Preparing your Environment

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

Clone the Data Science on Google Cloud Repository

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

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

Copied!

Change to the repository directory:

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

Copied!

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

```
mkdir data
cd data
```

Copied!

Retrieving Data From a Web Site

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 web site.

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
```
Copied!

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

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

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|...
```
Copied!

This file doesn't include header information so it is not clear what each field is for, and it appears to contain a lot of data that isn't really needed, but it demonstrates one way to acquire a starting data set.

Download custom data from a storage bucket

Snapshots of custom BTS data have been organized and saved in a public storage bucket.

The simplest way to ensure you get the data you need for this lab is to download it from the data-science-on-gcp public storage bucket. A script is provided in the repo to help achieve this.

In Cloud Shell, examine the ingest_from_crsbucket.sh script:

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

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 the year of 2015 to your destination bucket.

In order to run the script, create a bucket:

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

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

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

Test Completed Task

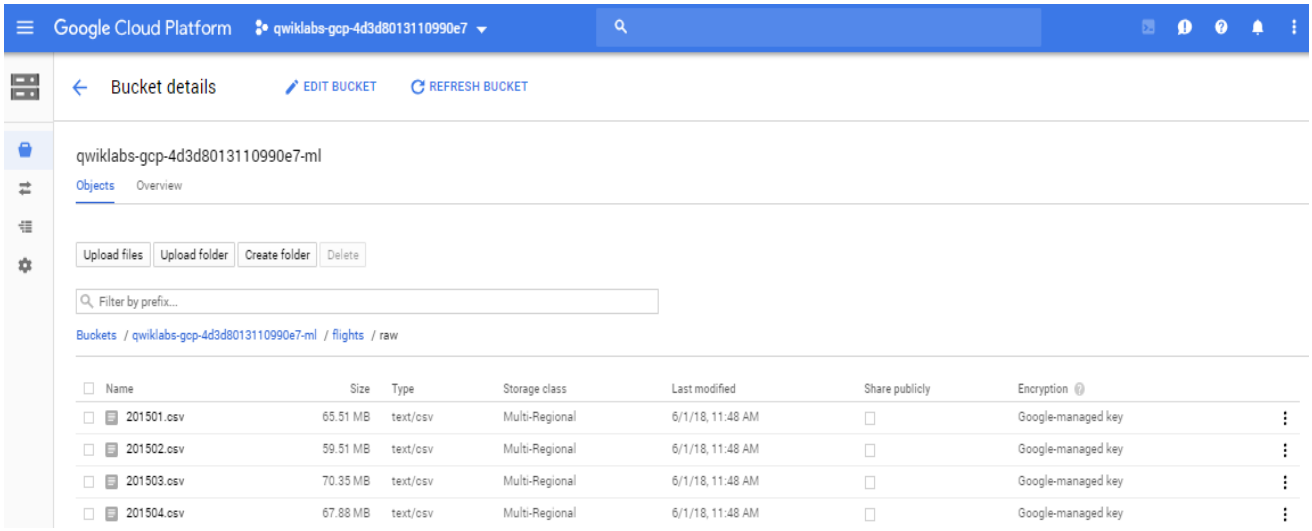
Click **Check my progress** to verify your performed task. If you have completed the task successfully you will granted with an assessment score.

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

There should be only one bucket with a name based on the lab project ID with a -ml suffix appended. Open the storage bucket by clicking on it.

Then open the flights and raw folders.

You will see the twelve processed .csv text have been copied to the storage bucket.



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. The data is then copied to a Cloud Storage bucket where it can be reused easily. All of the later labs in the Data Science on Google Cloud quest will start with these same raw data files already stored in a cloud storage bucket.

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.

True

Congratulations!