



---

Explore 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



# The most common ML models at Google are models that operate on structured data

Type of network	# of network layers	# of weights	% of deployed models
MLP0	5	20M	61%
MLP1	4	5M	
LSTM0	58	52M	29%
LSTM1	56	34M	
CNN0	16	8M	5%
CNN1	89	100M	

<https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu>



Our goal is to predict the weight of newborns so that all newborns can get the care they need



Predict the weight  
of newborns



Identify babies  
who may need  
special facilities



Get babies the  
care they need



# This is what we will build

### Baby weight predictor

*Example application to predict a baby's weight.*

Mother's age 27

Gestation weeks 38

Plurality Single ▾

Baby's gender ☐ Male ☒ Female ☐ Unknown

**PREDICT**

Prediction 7.19 lbs.

### Baby weight predictor

*Example application to predict a baby's weight.*

Mother's age 39

Gestation weeks 33

Plurality Twins ▾

Baby's gender ☐ Male ☐ Female ☒ Unknown

**PREDICT**

Prediction 4.36 lbs.



# An open dataset of births is available in BigQuery

Births recorded in the 50 states of the USA from 1969 to 2008.

<b>Table ID</b>	bigquery-public-data:samples.nativity
<b>Table Size</b>	21.9 GB
<b>Long Term Storage Size</b>	21.9 GB
<b>Number of Rows</b>	137,826,763



<https://bigquery.cloud.google.com/table/bigquery-public-data:samples.nativity>



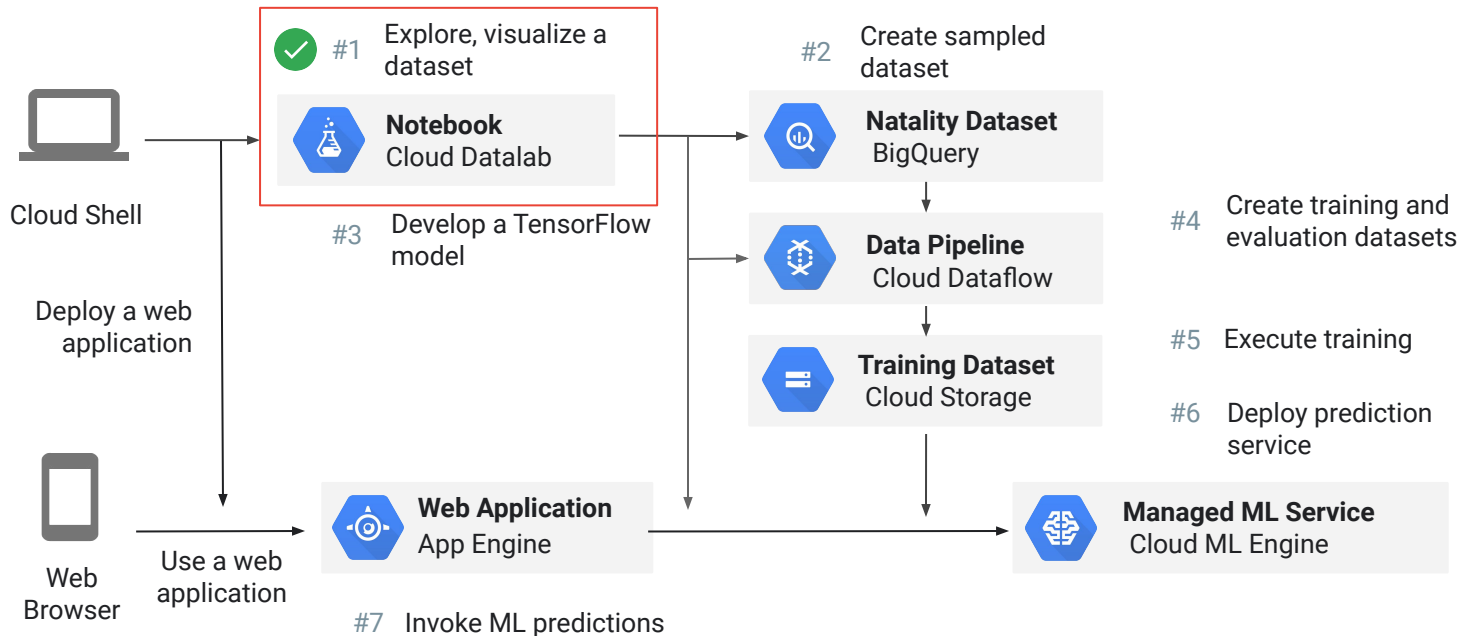
# The data set includes details about the pregnancy

Date of birth	<b>year</b>	INTEGER	NULLABLE	Four-digit year of the birth. Example: 1975.
	<b>month</b>	INTEGER	NULLABLE	Month index of the date of birth, where 1=January.
	<b>day</b>	INTEGER	NULLABLE	Day of birth, starting from 1.
	<b>wday</b>	INTEGER	NULLABLE	Day of the week, where 1 is Sunday and 7 is Saturday.
Location of birth (US state)	<b>state</b>	STRING	NULLABLE	The two character postal code for the state. Entries after 2004 do not include this value.
Baby's birth weight (lbs)	<b>weight_pounds</b>	FLOAT	NULLABLE	Weight of the child, in pounds.
Mother's age at birth	<b>mother_age</b>	INTEGER	NULLABLE	Reported age of the mother when giving birth.
Duration of pregnancy	<b>gestation_weeks</b>	INTEGER	NULLABLE	The number of weeks of the pregnancy.
Mother's weight gain (lbs)	<b>weight_gain_pounds</b>	INTEGER	NULLABLE	Number of pounds gained by the mother during pregnancy.





## The end-to-end machine learning set of labs



# BigQuery is a serverless data warehouse

- 1 Interactive analysis of petabyte scale databases.
- 2 Familiar, SQL 2011 query language and functions.
- 3 Many ways to ingest, transform, load, export data to/from BigQuery.
- 4 Nested and repeated fields, user-defined functions.
- 5 Data storage is inexpensive; queries charged on amount of data processed (or a monthly flat rate).



# Run a query from BigQuery web UI

The screenshot displays the BigQuery web interface. At the top is the 'Query editor' with a SQL query. Below the editor are buttons for 'Run query', 'Save query', 'Save view', and a 'More' dropdown menu. A status message indicates the query will process 1.45 GB. Below this is the 'Query results' section, which includes a 'SAVE AS' dropdown, 'EXPLORE IN DATA STUDIO', and tabs for 'Job information', 'Results', 'JSON', and 'Execution details'. The 'Results' tab is active, showing a table with 4 rows and 4 columns. Annotations with arrows point to various UI elements: 'Run' points to the 'Run query' button; 'Save/Share' points to the 'Save query' and 'Save view' buttons; 'More options' points to the 'More' dropdown; 'VALIDATE' points to the green checkmark in the status message; 'Cost' points to the 'SAVE AS' dropdown; and 'EXPORT' points to the 'EXPLORE IN DATA STUDIO' link.

```
1 SELECT
2   airline,
3   SUM(IF(arrival_delay > 0, 1, 0)) AS num_delayed,
4   COUNT(arrival_delay) AS total_flights
5 FROM
6   `bigquery-samples.airline_ontime_data.flights`
7 WHERE
8   arrival_airport='OKC'
9   AND departure_airport='DFW'
10 GROUP BY
11   airline
```

Run query | Save query | Save view | More

This query will process 1.45 GB when run. ✓

Query results | SAVE AS | EXPLORE IN DATA STUDIO

Query complete (1.602 sec elapsed, 1.45 GB processed)

Job information | Results | JSON | Execution details

Row	airline	num_delayed	total_flights
1	AA	10312	23060
2	OO	198	552
3	EV	756	1912
4	MQ	3884	7903

Analyze query performance

<https://bigquery.cloud.google.com/>

More options

VALIDATE

Cost

EXPORT



# Demo: Query large datasets in seconds

```
# standardsql

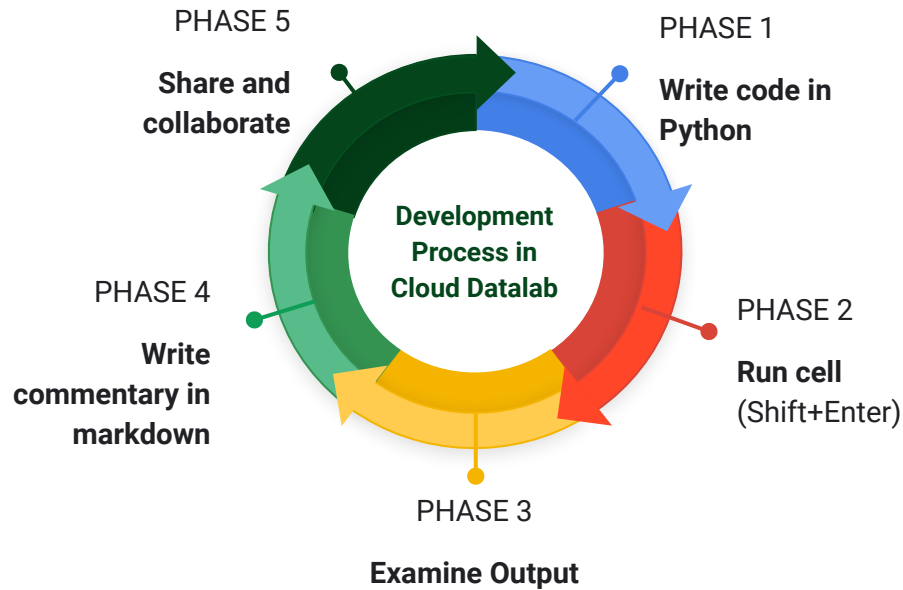
# medicare claims in 2014
SELECT
  npes_provider_state AS state,
  ROUND(SUM(total_claim_count) / 1e6) AS total_claim_count_millions
FROM
  `bigquery-public-data.medicare.part_d_prescriber_2014`
GROUP BY
  state
ORDER BY
  total_claim_count_millions DESC
LIMIT 5;
```

Row	state	total_claim_count_millions
1	CA	116.0
2	FL	91.0
3	NY	80.0
4	TX	76.0
5	PA	63.0

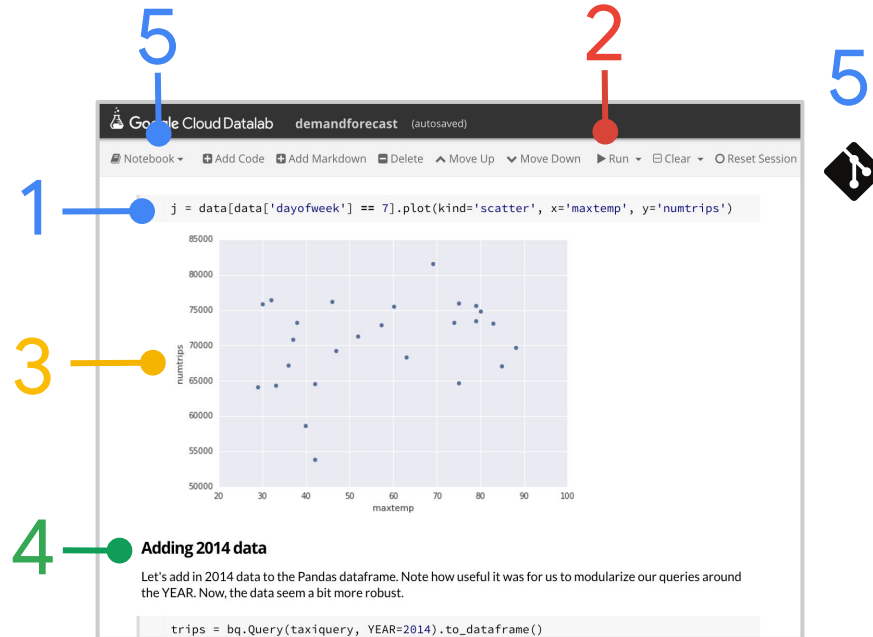
<https://bigquery.cloud.google.com/savedquery/663413318684:781a98ddf2264505af2b6a8fc398a80e>



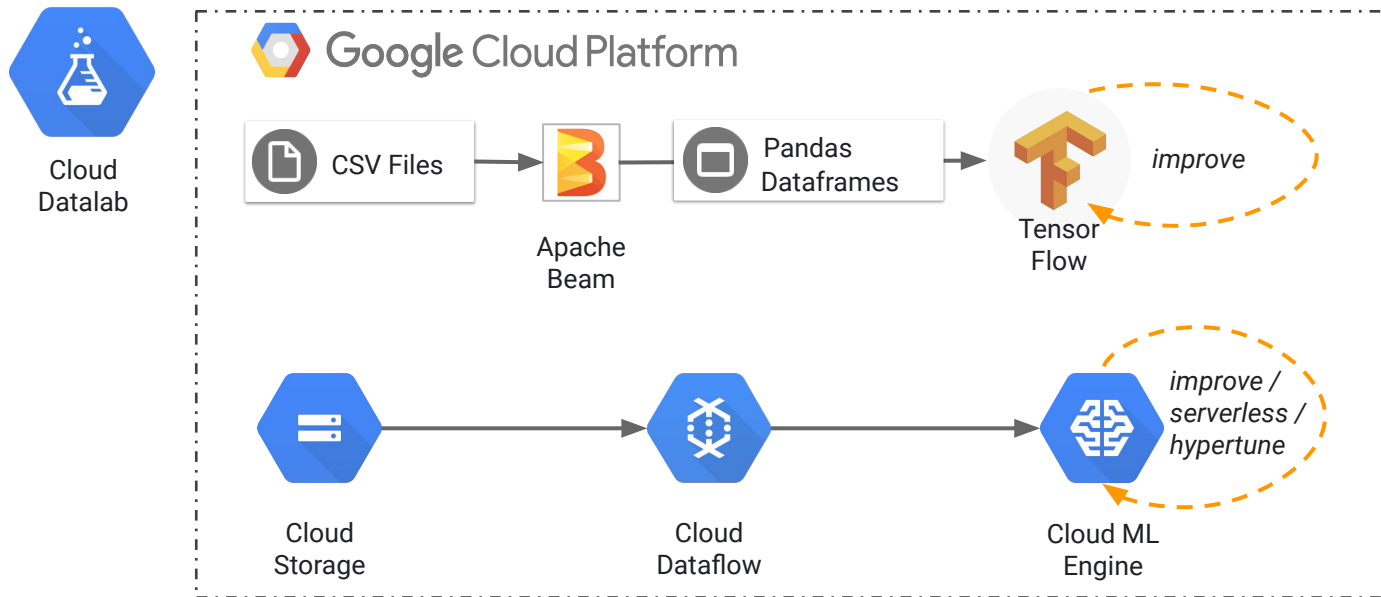
# Cloud Datalab notebooks are developed in an iterative, collaborative process



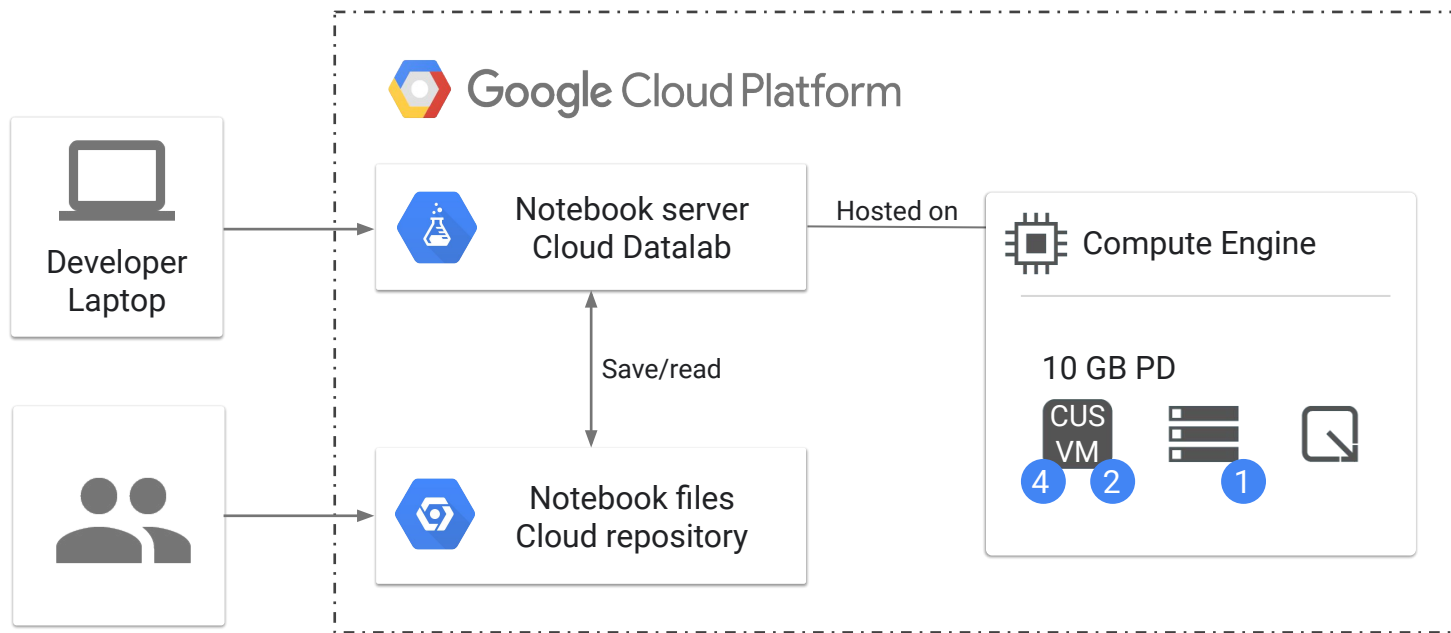
# Cloud Datalab notebooks are developed in an iterative, collaborative process



# You can develop locally with Cloud Datalab and then scale out data processing to the cloud



# Cloud Datalab notebooks let you change the underlying hardware





# Starting Cloud Datalab in Cloud Shell is simple

Google Cloud Platform Project

Activity

1 Activate Google Cloud Shell

```
datalab create my-datalab-vm \  
--machine-type n1-highmem-8 \  
--zone us-central1-a
```

2

3

4

Change Preview Port

Port Number: 8081

CANCEL CHANGE AND PREVIEW



# Preprocessing data at scale with BigQuery + Cloud Datalab



# BigQuery in Python to get a Pandas DF

```
query = """
SELECT
    weight_pounds,
    is_male,
    mother_age,
    plurality,
    gestation_weeks,
    ABS(FARM_FINGERPRINT(CONCAT(CAST(YEAR AS STRING), CAST(month AS STRING)))
FROM
    publicdata.samples.natality
WHERE year > 2000
"""
```

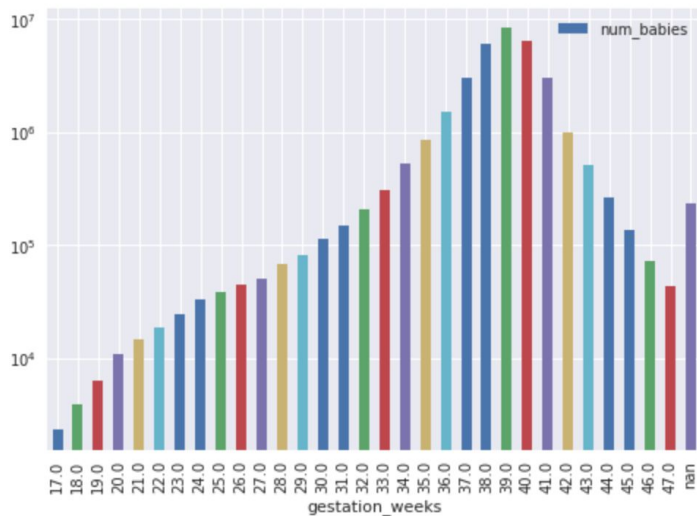
```
# Call BigQuery and examine in dataframe
import google.cloud.bigquery as bq
df = bq.Query(query + " LIMIT 100").execute().result().to_dataframe()
df.head()
```

	weight_pounds	is_male	mother_age	plurality	gestation_weeks	hashmonth
0	3.562670	True	25	1	30	1403073183891835564
1	3.999185	False	30	1	32	7146494315947640619



# Pandas + BigQuery in notebook rocks!

```
# Bar plot to see gestation_weeks with avg_wt linear and num_babies logarithmic
df = get_distinct_values(['gestation_weeks'])
df = df.sort_values('gestation_weeks')
df.plot(x='gestation_weeks', y='num_babies', logy=True, kind='bar');
df.plot(x='gestation_weeks', y='avg_wt', kind='bar');
```



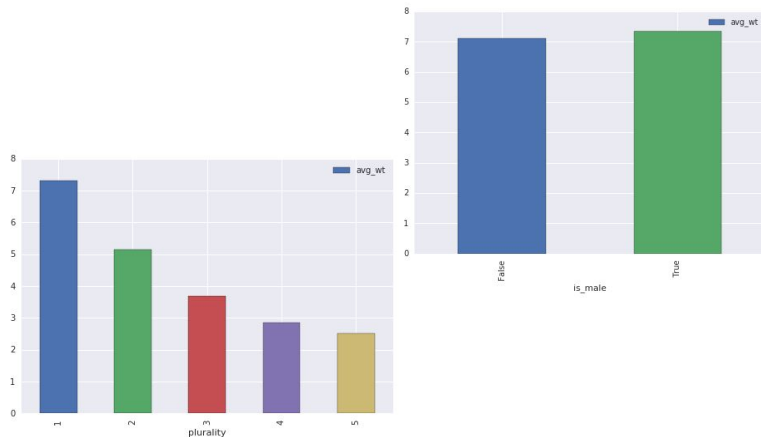
# Lab

---

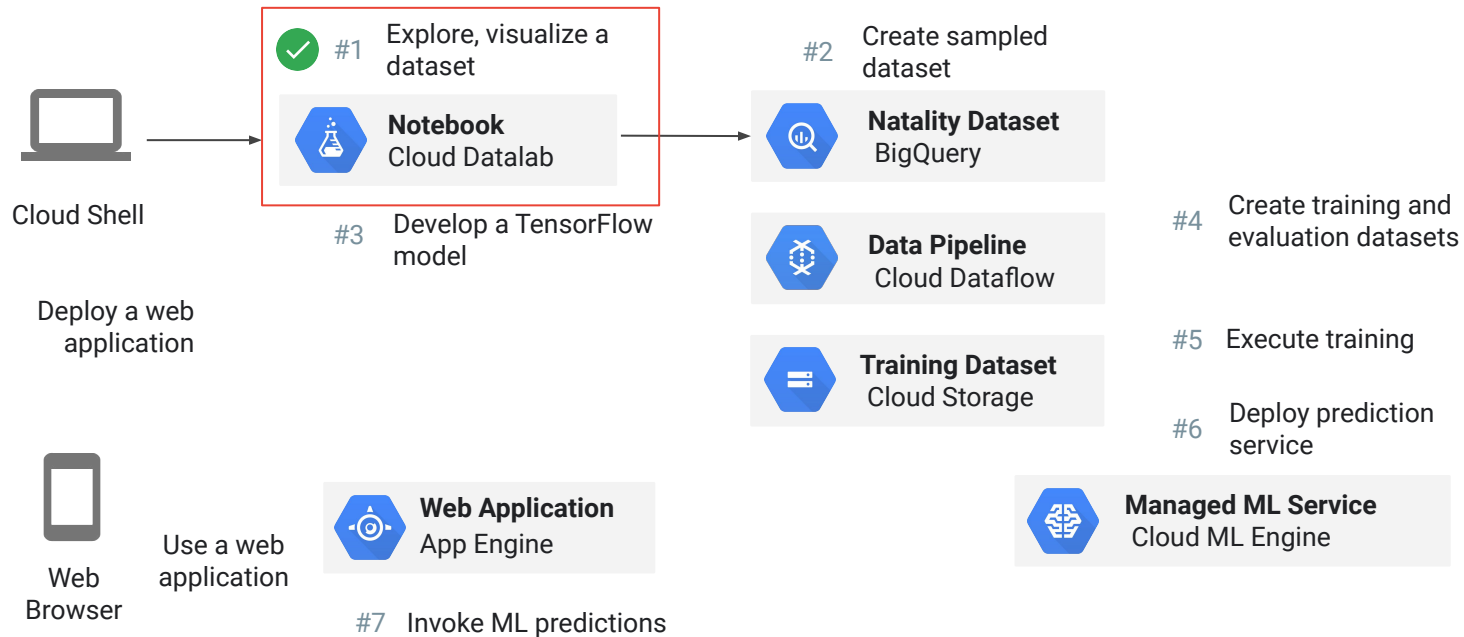
Explore a BigQuery dataset to find features to use in an ML model

In this lab, you will investigate which features have influence on what you want to predict: the baby's weight.

`publicdata.samples.natality`



# The end-to-end process



cloud.google.com

