Before using this project, you need to authenticate with Google Cloud using gcloud. Follow the instructions provided at https://cloud.google.com/sdk/docs/install-sdk for a one-time configuration. You can find the installation instruction based on the Machine Arch and OS installed in the above link.
i. To authenticate gcloud, execute the following commands:
    i. gcloud init
    ii. gcloud auth application-default login
j. Now, to run terraform execute the following :
    i. terraform init
    ii. terraform plan --var-file=input.tfvars
    iii. terraform apply --var-file=input.tfvars
k. Now, there must be a credential file created inside terraform/healthcheck/keys folder, you will need this credential file while providing configuration for the upload script

4. cd Data Uploader Script
5. cp upload_config.yaml.example upload_config.yaml
6. cd Data Uploader Script
7. Open upload_config.yaml file and edit it with your configuration :

a. jobs_data_gcs_bucket_path: "add the gcs bucket path where you have the csv files of JOBS Data"

b. jobs_timeline_data_gcs_bucket_path: "add the gcs bucket path where you have the csv files of JOBSTIMELINE Data"
c. billing_data_gcs_bucket_path: "add the gcs bucket path where you have the csv files of Billing Data"
d. You can also enter the same gcs bucket path for JOBS, JOBSTIMELINE and BILLING Data
e. upload_project: "add the project id where you want to upload the data to bigquery for performing Health Check"
f. upload_dataset: "add the dataset id where you want to upload the data to bigquery for performing Health Check. If the dataset does not exist then a new dataset with this name will be created"
g. jobs_table_name: "add the table name which you want to keep for the table where you want to load the JOBS Data"
h. jobs_timeline_table_name: "add the table name which you want to keep for the table where you want to load the JOBS TIMELINE Data"
i. billing_table_name: "add the table name which you want to keep for the table where you want to load the BILLING Data"
j. credential: "replace this with path of the credential file created in previous step through terraform"

Note : Please use a double backslash when defining credential path in download_config.yaml file for Windows Machine

Example :

Windows :

credential: "C:\\Users\\user1\\path\to\\authentication\\key\\project-1.json"

Mac/Linux :

credential: "/path/to/authentication/key/project-1.json"

8. Install required python packages using : pip3 install -r requirements.txt
9. Run upload script with below command :

```
python3 /path/to/upload_metadata.py --config_file
/path/to/upload_config.yaml
```

10. Tables with JOBS, JOBS_TIMELINE and BILLING metadata will be created in the defined BigQuery dataset.
11. To destroy the roles and key created by terraform, run :

a. cd ../terraform/healthcheck/
b. terraform destroy --var-file=input.tfvars

12. Now, configure these tables in unravel and start the data hydration.

Steps for creating a gcs bucket :

1. Go to the Cloud Storage Buckets page: https://console.cloud.google.com/storage/browser in the Google Cloud console.
2. Click + Create.
3. On the Create a bucket page, enter the following information:
    a. Name your bucket: Enter a unique name for your bucket. The name must start with a lowercase letter or number, and it can contain up to 63 characters. It can also contain dashes and periods.
    b. Choose where to store your data: Select a location for your bucket. You can choose a location in the same region as your project, or you can choose a different region.
    c. Choose a storage class for your data: Select a storage class for your bucket. The storage class determines how long your data is kept and how much it costs to store.
4. Click Create.
5. Create new folder in bucket

1.5 Video(s) :

- https://drive.google.com/file/d/1-J2z6hGpSwGB0Dbv1o8jd5V4tDuilIWU/view?usp=sharing
- https://drive.google.com/file/d/13trMfsay4S-lanw5KqvqotDwhdsleDCs/view?usp=sharing