Google Cloud

Create the Dataset

Advanced ML with TensorFlow on GCP

End-to-End Lab on Structured Data ML

Production ML Systems

Image Classification Models

Sequence Models

Recommendation Systems

Steps involved in doing ML on GCP

1 Explore the dataset

2 **Create the dataset**

3 Build the model

4 Operationalize the model

Building an ML model involves:

Creating the dataset

Building the model

Operationalizing the model

What makes a feature “good”?

1 Be related to the objective.

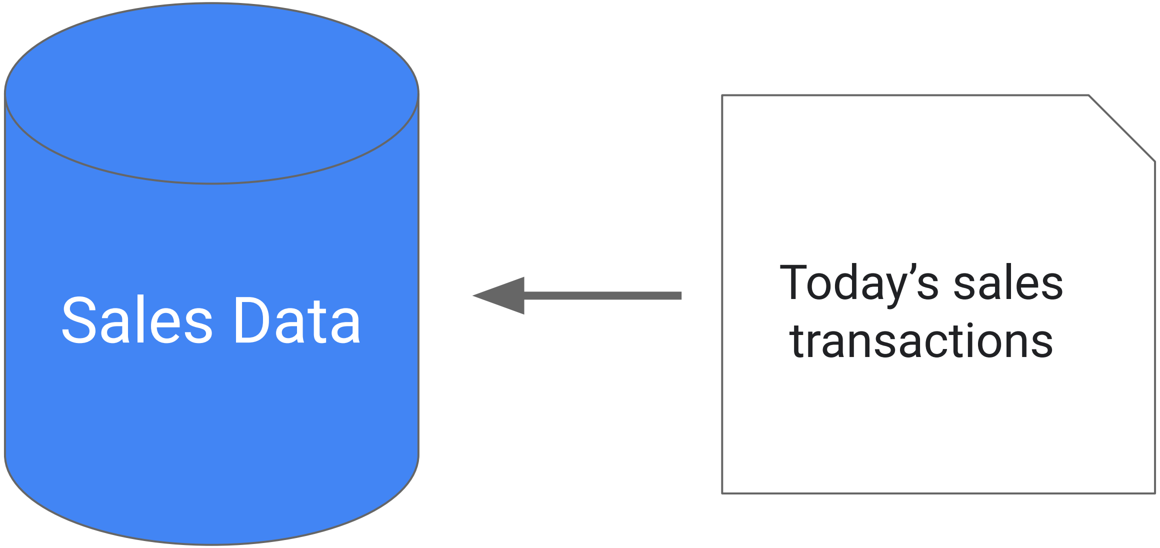
2 **Be known at prediction-time.**

3 Be numeric with meaningful magnitude.

4 Have enough examples.

5 Bring human insight to problem.

Some data could be known immediately, and some other data is not known in real time



Sales Data

Today’s sales transactions

Will we know all these things at prediction time?

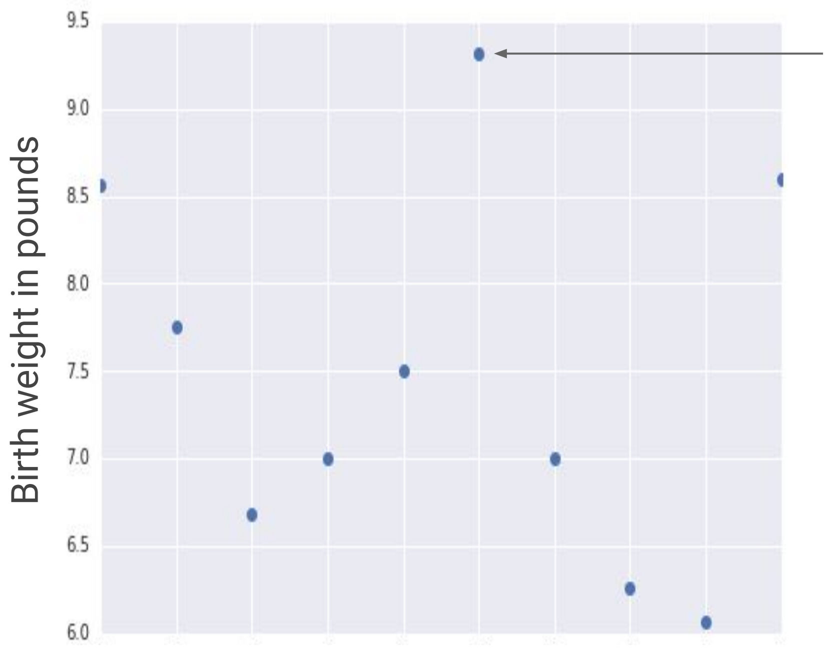
With ultrasound

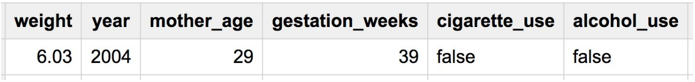
✓ Sex: Male/Female
Plurality: 1, 2, 3, 4, or 5

Without ultrasound

? Sex: Unknown
Plurality: Single/Multiple

The simplest option is to sample rows randomly





weight	year	mother_age	gestation_weeks	cigarette_use	alcohol_use
6.03	2004	29	39	false	false

Each data point is a birth record from the natality dataset.

Random sampling eliminates potential biases due to order of the training examples, but ...

Also ... what about triplets?

3 rows with essentially the same data!

How can we make this data unique?

How can we solve this?

Solution: Split a dataset into training/validation using hashing and modulo operators

```
#standardSQL
SELECT
  date,
  airline,
  departure_airport,
  departure_schedule,
  arrival_airport,
  arrival_delay
FROM
 `bigquery-samples.airline_ontime_data.flights`

WHERE
  MOD(ABS(FARM_FINGERPRINT(date)),10) < 8
```

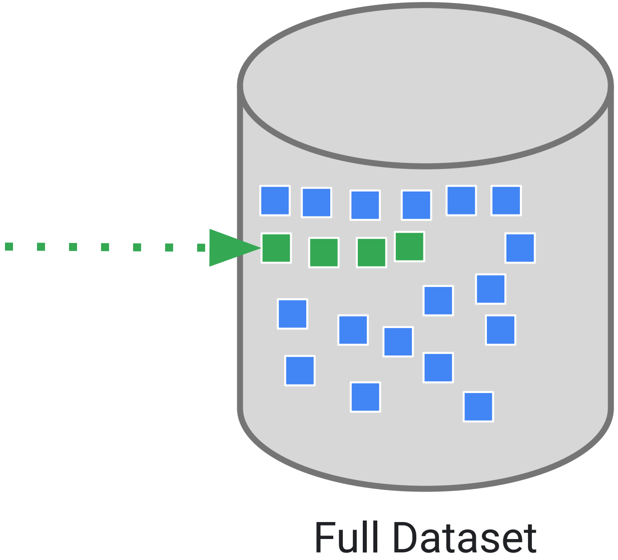
Note: Even though we select date, our model wouldn't actually use it during training.

Hash value on the Date will always return the same value.

Then we can use a modulo operator to only pull 80% of that data based on the last few hash digits.

Developing the ML model software on the entire dataset can be expensive; you want to develop on a smaller sample

Develop your TensorFlow code on a small subset of data, then scale it out to the cloud.



Solution: Sampling the split so that we have a small dataset to develop our code on

```
#standardSQL
SELECT
  date,
  airline,
  departure_airport,
  departure_schedule,
  arrival_airport,
  arrival_delay
FROM
 `bigquery-samples.airline_ontime_data.flights`

WHERE
  MOD(ABS(FARM_FINGERPRINT(date)),10) < 8  AND   RAND() < 0.01
```

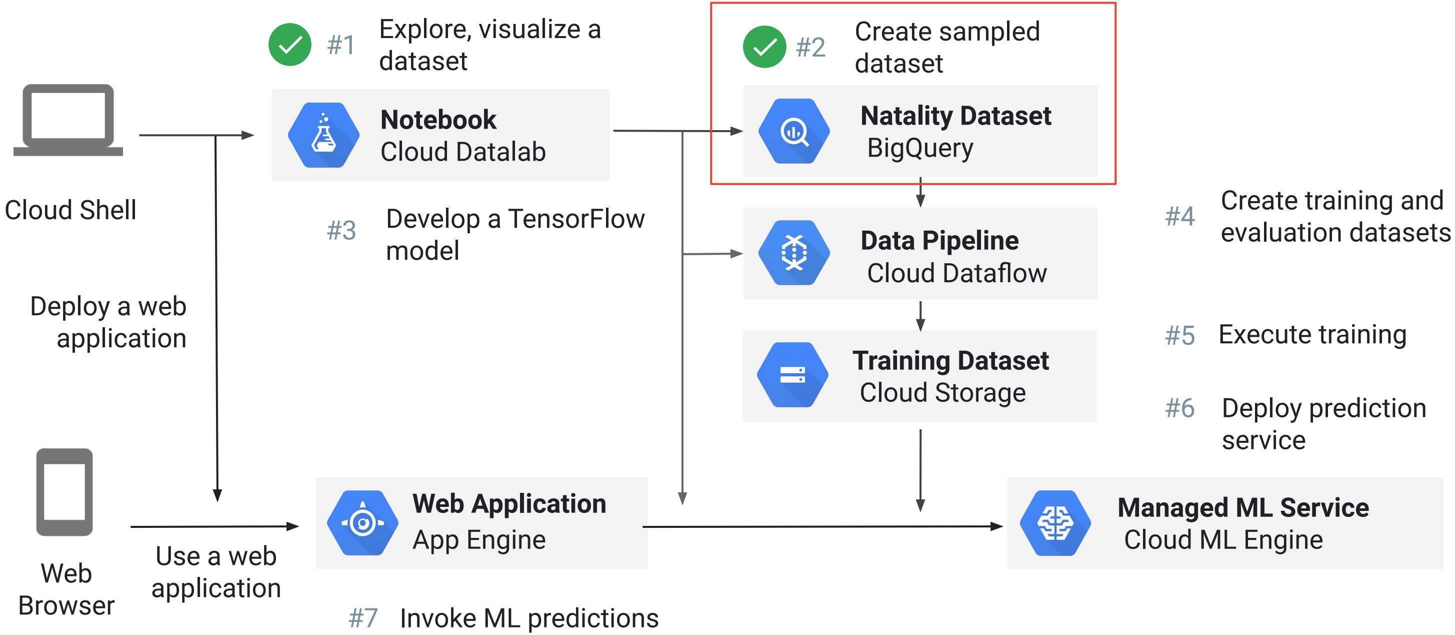
Lab

Creating a sampled dataset

In this lab, you will sample a BigQuery dataset to create datasets for ML, and preprocess data using Pandas.

https://www.oreilly.com/learning/repeatable-sampling-of-data-sets-in-bigquery-for-machine-learning

The end-to-end process



✓ #1 Explore, visualize a dataset

Notebook
Cloud Datalab

✓ #2 Create sampled dataset

Natality Dataset
BigQuery

Cloud Shell

#3 Develop a TensorFlow model

Data Pipeline
Cloud Dataflow

#4 Create training and evaluation datasets

Deploy a web application

Training Dataset
Cloud Storage

#5 Execute training

#6 Deploy prediction service

Web Application
App Engine

Managed ML Service
Cloud ML Engine

Web Browser

Use a web application

#7 Invoke ML predictions

cloud.google.com