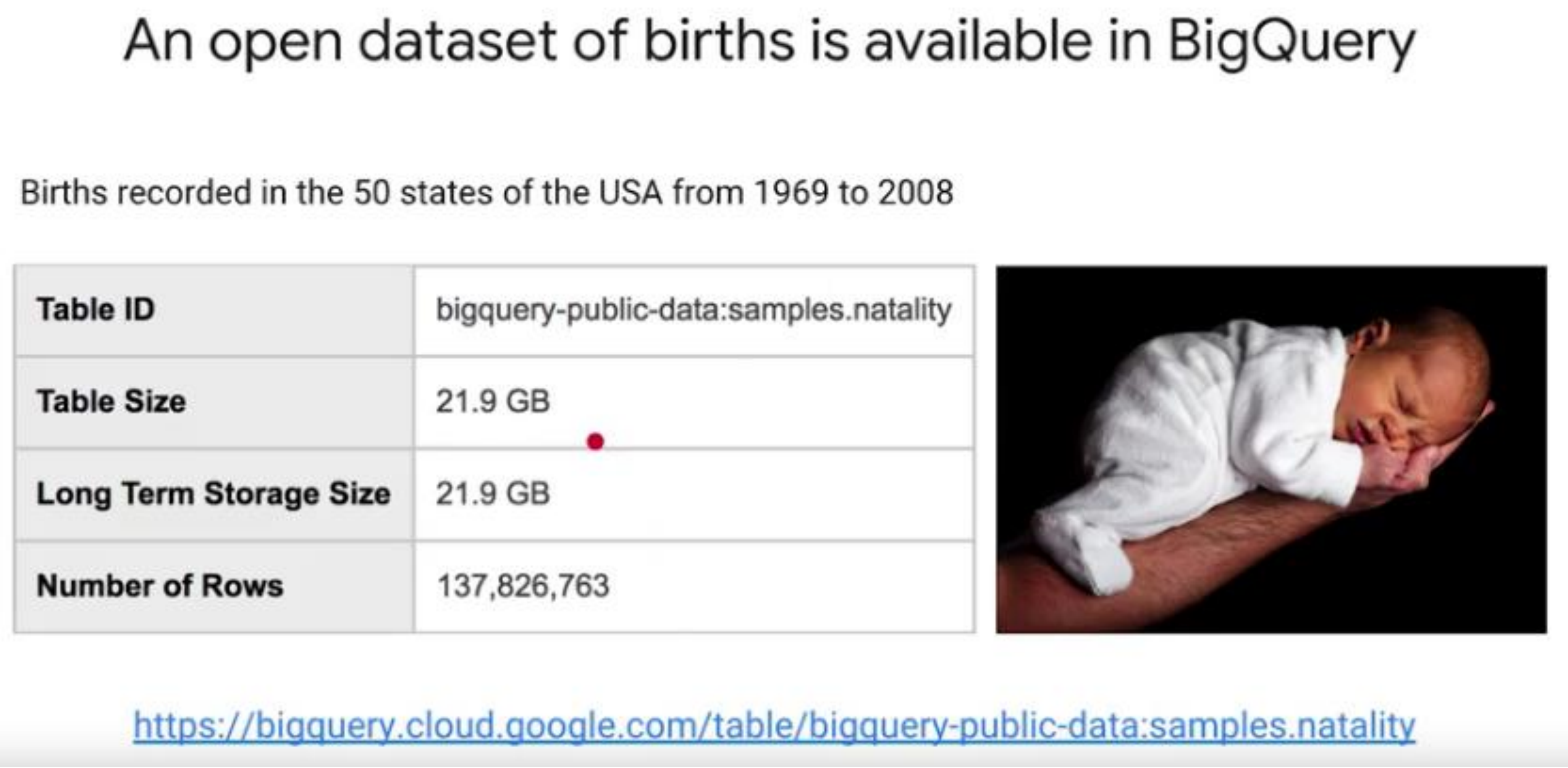
Explorar Dataset

TensorFlow

CARACTERÍSTICAS



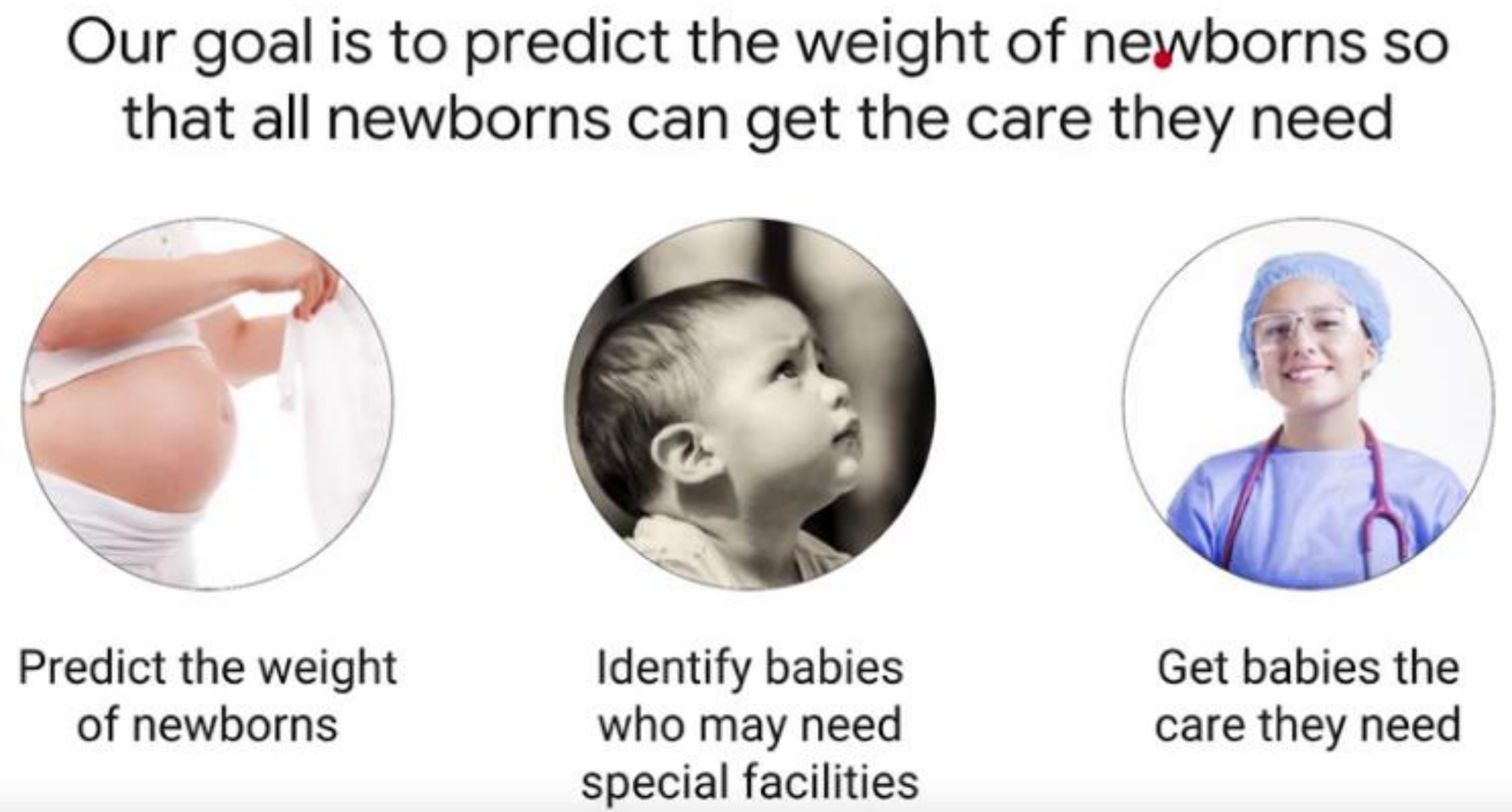
An open dataset of births is available in BigQuery

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

Table ID	bigquery-public-data:samples.natality
Table Size	21.9 GB
Long Term Storage Size	21.9 GB
Number of Rows	137,826,763

https://bigquery.cloud.google.com/table/bigquery-public-data:samples.natality

Por que explorar?



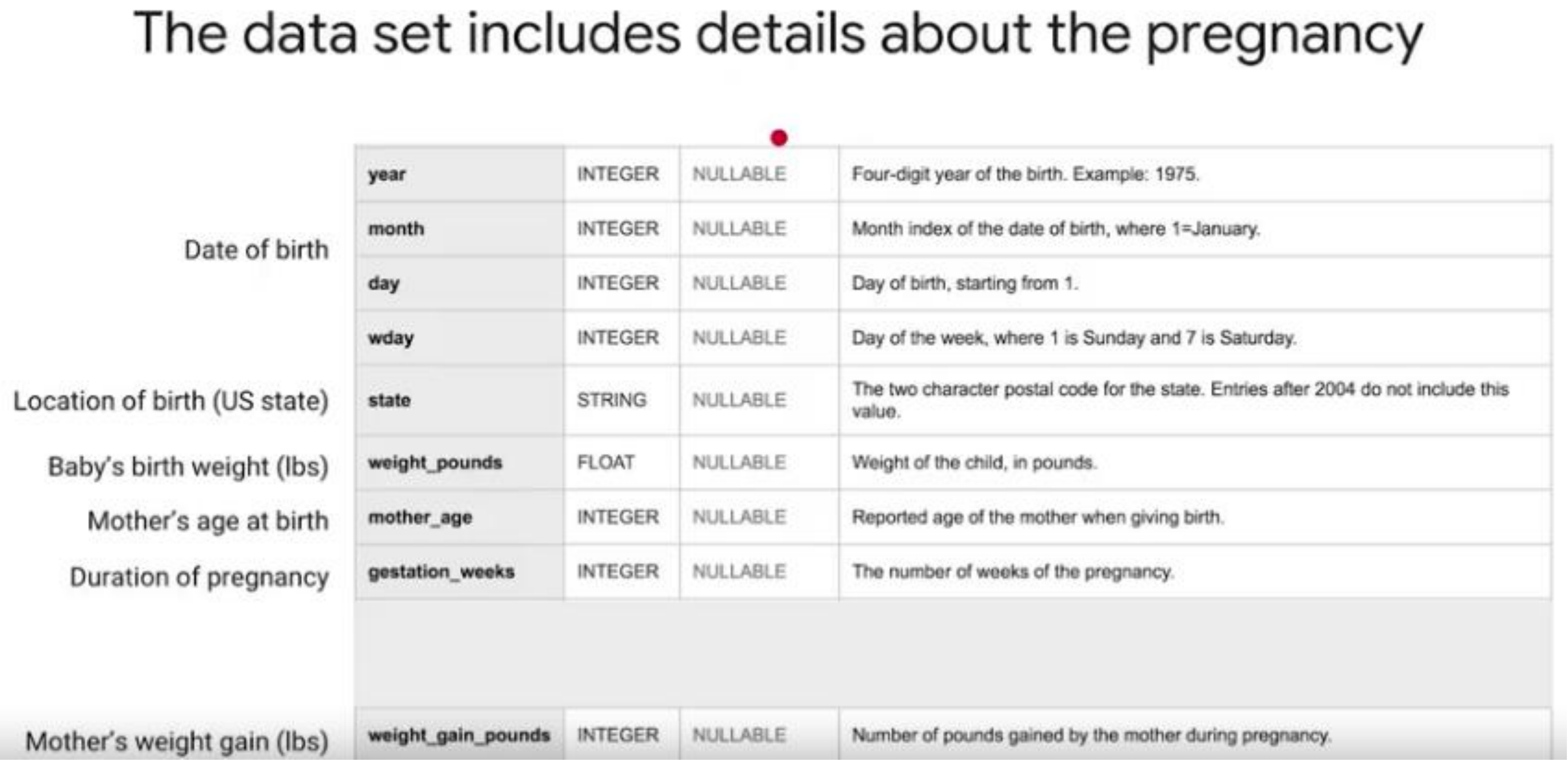
Our goal is to predict the weight of newborns so that all newborns can get the care they need

Predict the weight of newborns

Identify babies who may need special facilities

Get babies the care they need

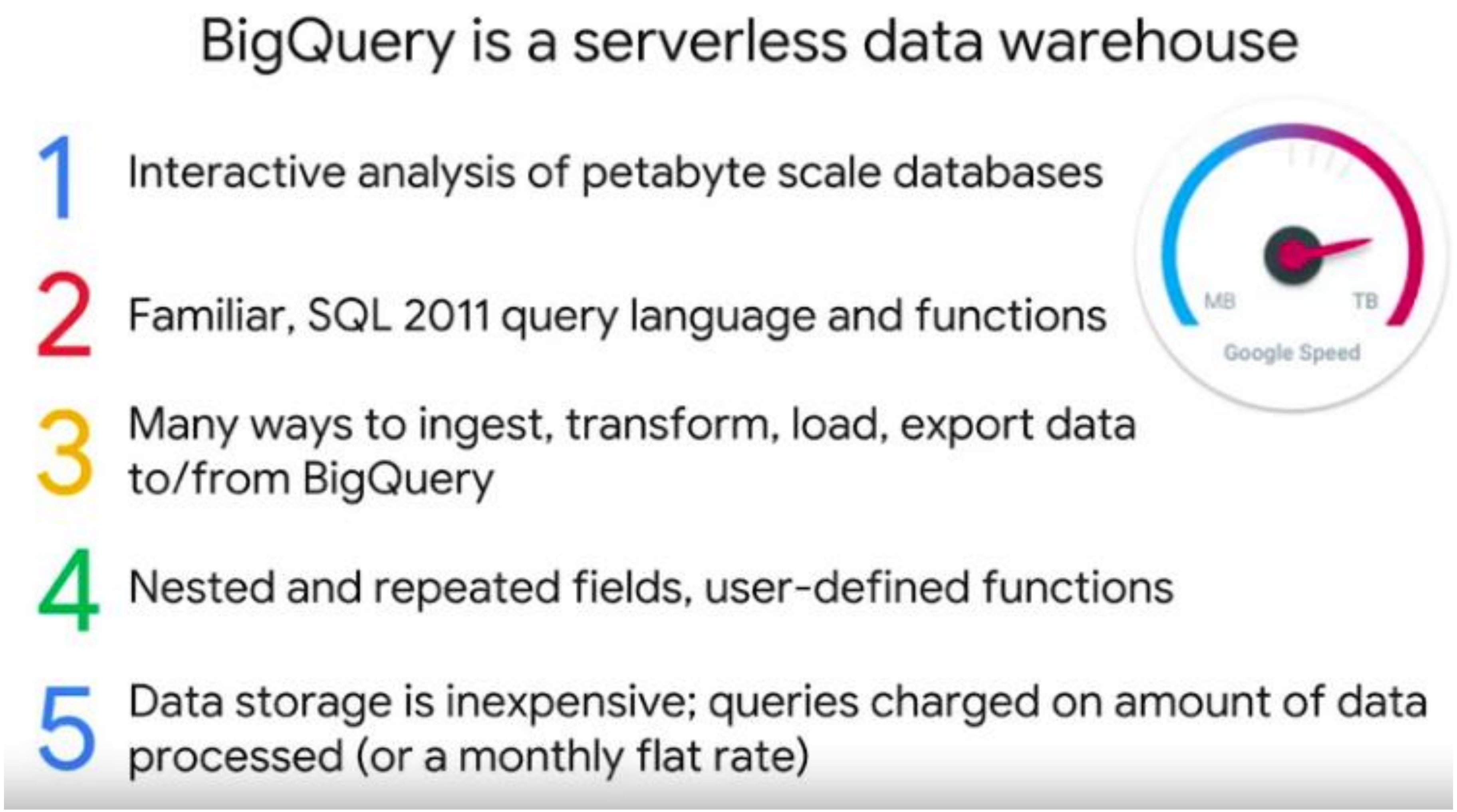
Detalhes da tabela



The data set includes details about the pregnancy

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

BigQuery



BigQuery is a serverless data warehouse

1 Interactive analysis of petabyte scale databases

2 Familiar, SQL 2011 query language and functions

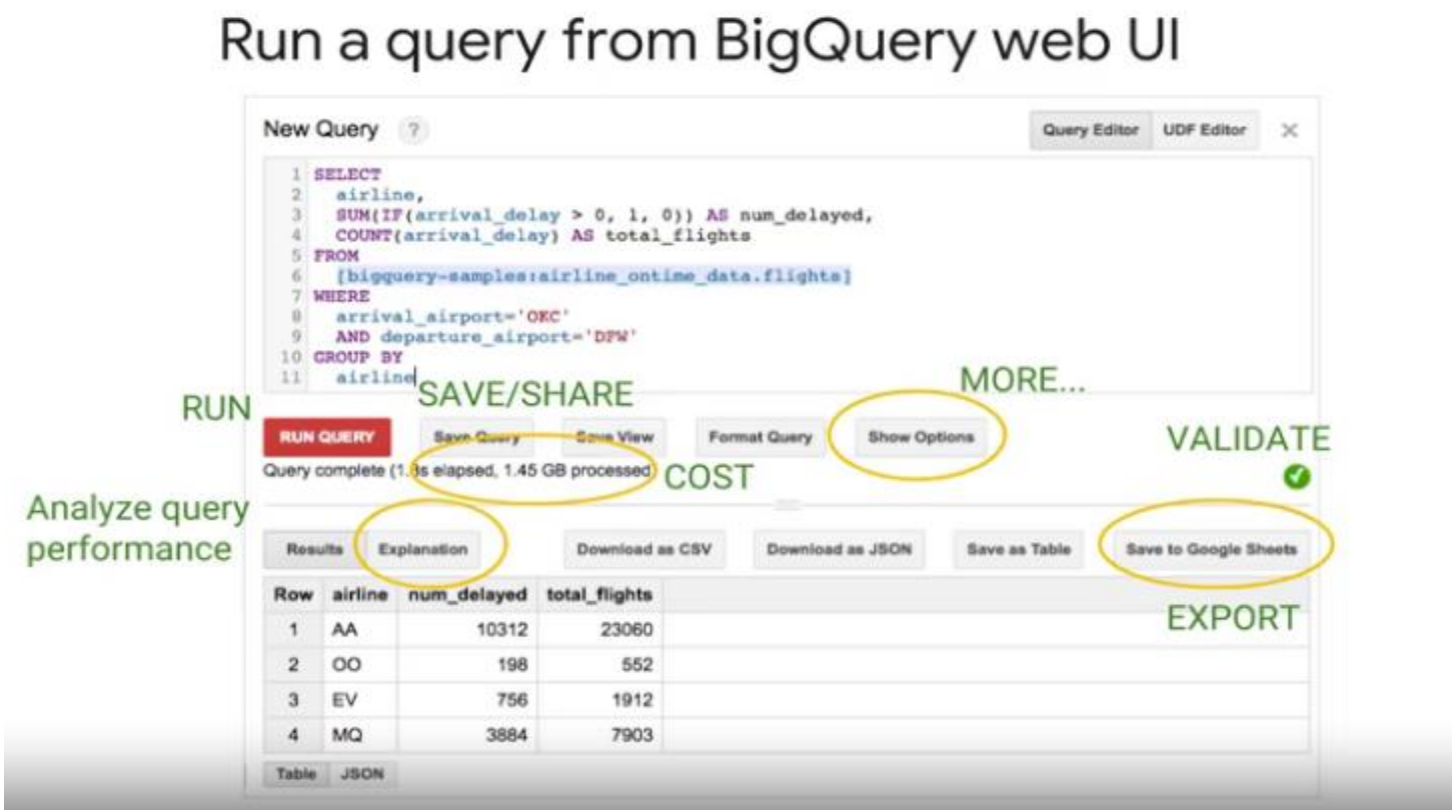
3 Many ways to ingest, transform, load, export data to/from BigQuery

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)

MB
TB
Google Speed

BigQuery



Run a query from BigQuery web UI

```
SELECT
  airline,
  SUM(IF(arrival_delay > 0, 1, 0)) AS num_delayed,
  COUNT(arrival_delay) AS total_flights
FROM
  [bigquery-samples:airline_ontime_data.flights]
WHERE
  arrival_airport='OKC'
  AND departure_airport='DFW'
GROUP BY
  airline
```

Query complete (1.8s elapsed, 1.45 GB processed)

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

RUN

SAVE/SHARE

MORE...

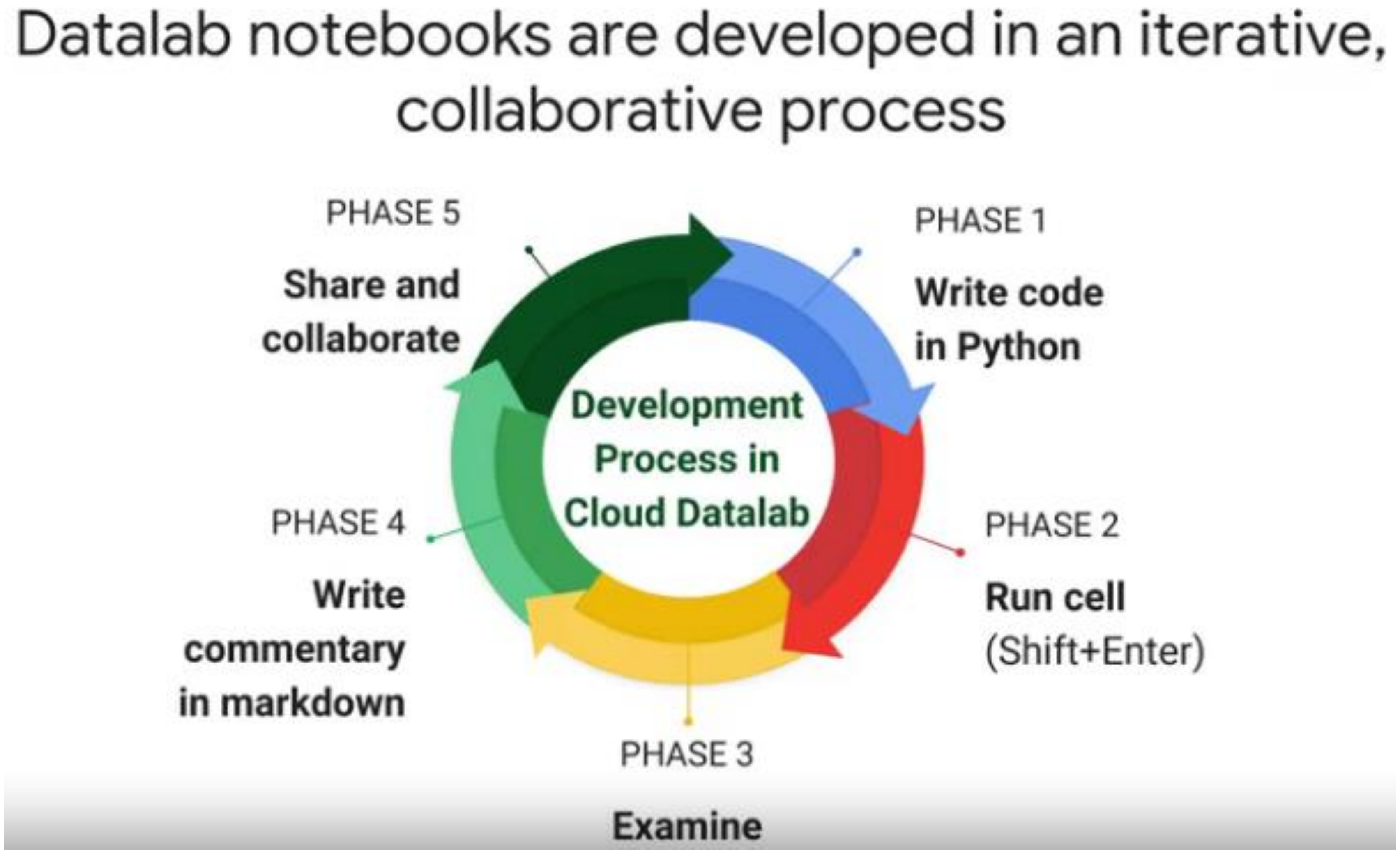
VALIDATE

COST

Analyze query performance

EXPORT

Datalab



Datalab notebooks are developed in an iterative, collaborative process

Development Process in Cloud Datalab

PHASE 1

Write code in Python

PHASE 2

Run cell (Shift+Enter)

PHASE 3

Examine

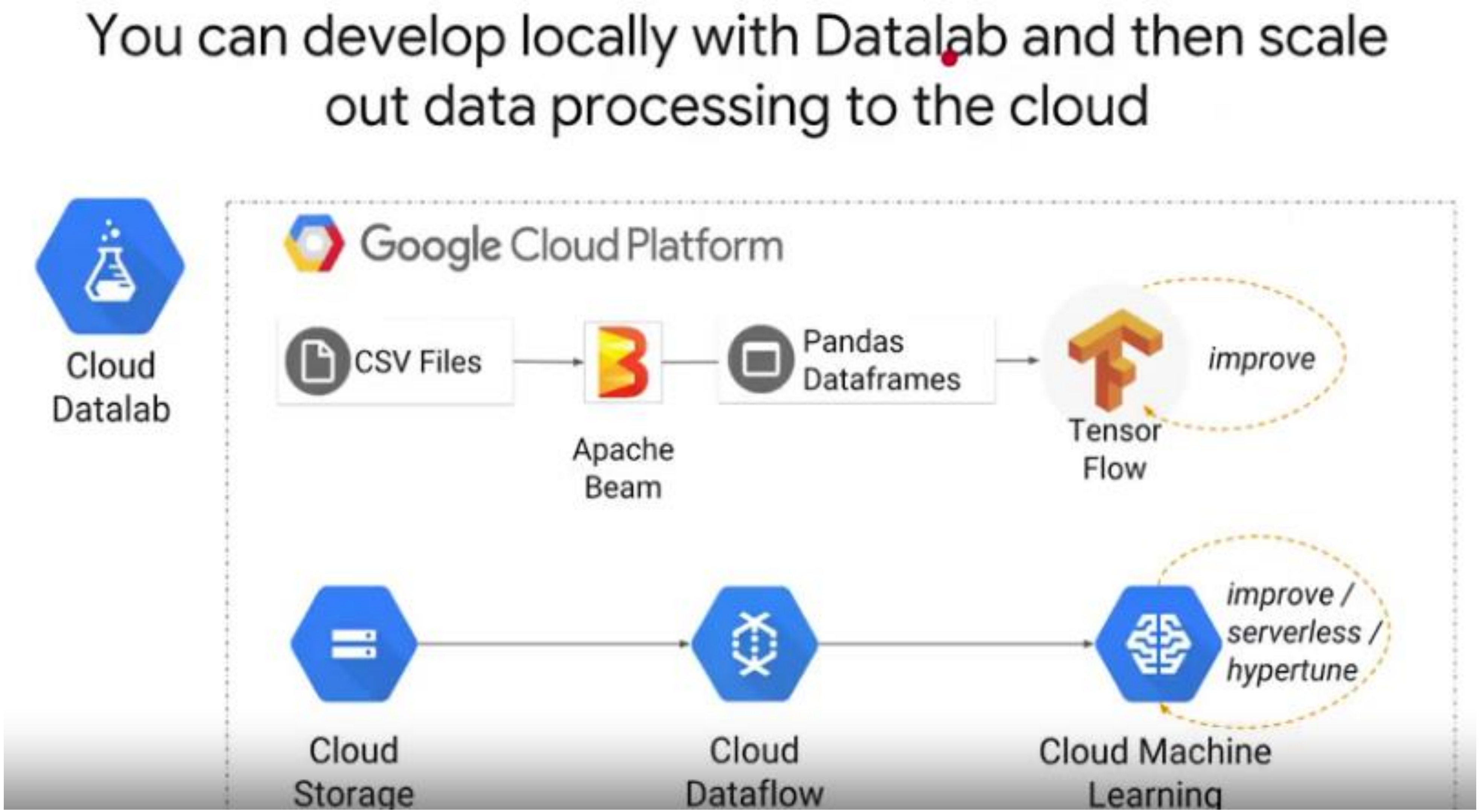
PHASE 4

Write commentary in markdown

PHASE 5

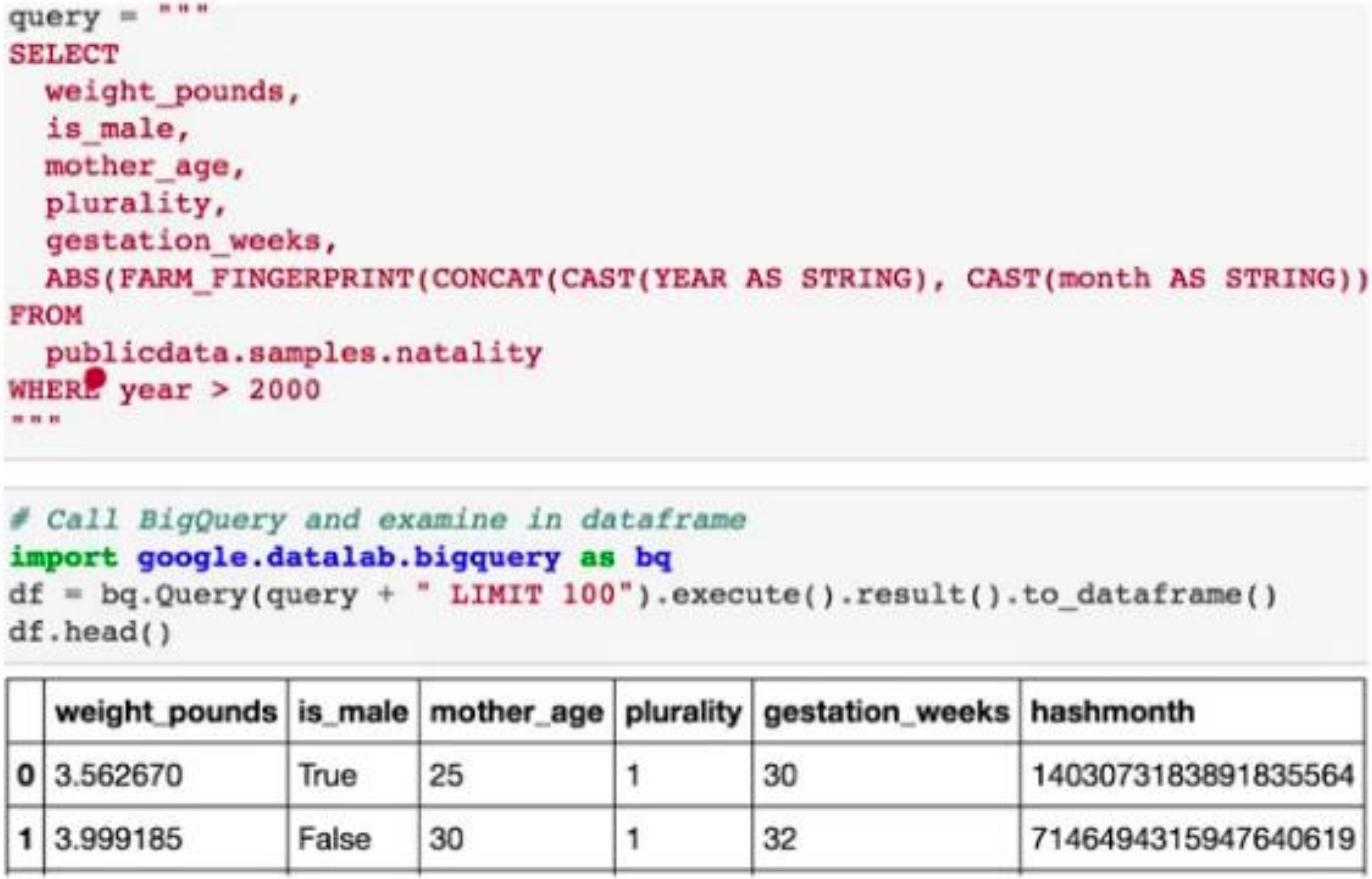
Share and collaborate

Datalab



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

Datalab + BigQuery



```
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
"""
```

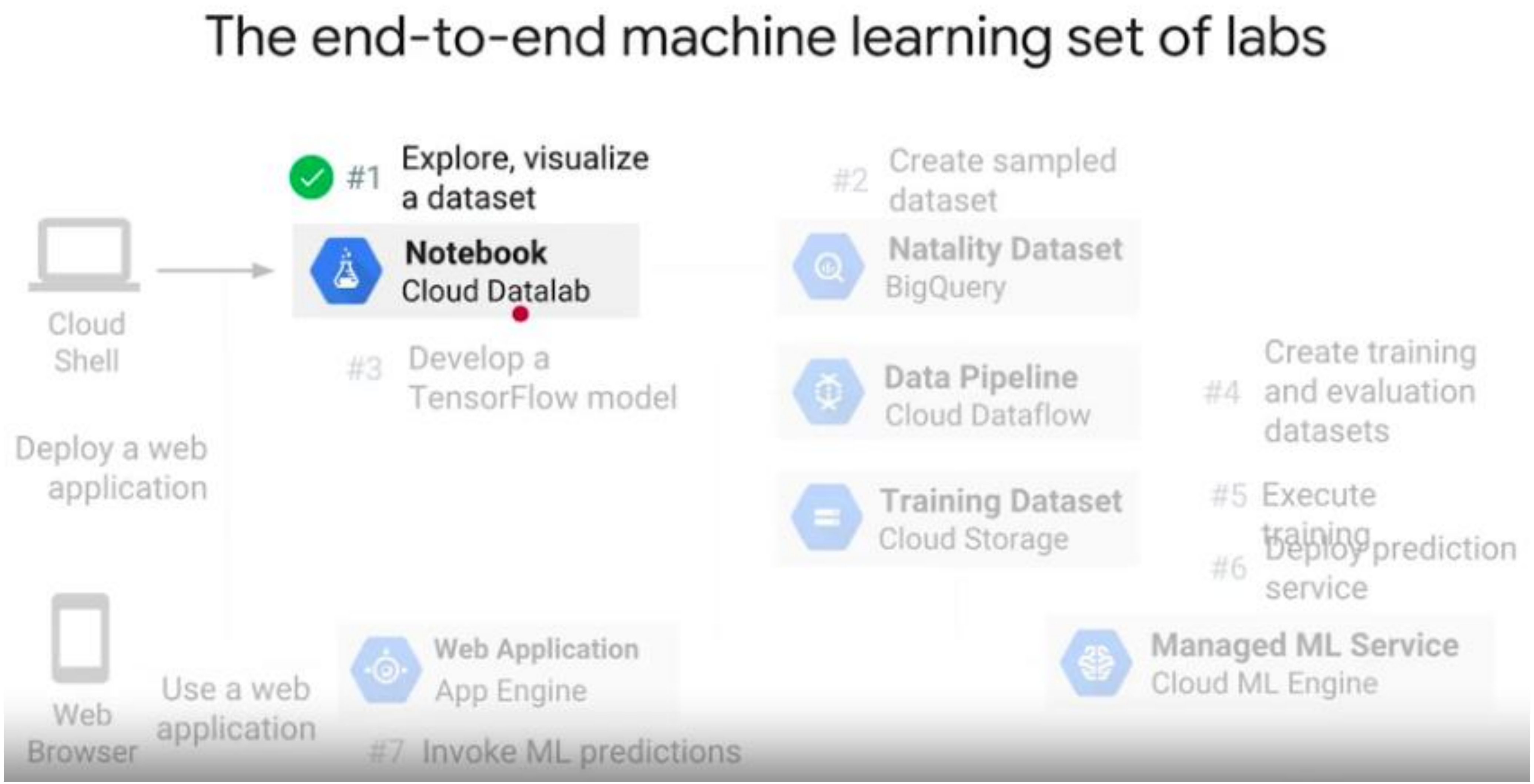
```
# Call BigQuery and examine in dataframe
import google.datalab.bigquery as bq
df = bq.Query(query + " LIMIT 100").execute().result().to_dataframe()
df.head()
```

	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

Lab

Vídeo “03-Explore Dataset” ou acesso ao lab

Onde estamos?



The end-to-end machine learning set of labs