Datastream MySQL to BigQuery

cloudskillsboost.google/games/2880/labs/17405

GSP840

Google Cloud Self-Paced Labs

Overview

Datastream is a serverless and easy-to-use Change Data Capture (CDC) and replication service that allows you to synchronize data across heterogeneous databases, storage systems, and applications reliably and with minimal latency. In this lab you'll learn how to replicate data from your OLTP workloads into BigQuery, in real time.

You will begin by deploying MySQL on Cloud SQL and import a dataset using the `gcloud` command line. Then, in the Cloud Console UI, you will create and start a Datastream stream and a Dataflow job for replication. The replication uses a Dataflow template to enable continuous replication of data, along with Cloud Storage and Pub/Sub for buffering data.

Although you can easily copy and paste commands from the lab to the appropriate place, students should type the commands themselves to reinforce their understanding of the core concepts.

What you'll do

- Prepare a MySQL Cloud SQL instance using the Google Cloud Console
- Create a GCS bucket to be used in replication
- Create a Pub/Sub topic, subscription, and a GCS Pub/Sub notification policy
- Import data into the Cloud SQL instance
- Create a Datastream connection profile referencing the MySQL DB
- Create a Datastream connection profile referencing the GCS destination
- Create a Pub/Sub resources and a GCS Pub/Sub notification policy
- Create a Datastream stream and start replication
- Create a BigQuery dataset
- Deploy a Dataflow job to replicate data

Prerequisites

- Familiarity with standard Linux environments
- Familiarity with change data capture (CDC) concepts

Setup

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**.

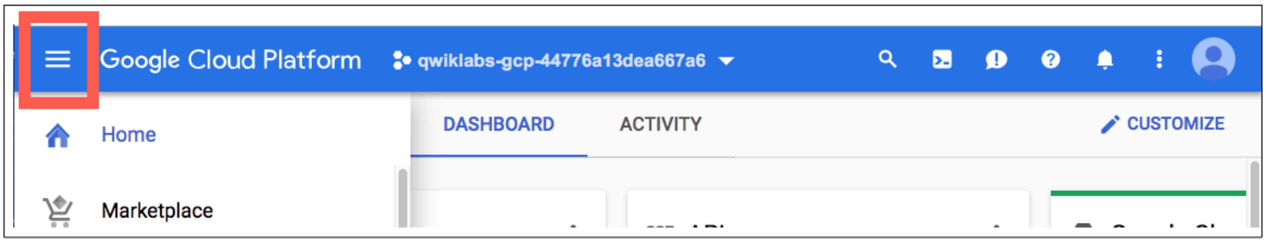
4. Copy the **Password** from the **Lab Details** panel and paste it into the **Welcome** dialog. Click **Next**.

   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.

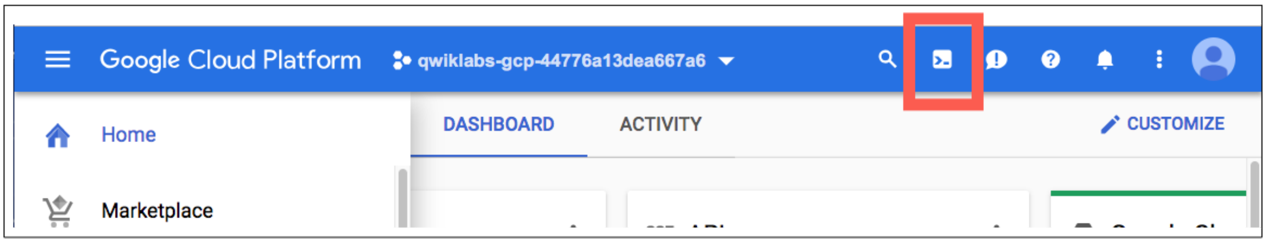
Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left.



Activate Cloud Shell

Cloud Shell is a virtual machine that is loaded with development tools. It offers a persistent 5GB home directory and runs on the Google Cloud. Cloud Shell provides command-line access to your Google Cloud resources.

1. In the Cloud Console, in the top right toolbar, click the **Activate Cloud Shell** button.



2. Click **Continue**.

It takes a few moments to provision and connect to the environment. When you are connected, you are already authenticated, and the project is set to your **PROJECT_ID**. The output contains a line that declares the **PROJECT_ID** for this session:

Your Cloud Platform project in this session is set to YOUR_PROJECT_ID
`gcloud` is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab-completion.

3. (Optional) You can list the active account name with this command:

```
gcloud auth list
```

(Output)

```
ACTIVE: * ACCOUNT: student-01-xxxxxxxxxxxx@qwiklabs.net To set the active account,
run: $ gcloud config set account `ACCOUNT`
```

4. (Optional) You can list the project ID with this command:

```
gcloud config list project
```

(Output)

```
[core] project = <project_ID>
```

(Example output)

```
[core] project = qwiklabs-gcp-44776a13dea667a6
```

For full documentation of `gcloud`, in Google Cloud, Cloud SDK documentation, see the gcloud command-line tool overview.
Set your Project ID variable:

```
PROJECT_ID=<Replace with your Project ID>
```

Create a Cloud SQL database instance

Run the following command in Cloud Shell to create a Cloud SQL for MySQL database instance:

```
MYSQL_INSTANCE=mysql-db
DATASTREAM_IPS=34.72.28.29,34.67.234.134,34.67.6.157,34.72.239.218,34.71.242.81
gcloud sql instances create ${MYSQL_INSTANCE} \ --cpu=2 --memory=10GB \ --
authorized-networks=${DATASTREAM_IPS} \ --enable-bin-log \ --region=us-central1 \
--database-version=MYSQL_8_0 \ --root-password password123
```

This script creates the database in `us-central1`. For other regions, be sure to replace the IPs below with the right Datastream Public IPs for your region.

Once the database instance is created, make a note of the instance's public IP - you'll need this later when creating Datastream's connection profile.

Click *Check my progress* to verify the objective. Create a Cloud SQL database instance

Create a Cloud Storage bucket

Run the following to create a Cloud Storage bucket with the same name as your project:

```
gsutil mb gs://${PROJECT_ID}
```

Create Pub/Sub resources

Run the following to create Pub/Sub resources:

```
gcloud pubsub topics create datastream gcloud pubsub subscriptions create datastream-
subscription --topic=datastream gsutil notification create -f "json" -p "data/" -t
"datastream" "gs://${PROJECT_ID}"
```

Click *Check my progress* to verify the objective. Create cloud storage bucket and Pub/Sub resources

Import a SQL file into MySQL

Open a file named **create_mysql.sql** in vim or your favorite editor, then copy the text below into your file:

```
CREATE DATABASE IF NOT EXISTS test; USE test; CREATE TABLE IF NOT EXISTS
test.example_table ( id INT NOT NULL AUTO_INCREMENT PRIMARY KEY, text_col
VARCHAR(50), int_col INT, created_at TIMESTAMP ); INSERT INTO
test.example_table (text_col, int_col, created_at) VALUES ('hello', 0, '2020-01-01
00:00:00'), ('goodbye', 1, NULL), ('name', -987, NOW()), ('other', 2786, '2021-01-01
00:00:00');
```

Next, copy this file into the Cloud Storage bucket you created above (make sure you do not load the file into the `data/` directory), make the file accessible to your Cloud SQL service account, and import the SQL command into your database:

```
SERVICE_ACCOUNT=$(gcloud sql instances describe ${MYSQL_INSTANCE} | grep
serviceAccountEmailAddress | awk '{print $2;}') gsutil cp create_mysql.sql
gs://${PROJECT_ID}/resources/create_mysql.sql gsutil iam ch
serviceAccount:${SERVICE_ACCOUNT}:objectViewer gs://${PROJECT_ID} gcloud sql
import sql ${MYSQL_INSTANCE} gs://${PROJECT_ID}/resources/create_mysql.sql --
quiet
```

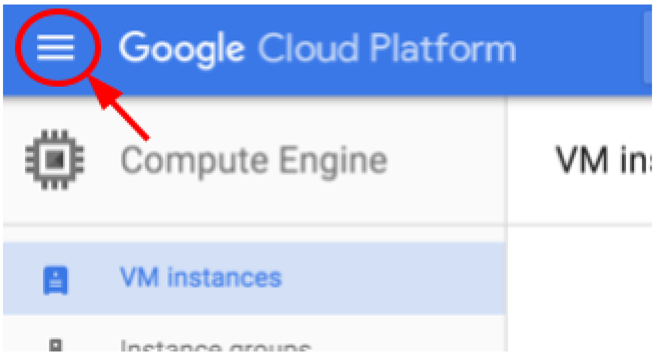
Click *Check my progress* to verify the objective. Import a SQL file into MySQL

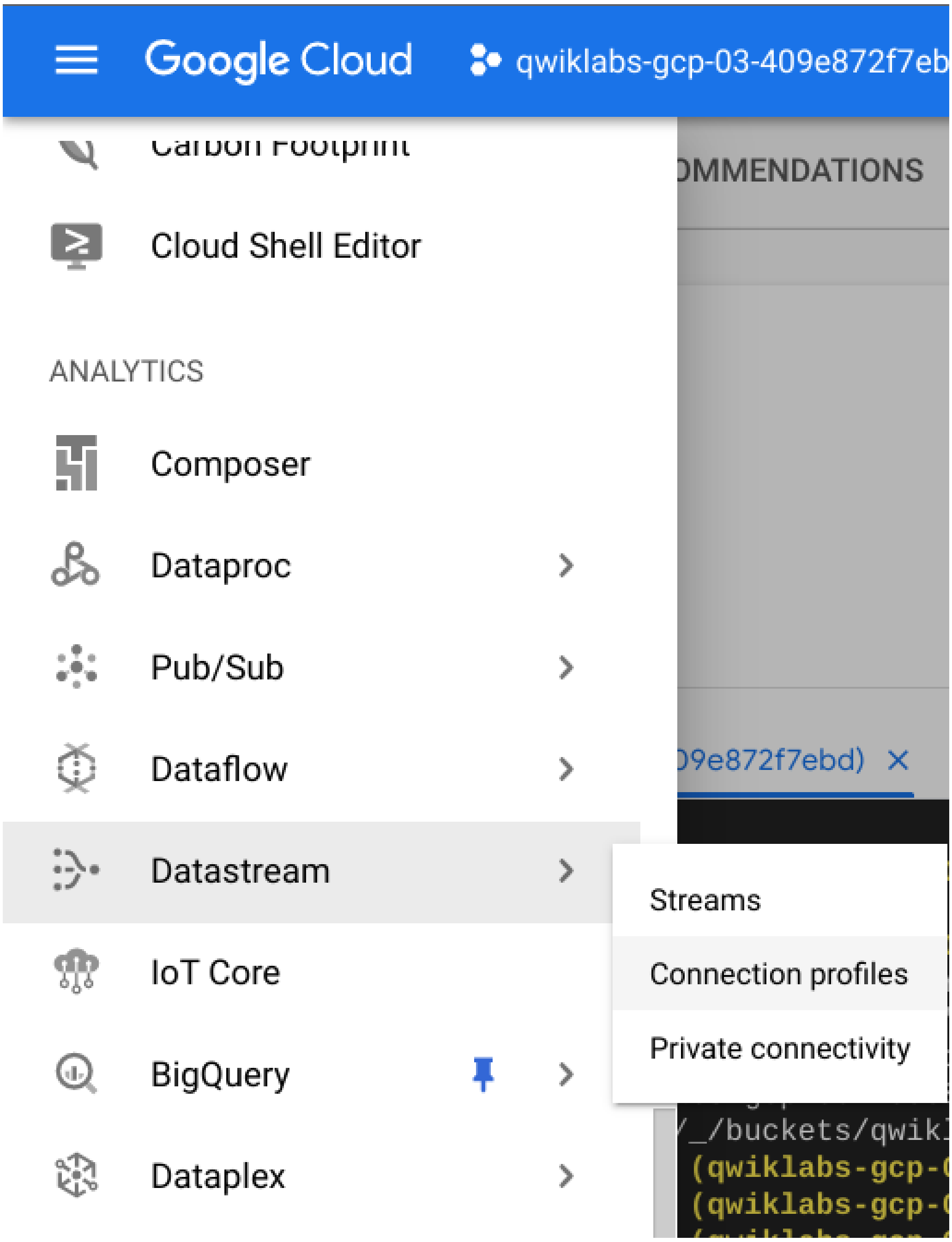
Create Datastream resources

Now that all the initial resources are deployed, create the Datastream connection profiles and stream to begin replication.

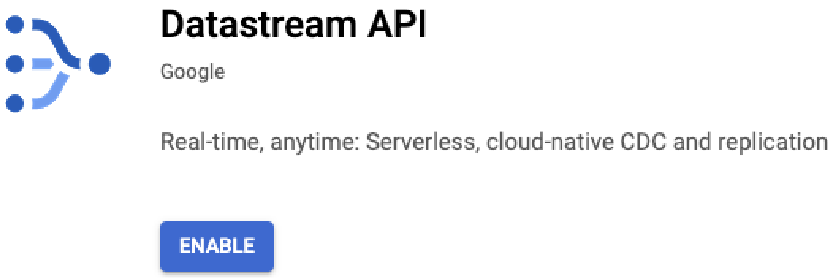
In the Cloud Console, click the **Navigation menu** icon on the top left of the screen:

Then navigate to **Analytics > Datastream > Connection Profiles**





Click **ENABLE** to enable the Datastream API.



Datastream API

Google

Real-time, anytime: Serverless, cloud-native CDC and replication

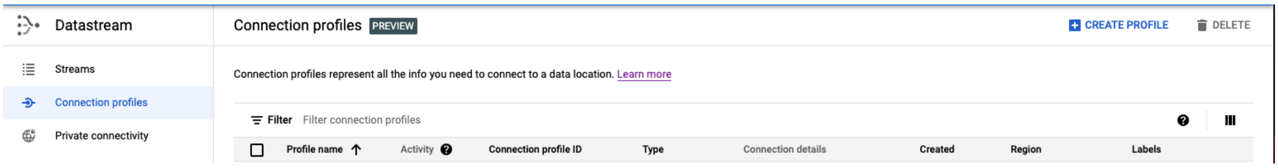
Click *Check my progress* to verify the objective. Enable Datastream API

Create Connection Profiles

Create two connection profiles, one for the MySQL source, and another for the Cloud Storage destination.

MySQL connection profile

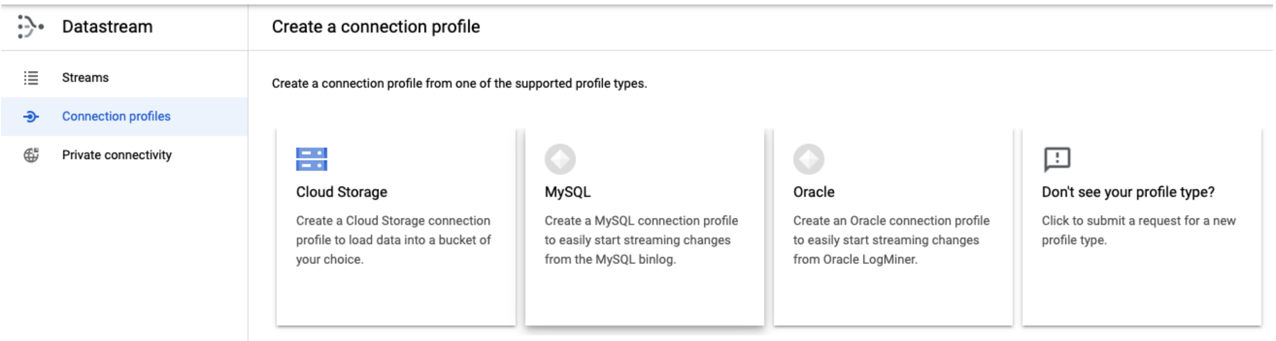
In the Cloud Console, navigate to the **Connection Profiles** tab and click **CREATE PROFILE**.



Connection profiles represent all the info you need to connect to a data location. Learn more

□	Profile name ↑	Activity	Connection profile ID	Type	Connection details	Created	Region	Labels

Select MySQL **Connection profile** type.



Create a connection profile

Create a connection profile from one of the supported profile types.

Cloud Storage

Create a Cloud Storage connection profile to load data into a bucket of your choice.

MySQL

Create a MySQL connection profile to easily start streaming changes from the MySQL binlog.

Oracle

Create an Oracle connection profile to easily start streaming changes from Oracle LogMiner.

Don't see your profile type?

Click to submit a request for a new profile type.

Use `mysql-cp` as the name and ID for your connection profile.

Enter the database connection details:

- The IP and port of the Cloud SQL for MySQL instance created earlier
- Username: `root` , password: `password123`

Click **CONTINUE**.

Leave the encryption as NONE. Click **CONTINUE**.

Select the **IP allowlisting** connectivity method, and click **CONTINUE**.

Click **RUN TEST** to make sure that Datastream is able to reach the database.

Click **CREATE**.

Cloud Storage connection profile

In the Cloud Console, navigate to the **Connection Profiles** tab and click **CREATE PROFILE**.

- Select **Cloud Storage** connection profile type.
- Use `gcs-cp` as the name and ID for your connection profile.
- Choose the bucket created earlier, and enter `/data` as the **connection profile path prefix**.

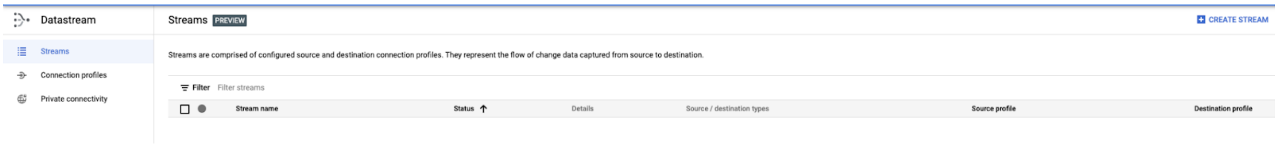
Click **CREATE**.

Click *Check my progress* to verify the objective. Create Connection Profiles

Create the stream

Create the stream which connects the connection profiles created above and defines the configuration for the data to stream from source to destination.

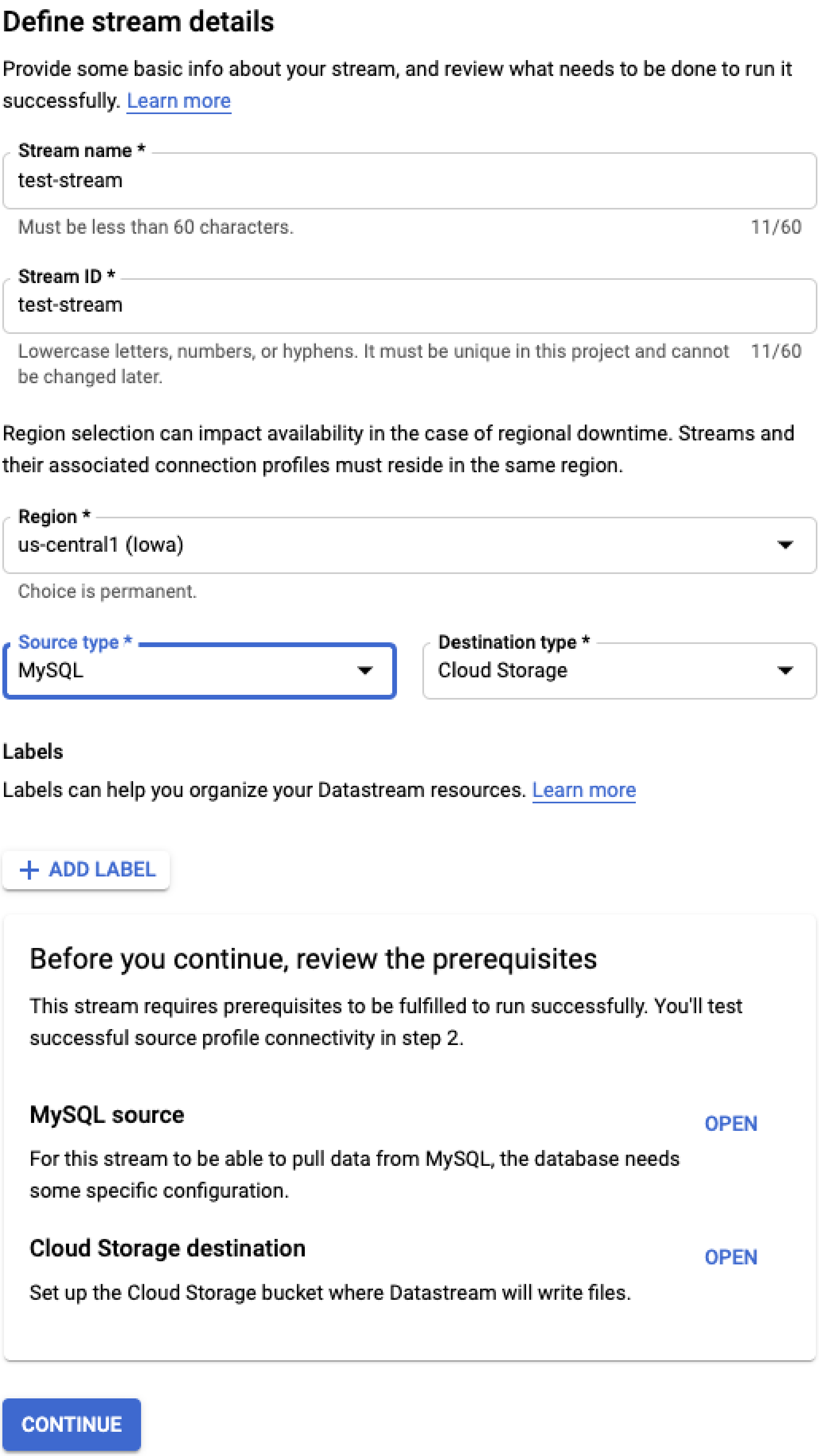
In the Cloud Console, navigate to the **Streams** tab and click **CREATE STREAM**.



Streams are comprised of configured source and destination connection profiles. They represent the flow of change data captured from source to destination.

□	●	Stream name	Status ↑	Details	Source / destination types	Source profile	Destination profile

- Use `test-stream` as the name and ID for your stream.
- Select **MySQL** as the source.
- Click **CONTINUE**.



Define stream details

Provide some basic info about your stream, and review what needs to be done to run it successfully. Learn more

Stream name *
test-stream

Must be less than 60 characters. 11/60

Stream ID *
test-stream

Lowercase letters, numbers, or hyphens. It must be unique in this project and cannot be changed later. 11/60

Region selection can impact availability in the case of regional downtime. Streams and their associated connection profiles must reside in the same region.

Region *
us-central1 (Iowa)

Choice is permanent.

Source type *
MySQL

Destination type *
Cloud Storage

Labels

Labels can help you organize your Datastream resources. Learn more

Before you continue, review the prerequisites

This stream requires prerequisites to be fulfilled to run successfully. You'll test successful source profile connectivity in step 2.

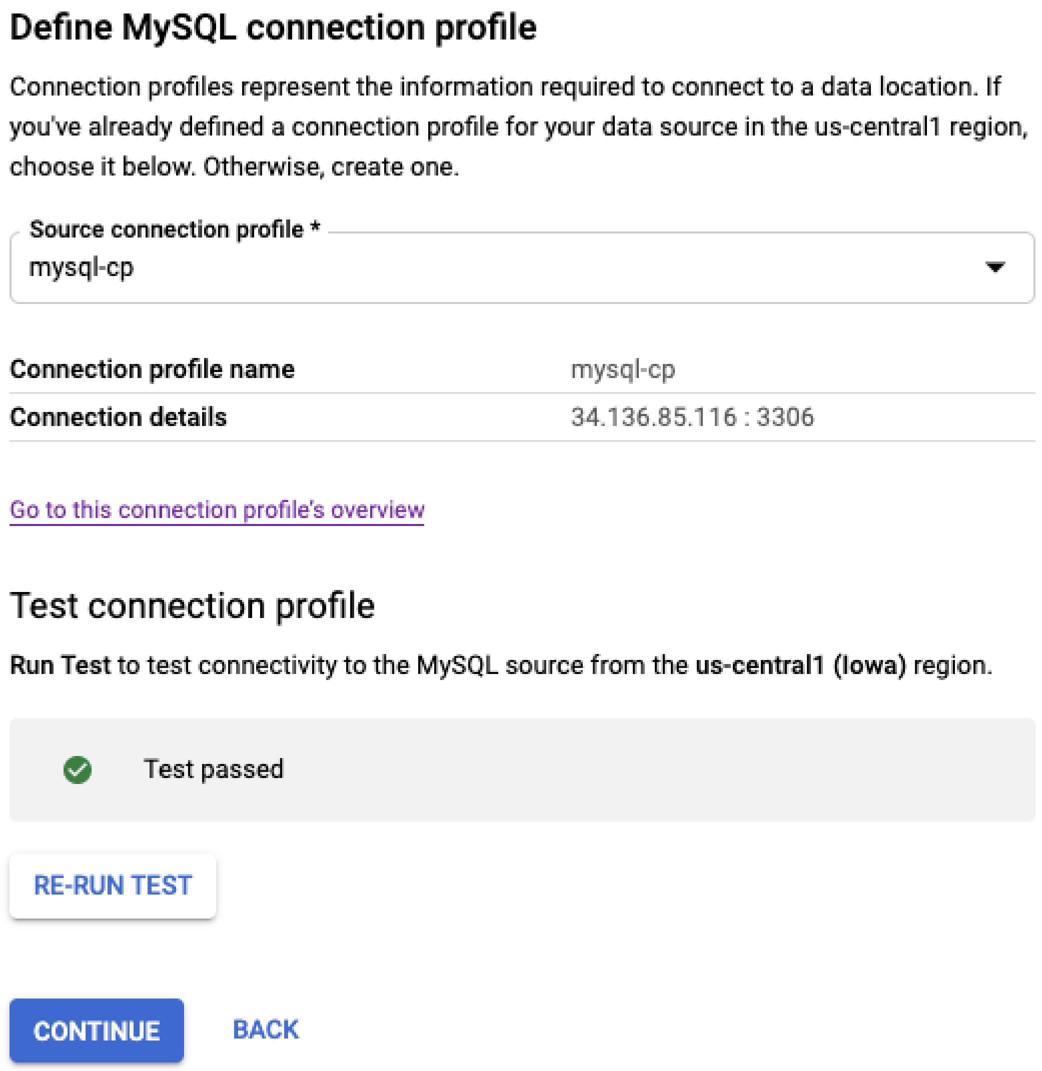
MySQL source

For this stream to be able to pull data from MySQL, the database needs some specific configuration.

Cloud Storage destination

Set up the Cloud Storage bucket where Datastream will write files.

Select the **mysql-cp** created in the previous step. You can test connectivity by clicking **RUN TEST**, then click **Continue** once the test passes.



Define MySQL connection profile

Connection profiles represent the information required to connect to a data location. If you've already defined a connection profile for your data source in the us-central1 region, choose it below. Otherwise, create one.

Source connection profile *
mysql-cp

Connection profile name	mysql-cp
Connection details	34.136.85.116 : 3306

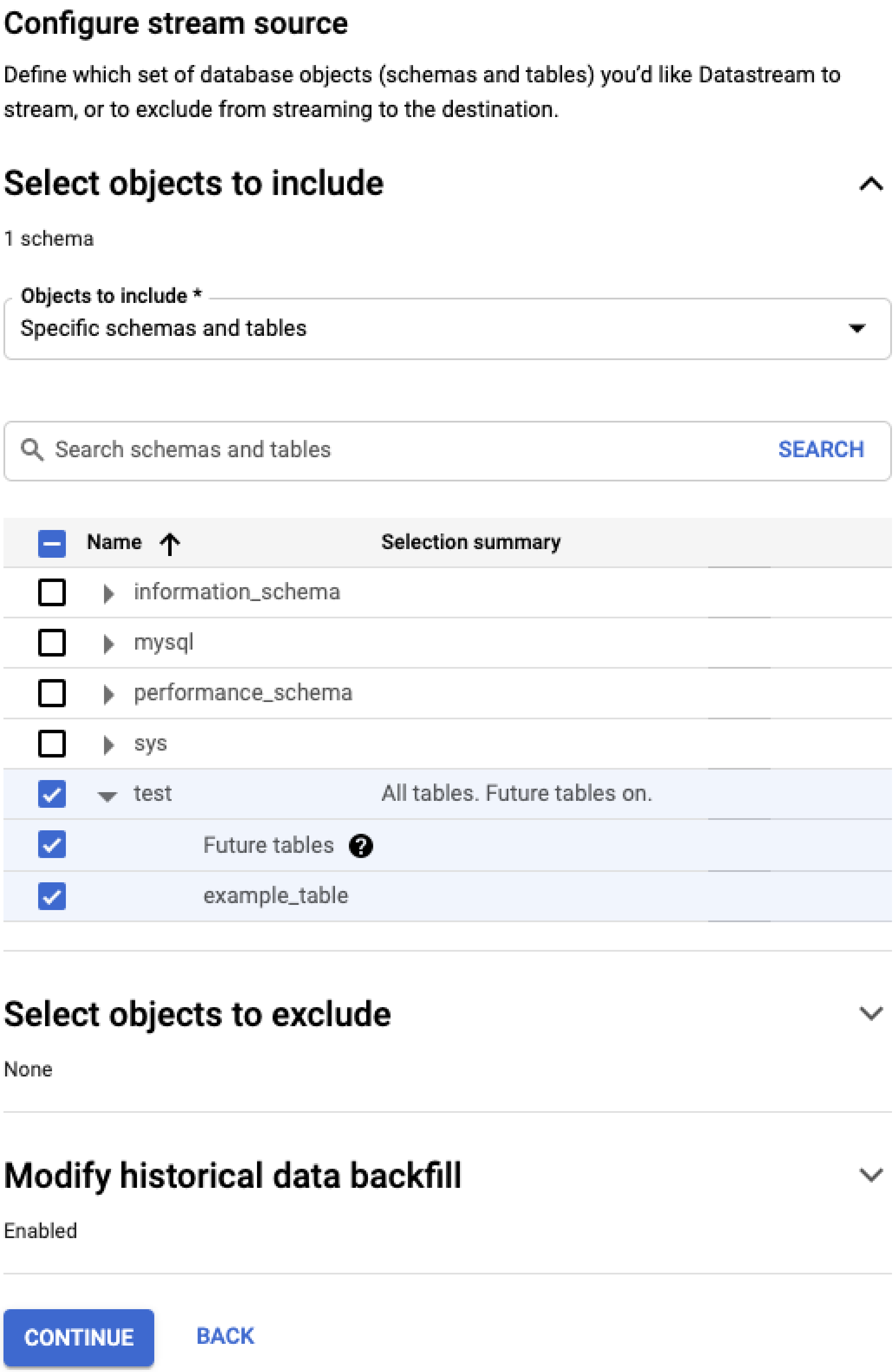
Go to this connection profile's overview

Test connection profile

Run Test to test connectivity to the MySQL source from the **us-central1 (Iowa)** region.

✓ Test passed

Mark the tables you want to replicate - for this lab, only replicate the test database, then click **CONTINUE**.



Configure stream source

Define which set of database objects (schemas and tables) you'd like Datastream to stream, or to exclude from streaming to the destination.

Select objects to include

1 schema

Objects to include *
Specific schemas and tables

Search schemas and tables SEARCH

☑	Name ↑	Selection summary
□	▸ information_schema	
□	▸ mysql	
□	▸ performance_schema	
□	▸ sys	
☑	▾ test	All tables. Future tables on.
☑	Future tables ?	
☑	example_table	

Select objects to exclude

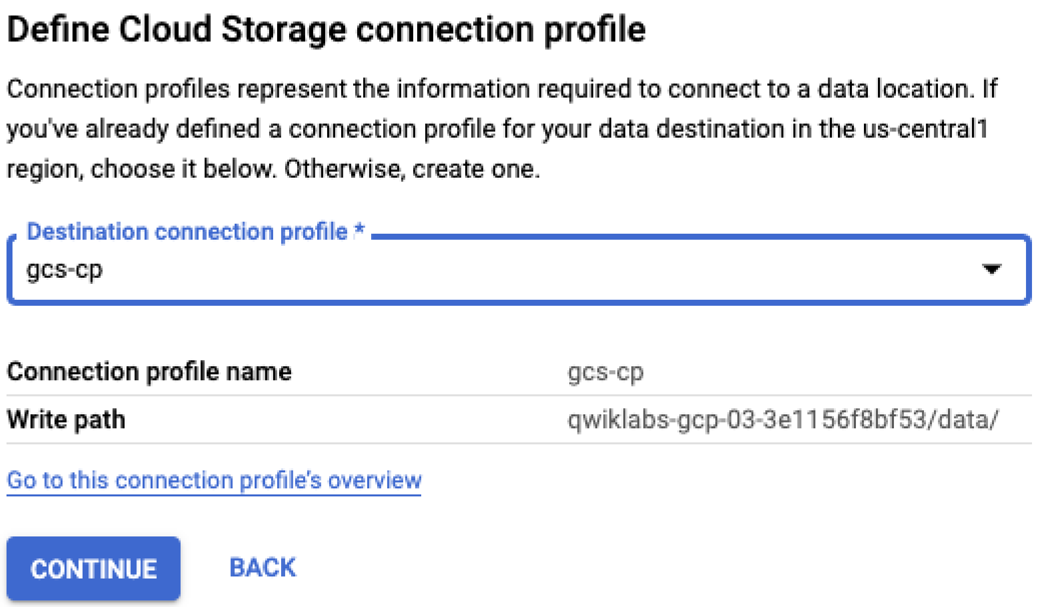
None

Modify historical data backfill

Enabled

CONTINUE BACK

Select the Cloud Storage bucket you created in the previous step, then click **CONTINUE**.



Define Cloud Storage connection profile

Connection profiles represent the information required to connect to a data location. If you've already defined a connection profile for your data destination in the us-central1 region, choose it below. Otherwise, create one.

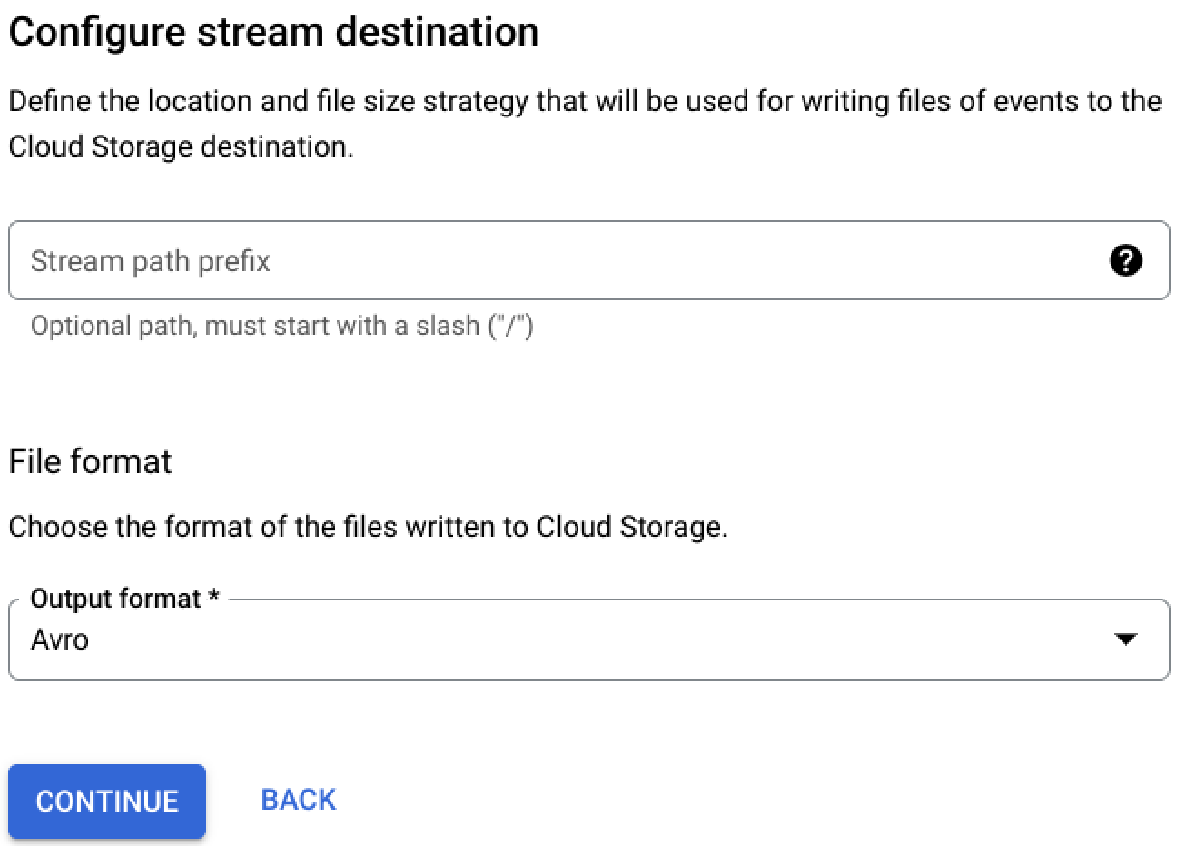
Destination connection profile *: gcs-cp

Connection profile name	gcs-cp
Write path	qwiklabs-gcp-03-3e1156f8bf53/data/

Go to this connection profile's overview

Do not add any stream path in the next step, you will use the path defined in the Connection Profile.

Click **CONTINUE**.



Configure stream destination

Define the location and file size strategy that will be used for writing files of events to the Cloud Storage destination.

Stream path prefix

Optional path, must start with a slash ("/")

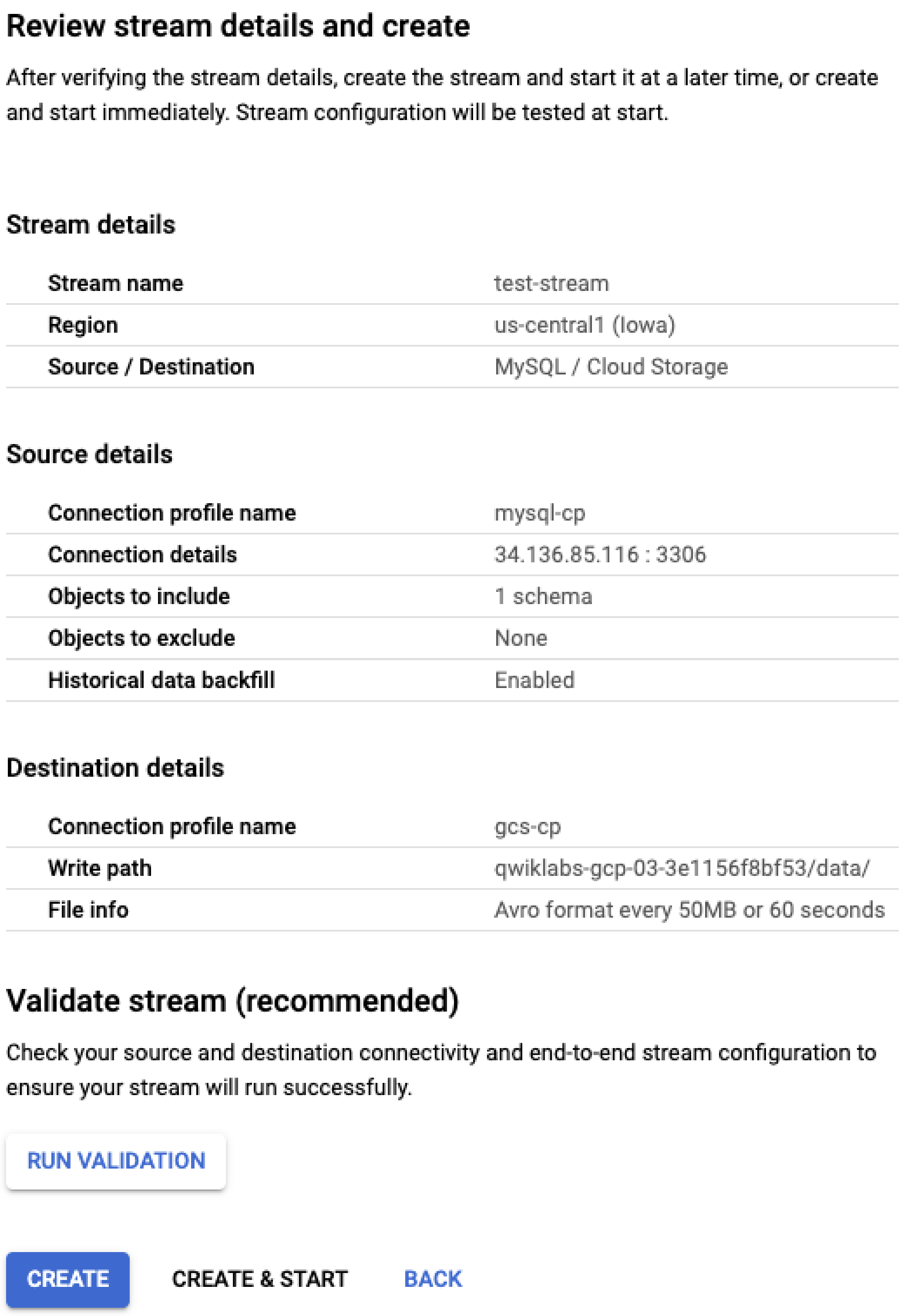
File format

Choose the format of the files written to Cloud Storage.

Output format *: Avro

Finally, validate the stream details by clicking on **RUN VALIDATION**. Once validation completes successfully, click **CREATE AND START**.

Click again **CREATE AND START**.



Review stream details and create

After verifying the stream details, create the stream and start it at a later time, or create and start immediately. Stream configuration will be tested at start.

Stream details

Stream name	test-stream
Region	us-central1 (Iowa)
Source / Destination	MySQL / Cloud Storage

Source details

Connection profile name	mysql-cp
Connection details	34.136.85.116 : 3306
Objects to include	1 schema
Objects to exclude	None
Historical data backfill	Enabled

Destination details

Connection profile name	gcs-cp
Write path	qwiklabs-gcp-03-3e1156f8bf53/data/
File info	Avro format every 50MB or 60 seconds

Validate stream (recommended)

Check your source and destination connectivity and end-to-end stream configuration to ensure your stream will run successfully.

Click *Check my progress* to verify the objective. Create Stream

Create a BigQuery dataset

Using Cloud Shell, run the following `bq` command.

bq mk dataset

Deploy Dataflow job

The Dataflow job can be created from the UI:

Navigation menu > **Analytics** > **Dataflow** > **Jobs** > **Create job from template**

However, use `gcloud` to ensure the variables are submitted correctly.

```
gcloud services enable dataflow.googleapis.com gcloud beta dataflow flex-template run datastream-replication \ --project="${PROJECT_ID}" --region="us-central1" \ --template-file-gcs-location="gs://dataflow-templates-us-central1/latest/flex/Cloud_Datastream_to_BigQuery" \ --enable-streaming-engine \ --parameters \ inputFilePattern="gs://${PROJECT_ID}/data/",\ gcsPubSubSubscription="projects/${PROJECT_ID}/subscriptions/datastream-subscription",\ outputProjectId="${PROJECT_ID}",\ outputStagingDatasetTemplate="dataset",\ outputDatasetTemplate="dataset",\ outputStagingTableNameTemplate="{_metadata_schema}_{_metadata_table}_log",\ outputTableNameTemplate="{_metadata_schema}_{_metadata_table}",\ deadLetterQueueDirectory="gs://${PROJECT_ID}/dlq/",\ maxNumWorkers=2,\ autoscalingAlgorithm="THROUGHPUT_BASED",\ mergeFrequencyMinutes=2,\ inputFileFormat="avro"
```

You can see your running job by navigating **Navigation Menu** > **Dataflow** > **Jobs**.

View the Data in BigQuery

The Dataflow job will replicate your data into the BigQuery dataset supplied. View these tables in the BigQuery UI: **Navigation Menu** > **BigQuery**.

Click *Check my progress* to verify the objective. Create a dataset and deploy Dataflow job

Congratulations!

Datastream is an important tool in your Data Migration and Analytics toolkit! You have learned the basics of MySQL to BigQuery with Datastream and Dataflow.

Google Cloud Training & 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 July 1, 2022

Lab Last Tested July 1, 2022