

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.