CHAPTER 39

Google Cloud Dataprep

Google Cloud Dataprep is a managed cloud service for quick data exploration and transformation. Dataprep makes it easy to clean and transform large datasets for analysis. It is auto-scalable as it takes advantage of the distributed processing capabilities of Google Cloud Dataflow.

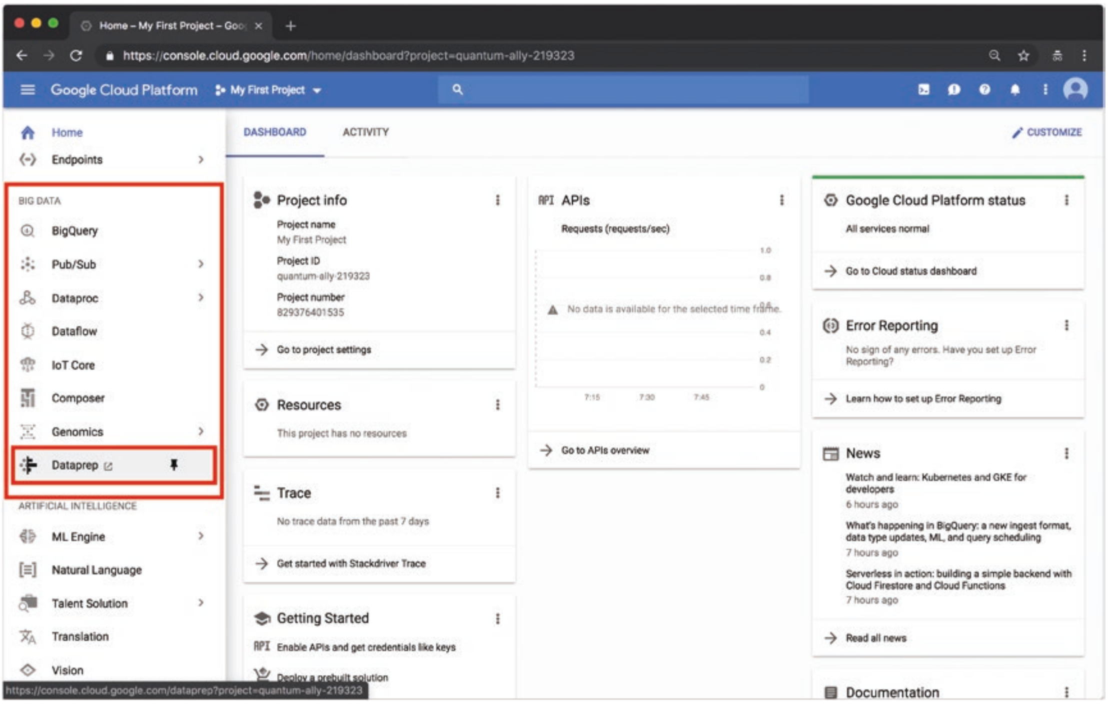
Typically Cloud Dataprep is aimed at easing the data preparation process. Datasets from real-world use cases are often messy and untidy. In this form, it cannot be used for downstream analytics or machine learning modeling. Hence, a large portion of the modeling process involves preparing and cleaning the data. Programming libraries earlier discussed like Pandas are centrally used for carrying out data preparation. However, Google Cloud Dataprep provides a simple visual interface for performing data cleaning. The ability to re-organize the dataset for modeling quickly without coding provides an instant appeal for Dataprep, as this can greatly speed up the time spent in data preparation as part of the overall modeling pipeline. The other good part is that Dataprep can work with petabyte scale data as it is built on a serverless infrastructure. Dataprep can be used for processing structured and unstructured datasets.

In this section, we'll go through a brief tour of Google Dataprep by using it to prepare our 'crypto_markets.csv' dataset already stored on Google Cloud Storage.

Getting Started with Cloud Dataprep

From the GCP dashboard, click the triple dash at the top-left corner and scroll down to 'Dataprep' under the **BIG DATA** section as seen in Figure 39-1.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_39



https://console.cloud.google.com/home/dashboard?project=quantum-ally-219323

Project name
My First Project

Project ID
quantum-ally-219323

Project number
829376401535

This project has no resources

No trace data from the past 7 days

No data is available for the selected time frame.

All services normal

No sign of any errors. Have you set up Error Reporting?

Watch and learn: Kubernetes and GKE for developers
6 hours ago

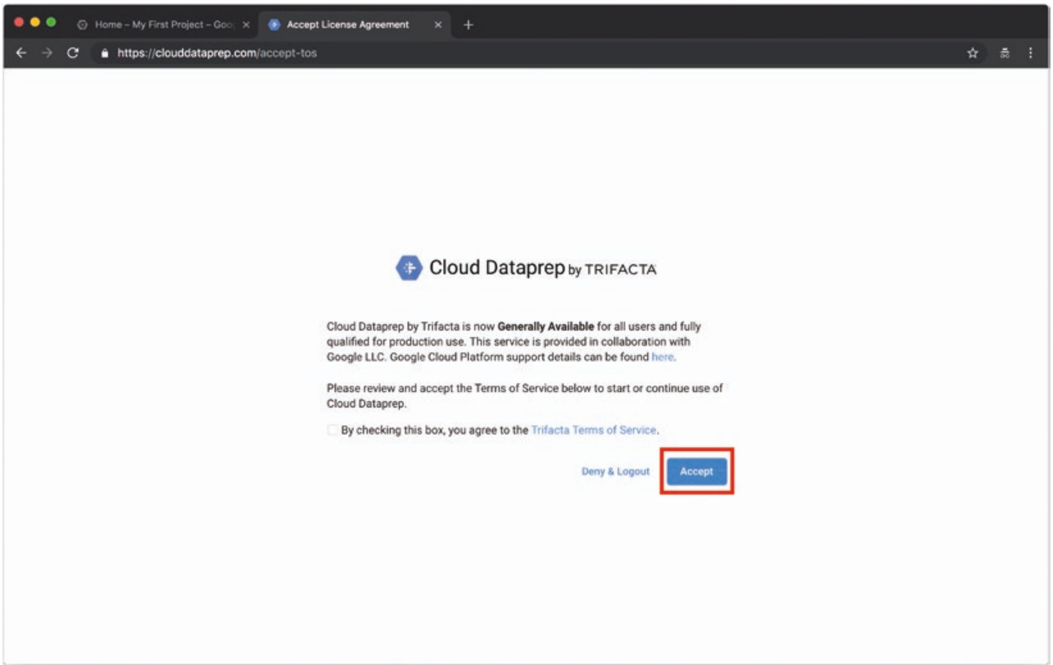
What's happening in BigQuery: a new ingest format, data type updates, ML, and query scheduling
7 hours ago

Serverless in action: building a simple backend with Cloud Firestore and Cloud Functions
7 hours ago

https://console.cloud.google.com/dataprep?project=quantum-ally-219323

Figure 39-1. Open Dataprep via the GCP dashboard

Dataprep is a service offered on GCP in alliance with the company Trifacta. To begin using Dataprep, agree and accept all the license agreements (see Figure 39-2). Dataprep creates a bucket on GCS to store the files that are uploaded to Dataprep and the outputs of its transformation (see Figure 39-3). The Dataprep dashboard is shown in Figure 39-4.


Home – My First Project – Goo
Accept License Agreement
https://clouddataprep.com/accept-tos

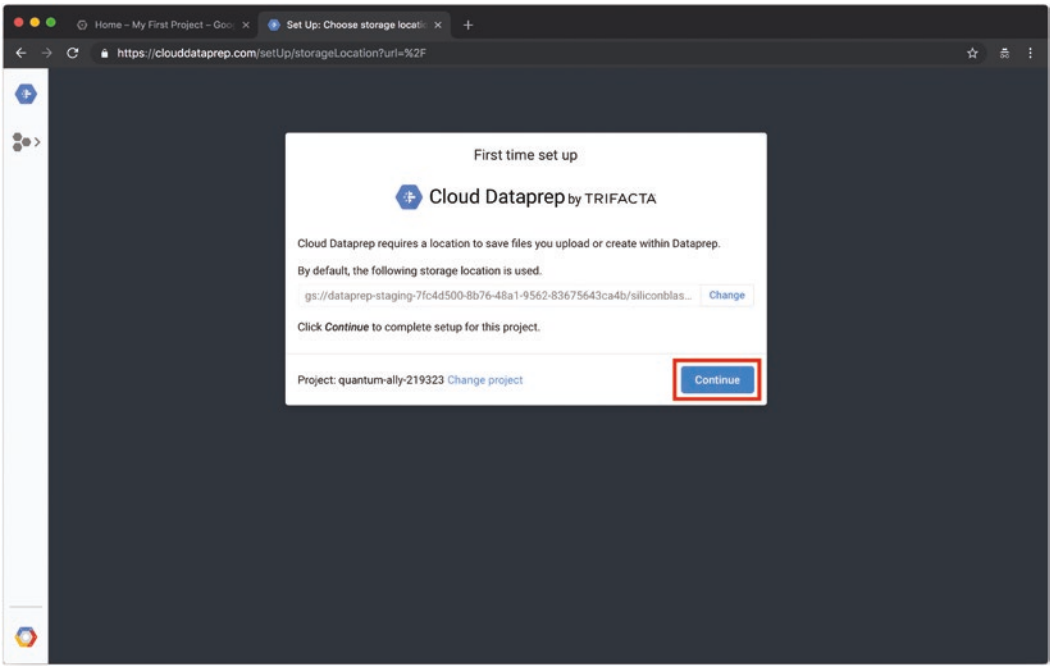
Cloud Dataprep by TRIFACTA

Cloud Dataprep by Trifacta is now **Generally Available** for all users and fully qualified for production use. This service is provided in collaboration with Google LLC. Google Cloud Platform support details can be found here.

Please review and accept the Terms of Service below to start or continue use of Cloud Dataprep.

□ By checking this box, you agree to the Trifacta Terms of Service.

Figure 39-2. *Trifacta license agreement*


Home – My First Project – Goo
Set Up: Choose storage locati
https://clouddataprep.com/setUp/storageLocation?url=%2F

First time set up

Cloud Dataprep by TRIFACTA

Cloud Dataprep requires a location to save files you upload or create within Dataprep.

By default, the following storage location is used.

gs://dataprep-staging-7fc4d500-8b76-48a1-9562-83675643ca4b/siliconblas...

Click *Continue* to complete setup for this project.

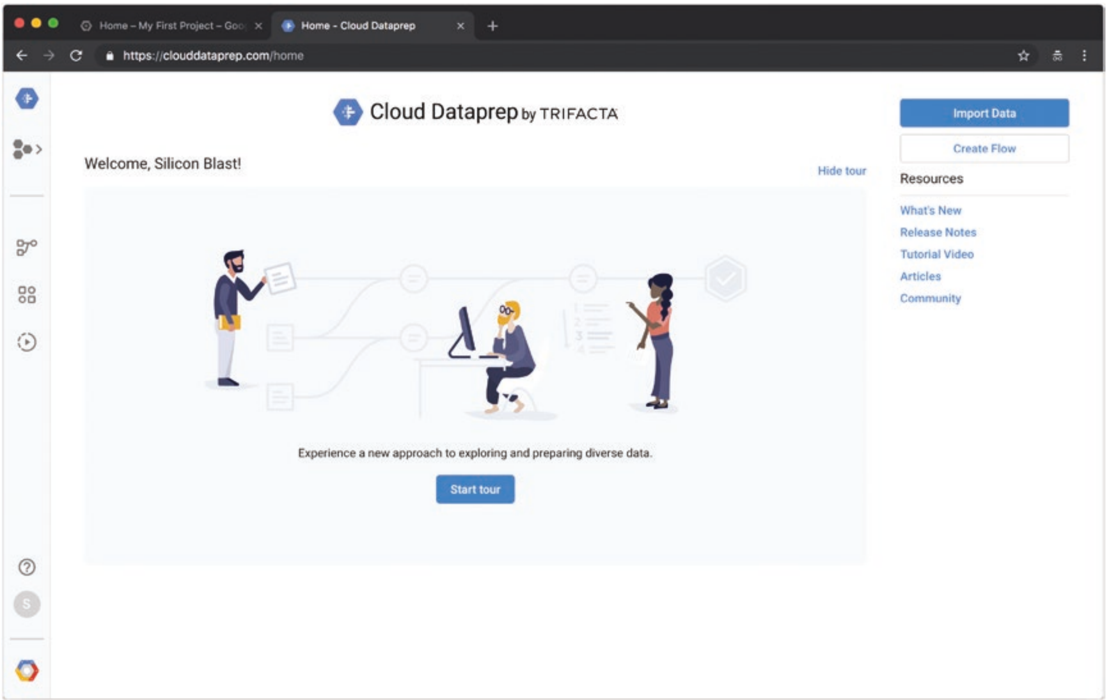
Project: quantum-ally-219323 Change project

Figure 39-3. *Dataprep GCS location setup*

Using Flows to Transform Data

A Dataprep flow is an object created to organize and manage the datasets and operations that are involved in data cleaning and transformation process:

1. We begin by creating a flow by clicking the ‘Create Flow’ button in the top-right corner of the Dataprep dashboard (see Figure 39-4). Enter the user-defined flow name and click ‘Create’ as shown in Figure 39-5. The Flow page is shown in Figure 39-6.



Welcome, Silicon Blast!

Experience a new approach to exploring and preparing diverse data.

Figure 39-4. *Dataprep dashboard*