www.cloudskillsboost.google /focuses/49746

Using BigQuery Omni with AWS | Google Cloud Skills Boost

Qwiklabs : 14-17 minutes

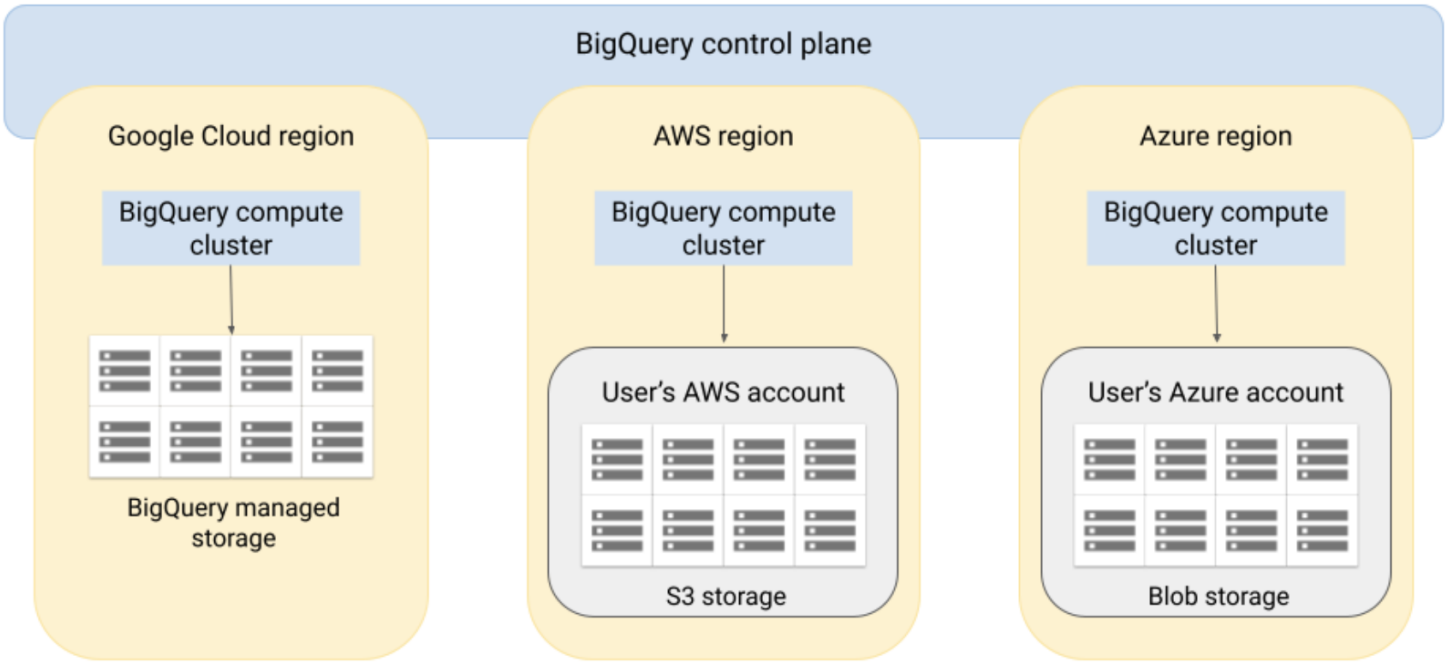
GSP1075

Google Cloud Self-Paced Labs

Overview

In this lab, you will learn how to use BigQuery Omni with AWS. BigQuery Omni lets you run BigQuery analytics on data stored in AWS S3. You will create an authorized connection between Google Cloud BigQuery and AWS S3, query data residing in S3 buckets without any data movement and write query results back to AWS S3 buckets. This lab is derived from a BigQuery Omni Guide published by Google.

High Level Architecture



Objectives

In this lab, you learn how to perform the following tasks:

- Create a connection between Google Cloud and AWS

- Authorize BigQuery Omni to read data in an AWS S3 bucket
- Create a BigQuery external table that references the raw data in AWS S3 bucket
- Run queries on AWS S3 data
- Export query results to AWS S3 bucket

Setup and requirements

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).

Note: Use an Incognito or private browser window to run this lab. This prevents any conflicts between your personal account and the Student account, which may cause extra charges incurred to your personal account.

- Time to complete the lab---remember, once you start, you cannot pause a lab.

Note: If you already have your own personal Google Cloud account or project, do not use it for this lab to avoid extra charges to your account.

How to start your lab and sign in to the Google Cloud Console

1. Click the **Start Lab** button. If you need to pay for the lab, a pop-up opens for you to select your payment method. On the left is the **Lab Details** panel with the following:
    - The **Open Google Console** button
    - Time remaining
    - The temporary credentials that you must use for this lab
    - Other information, if needed, to step through this lab
2. Click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.

    Tip: Arrange the tabs in separate windows, side-by-side.

    Note: If you see the **Choose an account** dialog, click **Use Another Account**.

3. If necessary, copy the **Username** from the **Lab Details** panel and paste it into the **Sign in** dialog. Click **Next**.

4. Copy the **Password** from the **Lab Details** panel and paste it into the **Welcome** dialog. Click **Next**.

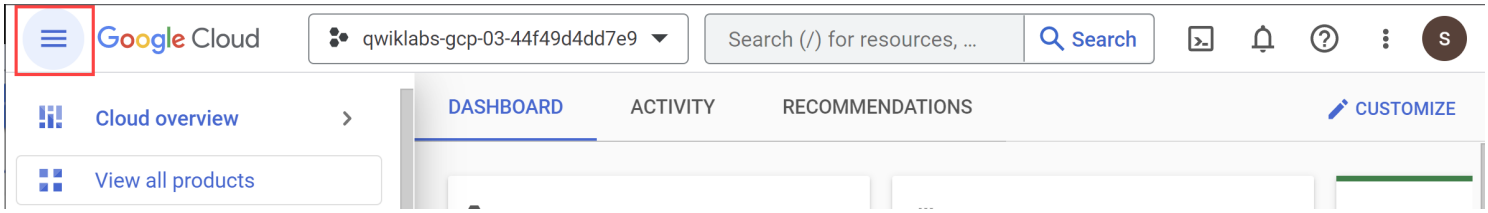
   Important: You must use the credentials from the left panel. Do not use your Google Cloud Skills Boost credentials. **Note:** Using your own Google Cloud account for this lab may incur extra charges.

5. Click through the subsequent pages:

   - Accept the terms and conditions.
   - Do not add recovery options or two-factor authentication (because this is a temporary account).
   - Do not sign up for free trials.

After a few moments, the Cloud Console opens in this tab.

Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left.



Task 1. Create a BigQuery AWS connection

BigQuery Omni accesses Amazon S3 data through authorized connections from Google Cloud. Each connection has its own unique Amazon Web Services (AWS) Identity and Access Management (IAM) user. You grant permissions to users through AWS IAM roles. The policies within the AWS IAM roles determine what data BigQuery can access for each connection.

Note: In this lab, you will be using both Google Cloud and AWS consoles, and switch between the two tabs/windows based on task descriptions.

Create an AWS IAM policy for BigQuery

1. Sign in to the AWS Management Console. Click the **Open AWS Console** button on the lab pane, and log in with the provided username and password.

2. Search for Amazon S3 in the Search bar at the top and select **S3**. A regional bucket named with data pre-populated is already available for this lab. **Copy** this bucket name for subsequent steps.

3. Search for AWS Identity and Access Management (IAM) in the Search bar at the top and select **IAM**.

4. From the left pane select **Policies** and click **bigquery-omni-connection-policy**.

5. Click **Edit policy > JSON** and paste the following into the editor. Replace all instances of `<BUCKET_NAME>` with your S3 bucket name copied from Step 2.

{ "Statement": [{ "Action": ["s3:ListBucket"], "Effect": "Allow", "Resource": ["arn:aws:s3:::<BUCKET_NAME>"], "Sid": "BucketLevelAccess" }, { "Action": ["s3:GetObject"], "Effect": "Allow", "Resource": ["arn:aws:s3:::<BUCKET_NAME>", "arn:aws:s3:::<BUCKET_NAME>/*"], "Sid": "ObjectLevelGetAccess" }, { "Action": ["s3:PutObject"], "Effect": "Allow", "Resource": ["arn:aws:s3:::<BUCKET_NAME>", "arn:aws:s3:::<BUCKET_NAME>/*"], "Sid": "ObjectLevelPutAccess" }], "Version": "2012-10-17" }

6. Click **Next**.

7. Click **Save changes**.

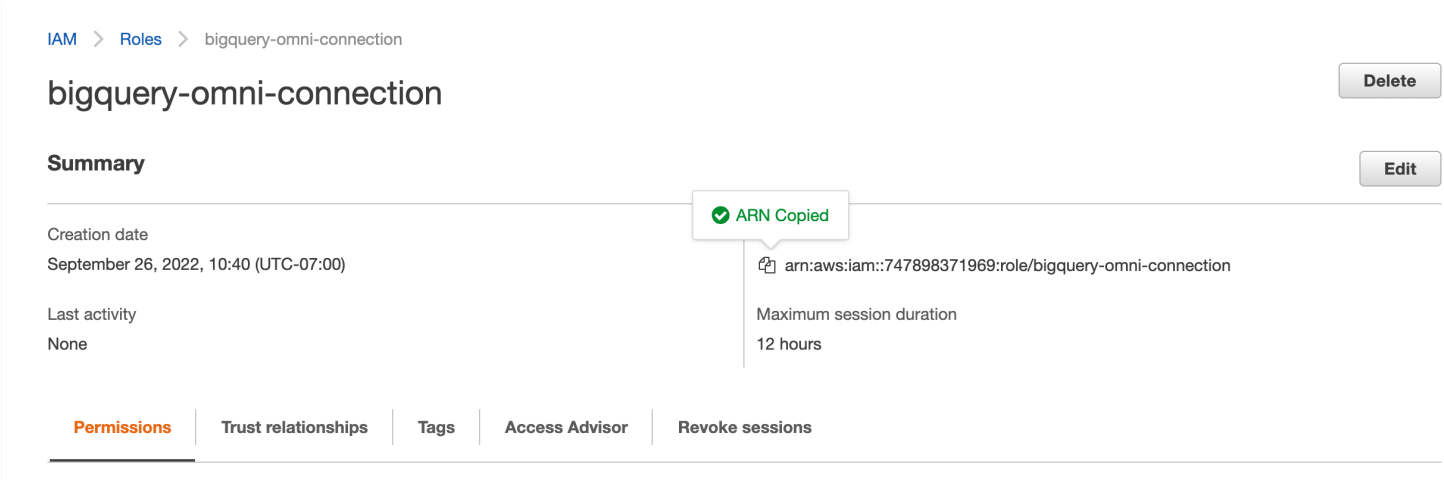
Validate the AWS IAM for BigQuery

1. From the left pane, select **Roles**.

2. Click the **bigquery-omni-connection** role.

3. Copy the **Role ARN**, it will be in the following format, where `<ACCOUNT_ID>` is your AWS Account ID.

arn:aws:iam::<ACCOUNT_ID>:role/bigquery-omni-connection



Create the BigQuery AWS connection

1. In the Google Cloud Console, from the **Navigation Menu**, go to **BigQuery** > **SQL Workspace**.

2. Click **+ADD**, then select **Connections to external data sources**.

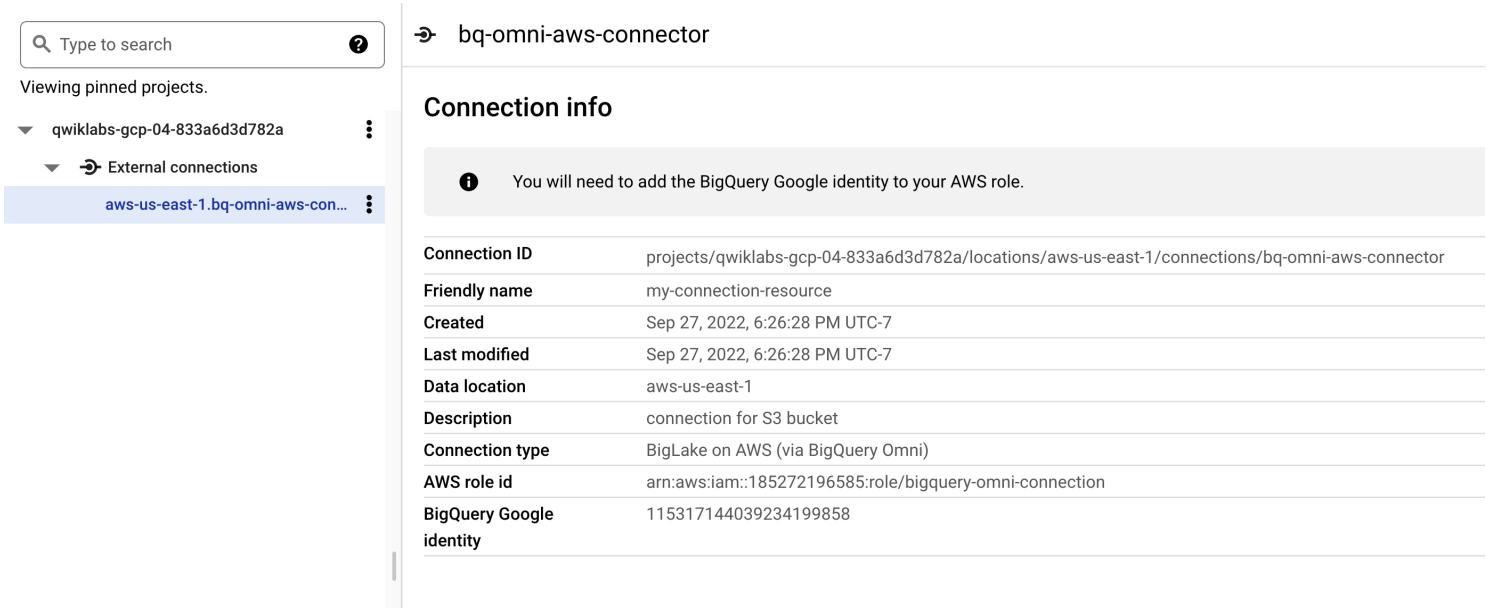
3. In the **External data source** pane, enter the following information:

    - For **Connection type**, select **BigLake on AWS (via BigQuery Omni)**.
    - For **Connection ID**, type `bq-omni-aws-connector` for an identifier for the connection resource.
    - For **Connection location**, select `aws-us-east-1`.
    - `Optional:` For **Friendly name**, enter a user-friendly name for the connection. The friendly name can be any value that helps you identify the connection resource if you need to modify it later.
    - `Optional:` For **Description**, enter a description for this connection resource.

- For **AWS role id**, enter the full IAM Role ARN that you copied in the previous step in this format: `arn:aws:iam::AWS_ACCOUNT_ID:role/ROLE_NAME`

4. Click **Create connection**.

5. In the BigQuery Explorer, click the dropdown next to your project name and navigate to the newly created connection in the **External Connections** list.



6. Note the **BigQuery Google identity**. This is a Google principle that is specific to each connection. Copy this BigQuery Google identity, it will be used in the next section.

Your BigQuery Google Identity should resemble the following:

BigQuery Google identity: 114999259451445753095

Click **Check my progress** to verify the objective. Create the BigQuery AWS connection

Add a Trust Relationship to the AWS role

The trust relationship lets the BigQuery AWS connection assume the role and access the S3 data as specified in the roles policy.

1. Navigate back to the **AWS IAM console**.

2. From the left pane, select **Roles**.

3. Select the **bigquery-omni-connection** role.

4. Click **Edit** and then do the following:

    - Verify if **Maximum session duration** is set to **12 hours**. As each query can run for up to six hours, this duration allows for one additional retry. Increasing the session duration beyond 12 hours will

not allow for additional retries. For more information, see the query/multi-statement query execution-time limit.
    - Click **Save changes**.

5. Select **Trust Relationships** tab and click **Edit trust policy**.

6. Replace the policy content with the following, replacing `<00000>` with the **BigQuery Google identity** you copied in the previous section.

{ "Version": "2012-10-17", "Statement": [{ "Effect": "Allow", "Principal": { "Federated": "accounts.google.com" }, "Action": "sts:AssumeRoleWithWebIdentity", "Condition": { "StringEquals": { "accounts.google.com:sub": "00000" } } }] }

7. Click **Update Policy**.

The connection is now ready to use.

Note: There may be a propagation delay for role assignment in AWS. If you receive an error of this type when using a new connection, waiting and trying again later may resolve the issue.

Task 2. Run queries on the AWS S3 external table

BigQuery Omni does not manage data stored in Amazon S3. To access S3 data, define an external table. This table is called an external table because the data is not stored in BigQuery managed storage. For more information about external tables, see External tables.

Create a BigQuery dataset

In this section, you will create a BigQuery dataset in the same region as your AWS S3 bucket.

1. In the Google Cloud Console, go to the BigQuery page to create a dataset.

2. Click the 3 dots next to your project name and select **Create dataset**.

3. On the **Create dataset** page, enter the following information:

    - For **Dataset ID**, enter `bq_omni_demo`.
    - For **Location type**, select `Region`.
    - For **Data location**, choose `aws-us-east-1`.

4. Click **Create dataset**.

Create an external table

In this section you will create an external table in the above dataset.

1. In this BigQuery explorer, expand your project and select the `bq_omni_demo` dataset created.

2. In the details panel, click **Create table**.

3. On the **Create table** page, in the **Source** section, do the following:

    - For **Create table from**, select **Amazon S3**.
    - For **Select S3 path**, enter `s3://[S3 bucket name]/taxi-data_green_trips_table.csv`.
        - Replace `[S3 bucket name]` with .
    - For **File format**, select **CSV**. **Note:** supported formats are AVRO, PARQUET, ORC, CSV, NEWLINE_DELIMITED_JSON, and Google Sheets.

Important: make sure to remove any spaces in the S3 path as this will cause errors!

4. On the **Create table** page, in the **Destination** section, do the following:

    - For **Dataset name**, choose `bq_omni_demo`.
    - In the **Table name** field, use `bq-omni-table`.
    - Verify that **Table type** is set to **External table**.
    - For **Connection ID**, choose the appropriate Connection ID from the dropdown.
    - In the **Schema** section, select the **Auto detect** checkbox.

5. Click **Create table.**

Note: You do not need to specify table schema since it is autodetected from the source file.

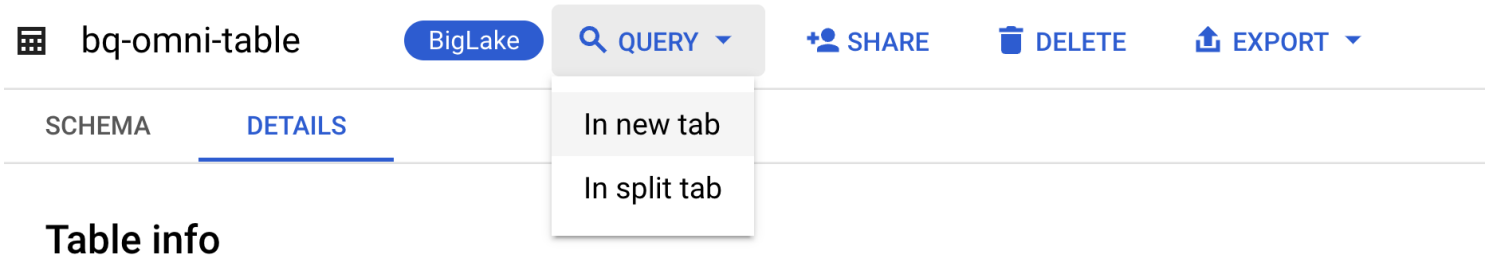
Click **Check my progress** to verify the objective. Create the BigQuery Dataset and External Table

Task 3. Create an external table and query AWS S3 data

BigQuery Omni lets you query the external table like any BigQuery table. The maximum result size for interactive queries is 10 GB (preview). For more information, see Limitations. If your query result is larger than 10 GB, then we recommend that you export it to Amazon S3. The query result is stored in a BigQuery temporary table.

Query the external table

1. From the **bq-omni-table** details page, select **Query > In new tab**.



2. In the Query editor, execute the following statement:

```
SELECT * FROM `{{{project_0.project_id|S3 bucket name}}}.bq_omni_demo.bq-omni-table`
```

3. Click **Run**.

You should see the following output:

Query results

SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

Row	Vendor_Id	Pickup_DateTime	Dropoff_DateTime	St...	Rate_Code_Id	PULocationID	DOLocationID	Passenger_...	Trip_Distance	Fare_
1	1	2021-01-27 09:55:10 UTC	2021-01-27 10:01:54 UTC	fal...	1	152	41	1	1.1	
2	2	2021-01-03 08:11:16 UTC	2021-01-03 08:30:06 UTC	fal...	1	95	82	1	3.89	
3	2	2021-01-26 16:11:00 UTC	2021-01-26 16:24:00 UTC	null	null	98	16	null	2.46	
4	2	2021-01-24 11:12:00 UTC	2021-01-24 11:21:00 UTC	null	null	57	92	null	4.35	
5	2	2021-01-26 09:54:00 UTC	2021-01-26 10:06:00 UTC	null	null	215	121	null	2.5	
6	2	2021-01-17 14:26:00 UTC	2021-01-17 14:38:00 UTC	null	null	16	171	null	2.59	
7	2	2021-01-08 20:21:46 UTC	2021-01-08 20:37:22 UTC	fal...	1	116	246	3	6.54	
8	2	2021-01-13 11:54:44 UTC	2021-01-13 12:15:47 UTC	fal...	1	65	225	1	2.68	
9	1	2021-01-07 12:09:25 UTC	2021-01-07 12:23:24 UTC	fal...	1	244	47	1	3.2	
10	2	2021-01-22 21:34:43 UTC	2021-01-22 21:40:57 UTC	fal...	1	116	152	1	1.17	
11	2	2021-01-07 10:35:00 UTC	2021-01-07 11:08:00 UTC	null	null	14	77	null	17.47	
12	2	2021-01-12 10:29:00 UTC	2021-01-12 11:05:00 UTC	null	null	42	232	null	9.29	
13	2	2021-01-10 14:43:11 UTC	2021-01-10 15:09:36 UTC	fal...	1	226	218	1	13.73	
14	2	2021-01-29 19:51:29 UTC	2021-01-29 19:52:18 UTC	fal...	5	130	130	1	0.0	

Click **Check my progress** to verify the objective. Query the External Table

Task 4. Export query results to AWS S3

BigQuery Omni lets you export the result of a query against a BigQuery external table to Amazon S3.

Export Query Results

BigQuery Omni writes to the specified Amazon S3 location irrespective of any existing content. The export query can overwrite existing data or mix the query result with existing data. In the Query editor field, you will need to run a Google Standard SQL export query. Google Standard SQL is the default syntax in the Google Cloud console. The following is the template for what you will need to write:

EXPORT DATA WITH CONNECTION `CONNECTION_REGION.CONNECTION_NAME` \
OPTIONS(uri="s3://BUCKET_NAME/PATH", format="FORMAT", ...) \ AS QUERY

You will need to replace the following:

- `CONNECTION_REGION`: the region where the connection was created.
- `CONNECTION_NAME`: the connection name that you created with the necessary permission to write to the S3 bucket.
- `BUCKET_NAME`: the Amazon S3 bucket where you want to write the data.
- `PATH`: the path where you want to write the exported file to
- `FORMAT`: supported formats are JSON, AVRO, and CSV.
- `QUERY`: the query to analyze the data that is stored in a BigQuery external table.

1. For this lab, the query has been pre-populated for you. Paste this query into the editor:

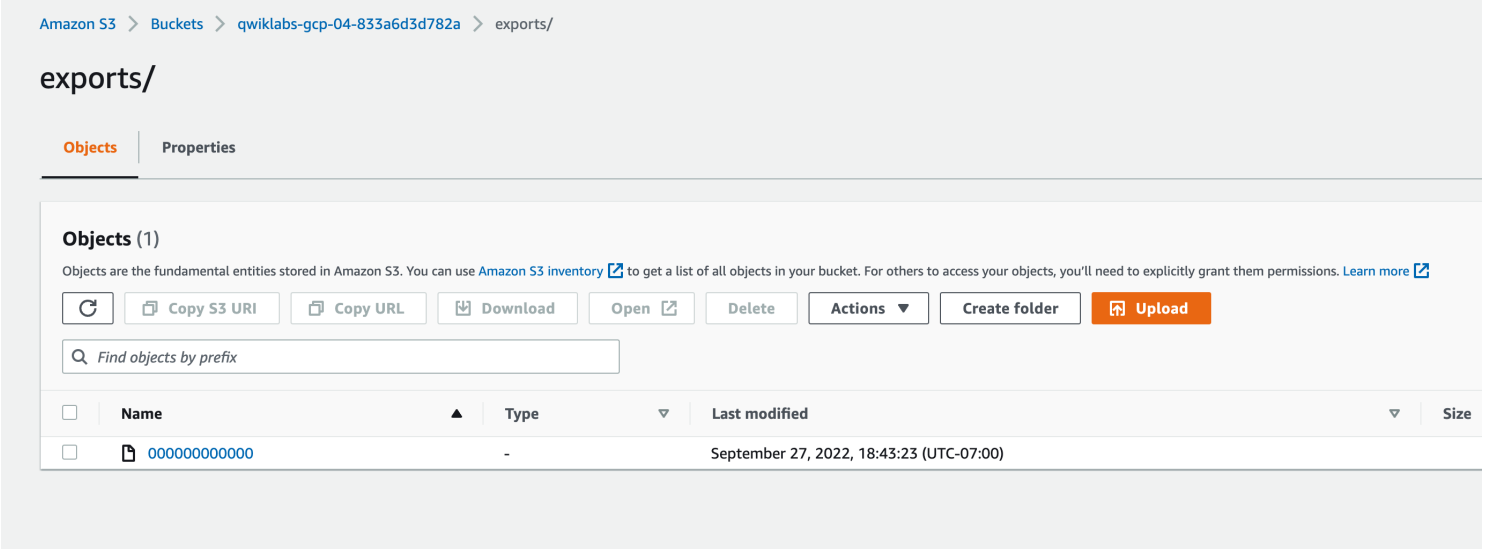
EXPORT DATA WITH CONNECTION `aws-us-east-1.bq-omni-aws-connector`
OPTIONS(uri="s3://{{{project_0.project_id|S3 bucket name}}}/exports/*", format="CSV") AS SELECT * FROM
`{{{project_0.project_id|S3 bucket name}}}.bq_omni_demo.bq-omni-table`

2. Click **Run**.

You should see the following output:

Successfully exported 15417 rows into 1 files.

3. Navigate to your S3 bucket and verify the data has been exported in the `exports` directory.



Great! You have successfully executed an export query and created a file in your S3 bucket.

Click **Check my progress** to verify the objective. Export query results to AWS S3

Congratulations!

In this lab you created a connection between Google Cloud and AWS and authorized BigQuery Omni to read the data in an AWS S3 bucket. You then created a BigQuery external table that references the raw data in the AWS S3 bucket, ran queries on the data, and exported query results back to an AWS S3 bucket.

Next steps / learn more

Be sure to check out the following to receive more information about BigQuery Omni:

- BigQuery Omni Overview
- BigQuery Omni Pricing
- Setting up VPC Service Controls for BigQuery Omni
- BigQuery Omni for cross cloud data analytics

Google Cloud training and certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.

Manual Last Updated June 2, 2023

Lab Last Tested June 2, 2023