Ingesting Data Into The Cloud

cloudskillsboost.google/games/2854/labs/17212

GSP194

Overview

In this lab you use a bash script to download selected data from a large public data set available on the internet. This data, made available on the US Bureau of Transport Statistics (BTS) website, provides historic information about internal flights in the United States.

The techniques used to ingest this data from the website into the cloud can be applied to other data sets that provide comprehensive real world data but must be parsed and cleaned before to be usefull.

Objectives

- Retrieve initial data from the BTS website
- Store the data in Cloud Storage
- Load data into Google BigQuery

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.

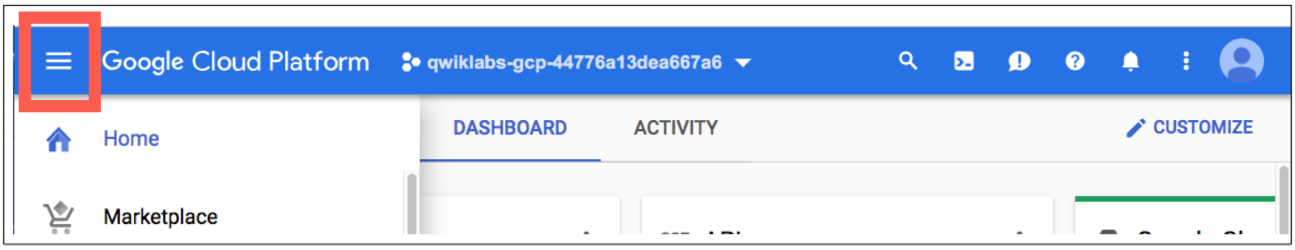
   Note: If you see the **Choose an account** dialog, click **Use Another Account**.
3. If necessary, copy the **Username** from the **Lab Details** panel and paste it into the **Sign in** dialog. Click **Next**.
4. Copy the **Password** from the **Lab Details** panel and paste it into the **Welcome** dialog. 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.

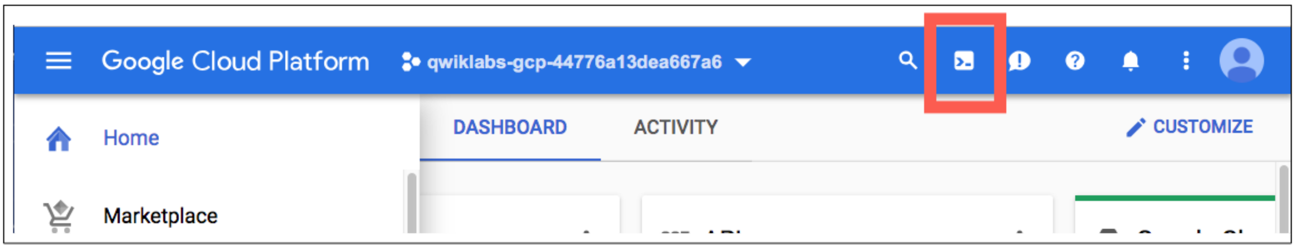
Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left.



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.

Task 1. Prepare your environment

This lab uses a set of code samples and scripts developed for *Data Science on Google Cloud* from O'Reilly Media, Inc. You clone the sample repository used in Chapter 2 from Github to Cloud Shell and then carry out the lab tasks from there.

Clone the Data Science on Google Cloud repository

1. In Cloud Shell enter the following commands to clone the repository:

```
git clone \ https://github.com/GoogleCloudPlatform/data-science-on-gcp/
```

2. Change to the repository directory:

```
cd data-science-on-gcp
```

3. Make a directory to store working data and change into that directory:

```
mkdir data cd data
```

Task 2. Retrieve data from a website

Fetch a sample data file using curl

You will use `curl` to fetch the monthly CSV files that contain the raw data that will be used to build your complete data set. The data set is called the On-Time performance data. You can download a pre-configured data file for each month in any given year from this website.

1. For example, use the following `curl` command to fetch the data from January 2015:

```
curl https://www.bts.dot.gov/sites/bts.dot.gov/files/docs/legacy/additional-attachment-
files/ONTIME.TD.201501.REL02.04APR2015.zip --output data.zip
```

2. Explore the downloaded data file to see what it looks like:

```
unzip data.zip head ontime.td.201501.asc
```

You'll see something similar to the following:

```
AA|1|JFK|LAX|20150101|4|900|900|855|1230|1230|1237|0|0|... AA|1|JFK|LAX|20150102|5|900|900|850|1230|1230|1211|0|0|...
```

Although this file doesn't include header information and appears to contain data that's not needed, it demonstrates one way to acquire a starting data set.

Download custom data from a storage bucket

For this lab, snapshots of custom BTS data have been organized and saved in a public storage bucket. Download it from the `data-science-on-gcp` public storage bucket. A script is provided in the repo to help achieve this.

1. In Cloud Shell, examine the `ingest_from_crsbucket.sh` script:

```
cat ../02_ingest/ingest_from_crsbucket.sh
```

Output:

```
#!/bin/bash if [ "$#" -ne 1 ]; then echo "Usage: ./ingest_from_crsbucket.sh destination-bucket-name" exit fi BUCKET=$1
FROM=gs://data-science-on-gcp/flights/raw TO=gs://$BUCKET/flights/raw CMD="gsutil -m cp " for MONTH in `seq -w 1 12`; do
CMD="$CMD ${FROM}/2015${MONTH}.csv" done CMD="$CMD ${FROM}/201601.csv $TO" echo $CMD $CMD
```

This script copies the monthly BTS data from year 2015 to your destination bucket.

2. To run the script, create a single-region bucket:

```
export PROJECT_ID=$(gcloud info --format='value(config.project)') gsutil mb -l us-central1 gs://${PROJECT_ID}-ml
```

3. Run the download script using your bucket name as the argument:

```
bash ../02_ingest/ingest_from_crsbucket.sh ${PROJECT_ID}-ml
```

Test completed task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will granted with an assessment score.

Create a new Cloud Storage bucket. Copy data files to the storage bucket.

4. In the Cloud Console tab, open **Cloud Storage** > **Browser**.

There should be only one bucket with a name based on the lab project ID appended with `-ml` , for example `qwiklabs-gcp-04-495910fcff5c-ml` . Click the storage bucket to open it.

5. Open the `flights` and `raw` folders.

You will see that the processed `.csv` texts have been copied to the storage bucket.

Task 3. Load data into Google BigQuery

For larger files, it's better to use `gsutil` to ingest the files into Cloud Storage because `gsutil` takes advantage of multithreaded, resumable uploads and is better suited to the public internet.

This is what you did in the previous section when you used `gsutil` to copy the extracted flights CSV files to Cloud Storage. Return to Cloud Shell to load the CSV files into BigQuery.

1. In Cloud Shell, examine the `bqload.sh` script:

```
cat ../02_ingest/bqload.sh
```

Output:

```
#!/bin/bash if [ "$#" -ne 2 ]; then echo "Usage: ./bqload.sh csv-bucket-name YEAR" exit fi BUCKET=$1 YEAR=$2
SCHEMA=Year:STRING,Quarter:STRING,Month:STRING,DayofMonth:STRING,DayOfWeek:STRING,FlightDate:DATE,Reporting_Airline:
# create dataset if not exists PROJECT=$(gcloud config get-value project) #bq --project_id $PROJECT rm -f
${PROJECT}:dsongcp.flights_raw bq --project_id $PROJECT show dsongcp || bq mk --sync dsongcp for MONTH in `seq -w 1 12`; do
CSVFILE=gs://${BUCKET}/flights/raw/${YEAR}${MONTH}.csv bq --project_id $PROJECT --sync \ load --
time_partitioning_field=FlightDate --time_partitioning_type=MONTH \ --source_format=CSV --ignore_unknown_values --
skip_leading_rows=1 --schema=$SCHEMA \ --replace ${PROJECT}:dsongcp.flights_raw\$${YEAR}${MONTH} $CSVFILE done
```

This script loads the data from the Cloud Storage bucket to BigQuery.

2. Run the script using your bucket name as argument:

```
bash ../02_ingest/bqload.sh ${PROJECT_ID}-ml 2015
```

At this point, the CSV files are in Cloud Storage and the raw data in BigQuery.

Load Data into Google BigQuery
This lab has demonstrated how to:

- Fetch raw data from a website in CSV format and perform some basic text actions to tidy it up.
- Copy the data to a Cloud Storage bucket.
- Load the data from a Cloud Storage bucket to BiqQuery where it can be reused easily.

Test your understanding

Below are multiple-choice questions to reinforce your understanding of this lab's concepts. Answer them to the best of your abilities.

Congratulations!

You now know how to copy a large text data set from a website and store it in a cloud storage bucket.

Finish your Quest

This self-paced lab is part of the Data Science on Google Cloud Quest. A Quest is a series of related labs that form a learning path. Completing this Quest earns you the badge above, to recognize your achievement. You can make your badge (or badges) public and link to them in your online resume or social media account. Enroll in this Quest and get immediate completion credit if you've taken this lab. See other available Quests.

Take your next lab

Continue your Quest with Loading Data into Google Cloud SQL, or check out these suggestions:

- Visualizing Data with Google DataStudio
- Processing Data with Google Cloud Dataflow

Next steps / learn more

Here are some follow-up steps:

- Data Science on the Google Cloud Platform, 2nd Edition: O'Reilly Media, Inc..
- Airline Service Quality Performance (On-Time performance data) can be downloaded from this website.

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 April 4, 2022

Lab Last Tested March 3, 2022