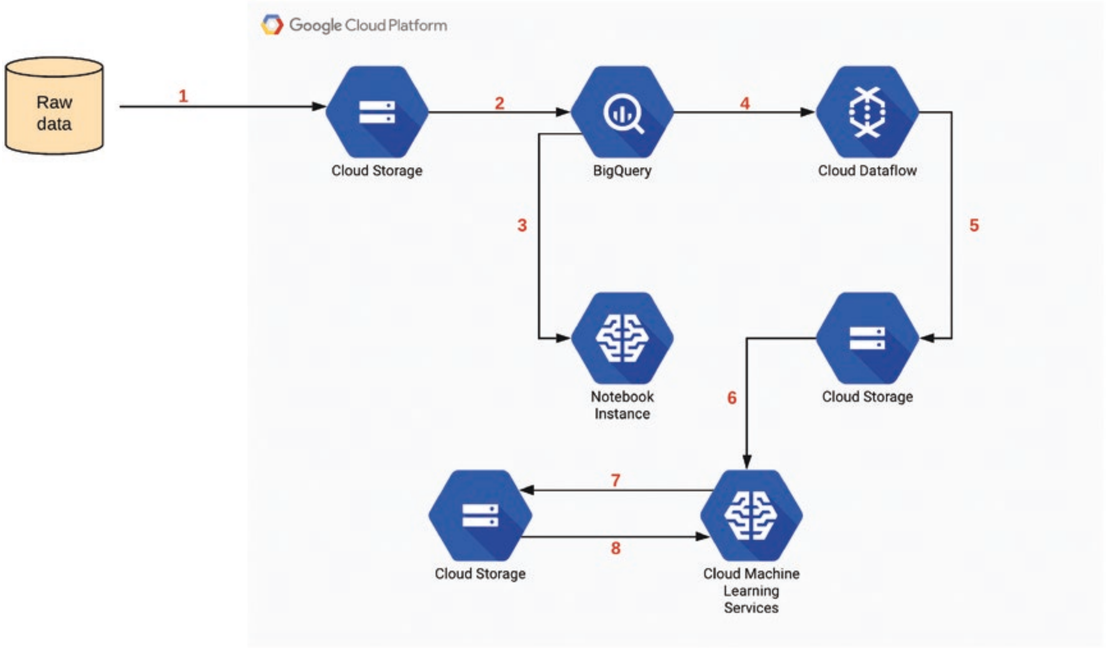

Figure 44-1. *Modeling architecture on GCP*

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
```

- Load raw data from GCS as a Table into the newly created BigQuery Dataset.

```
bq --location=US load --autodetect --source_format=CSV super
conductor.superconductor gs://superconductor/raw-data/train.csv
```

- View created Table schema on BigQuery.

```
bq show superconductor.superconductor

Last modified        Schema         Total Rows   Total Bytes
Expiration   Time Partitioning   Labels
------------- -------------------- ---------- -------------
---------- ------------------ --------
  08 Dec 01:16:51   |- number_of_elements: string
21264       25582000
                    |- mean_atomic_mass: string
                    |- wtd_mean_atomic_mass: string
                    |- wtd_mean_atomic_radius: string
                    |- gmean_atomic_radius: string
                    |- wtd_gmean_atomic_radius: string
                    |- entropy_atomic_radius: string
                    |- wtd_entropy_atomic_radius: string
                    ...
                    |- range_ThermalConductivity: string
                    |- wtd_range_ThermalConductivity: string
                    |- std_ThermalConductivity: string
                    |- wtd_std_ThermalConductivity: string
                    |- mean_Valence: string
                    |- wtd_std_Valence: string
                    |- critical_temp: string
```

Exploratory Data Analysis

The Table in BigQuery contains 21,264 rows. In the interest of speed and rapid iteration, we will not operate on all the rows of this dataset, but rather, we will select a thousand rows for data exploration, transformation, and machine learning spot checking.

```
import pandas as pd
%%bigquery --project ekabasandbox super_cond_df
WITH super_df AS (
SELECT
  number_of_elements, mean_atomic_mass, wtd_mean_atomic_mass,
  gmean_atomic_mass, wtd_gmean_atomic_mass, entropy_atomic_mass,
  wtd_entropy_atomic_mass, range_atomic_mass, wtd_range_atomic_mass,
  std_atomic_mass, wtd_std_atomic_mass, mean_fie, wtd_mean_fie,
  gmean_fie, wtd_gmean_fie, entropy_fie, wtd_entropy_fie, range_fie,
  wtd_range_fie, std_fie, wtd_std_fie, mean_atomic_radius, wtd_mean_atomic_
  radius,
  gmean_atomic_radius, wtd_gmean_atomic_radius, entropy_atomic_radius,
  wtd_entropy_atomic_radius, range_atomic_radius, wtd_range_atomic_radius,
  std_atomic_radius, wtd_std_atomic_radius, mean_Density, wtd_mean_Density,
  gmean_Density, wtd_gmean_Density, entropy_Density, wtd_entropy_Density,
  range_Density, wtd_range_Density, std_Density, wtd_std_Density, mean_
  ElectronAffinity,
  wtd_mean_ElectronAffinity, gmean_ElectronAffinity, wtd_gmean_
  ElectronAffinity
  entropy_ElectronAffinity, wtd_entropy_ElectronAffinity, range_
  ElectronAffinity,
  wtd_range_ElectronAffinity, std_ElectronAffinity, wtd_std_
  ElectronAffinity,
  mean_FusionHeat, wtd_mean_FusionHeat, gmean_FusionHeat, wtd_gmean_
  FusionHeat,
  entropy_FusionHeat, wtd_entropy_FusionHeat, range_FusionHeat,
  wtd_range_FusionHeat, std_FusionHeat, wtd_std_FusionHeat, mean_
  ThermalConductivity,
  wtd_mean_ThermalConductivity, gmean_ThermalConductivity, wtd_gmean_
  ThermalConductivity,
```