This dataset is made available by Kam Hamidieh of the University of Pennsylvania and submitted to the UCI Machine Learning Repository. The goal of this section is to demonstrate delivering an end-to-end machine learning modeling pipeline on Google Cloud Platform.

The Modeling Architecture on GCP

The goal of this end-to-end project is to demonstrate building a large-scale learning model on GCP using the components already discussed in this book. The modeling architecture is illustrated in Figure 44-1. Let's briefly explain the connections:

1. Stage the raw data on GCS.
2. Load data into BigQuery for analytics.
3. Exploratory data analysis.
4. Large-scale data processing with Dataflow.
5. Place transformed training and evaluation data on GCS.
6. Train the model on Cloud MLE.
7. Place the trained model output on GCS.
8. Deploy the model for inference on Cloud MLE.

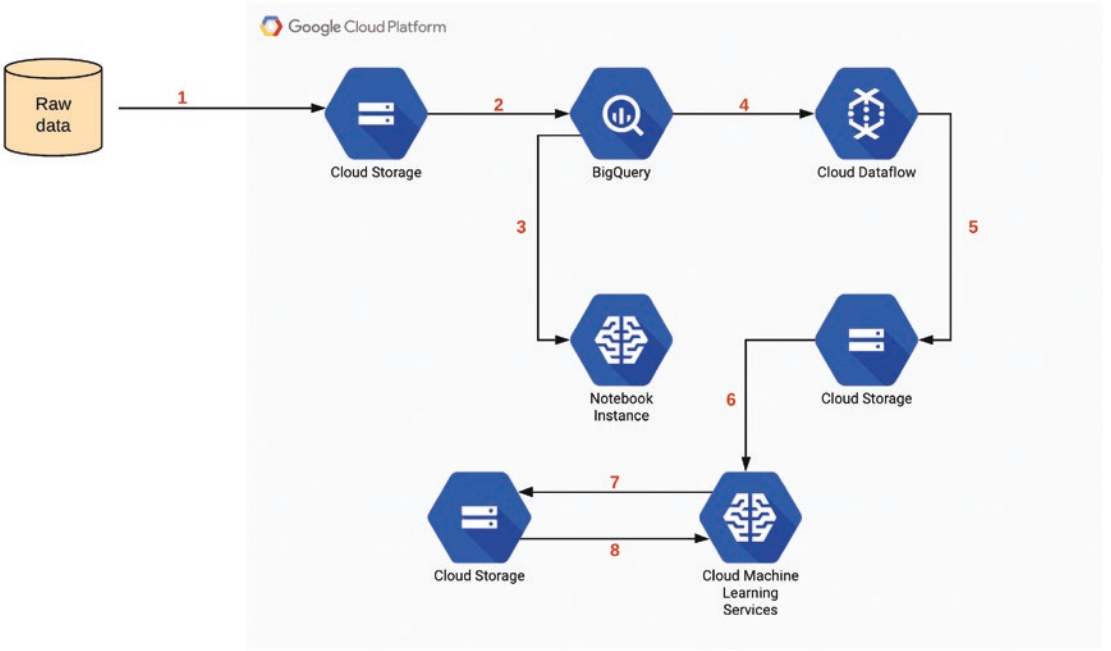

Figure 44-1. *Modeling architecture on GCP*

Stage Raw Data in GCS

Retrieve the raw data from the book code repository for modeling:

- Create a GCS bucket.

  ```
  gsutil mb gs://superconductor
  ```

- Navigate to the chapter folder and transfer the raw data to GCS.

  ```
  gsutil cp train.csv gs://superconductor/raw-data/
  ```

Load Data into BigQuery for Analytics

Move the dataset from Google Cloud Storage to BigQuery:

- Create a Dataset in BigQuery.

  ```
  bq mk superconductor
  ```