Seminar

July 25th, 2020

DATA SCIENCE INDONESIA

Data Science
Indonesia
East Java Region

# JATIM CAMP #5

Build Data Ecosystem for Better Analytics

## Data Pipelining for Creating Great Data Ecosystem

with

Rendy Bambang Junior
Senior Data Manager, Ruangguru

Muhammad Iqbal Tawakal
Data Scientist, Gojek

Supported by

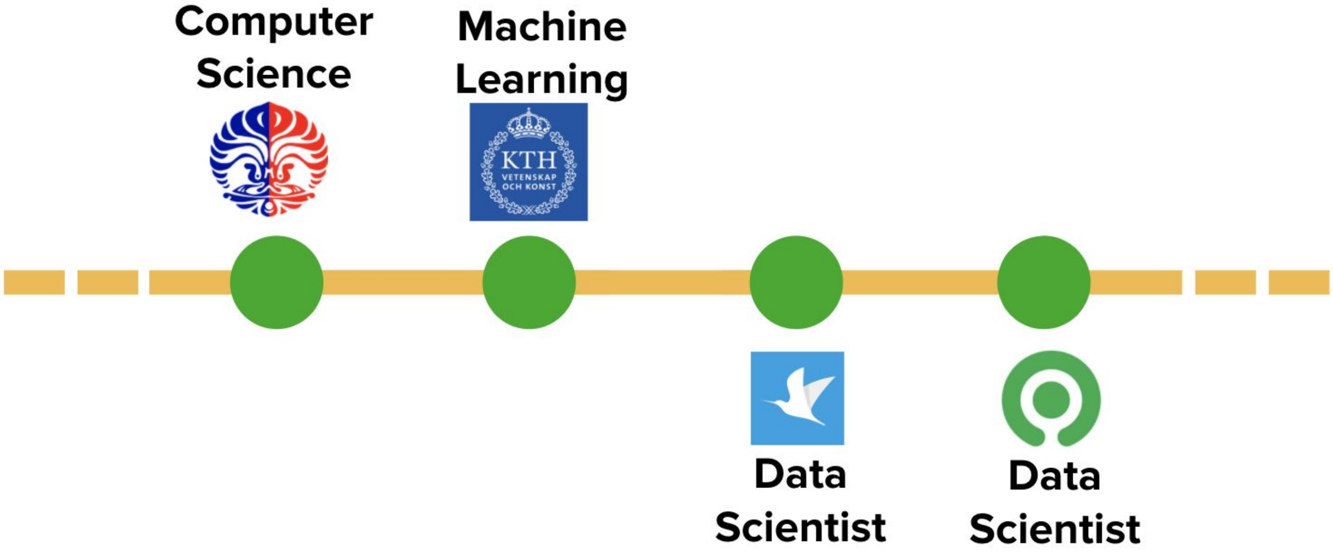
holistics

# Practical Data Pipeline for Data Science

## Case study: BigQuery in Gojek

M Iqbal Tawakal (@mitbal)
DS@gojek

# Self (Re)Introduction



Computer Science

Machine Learning

Data Scientist

Data Scientist

/mitbal

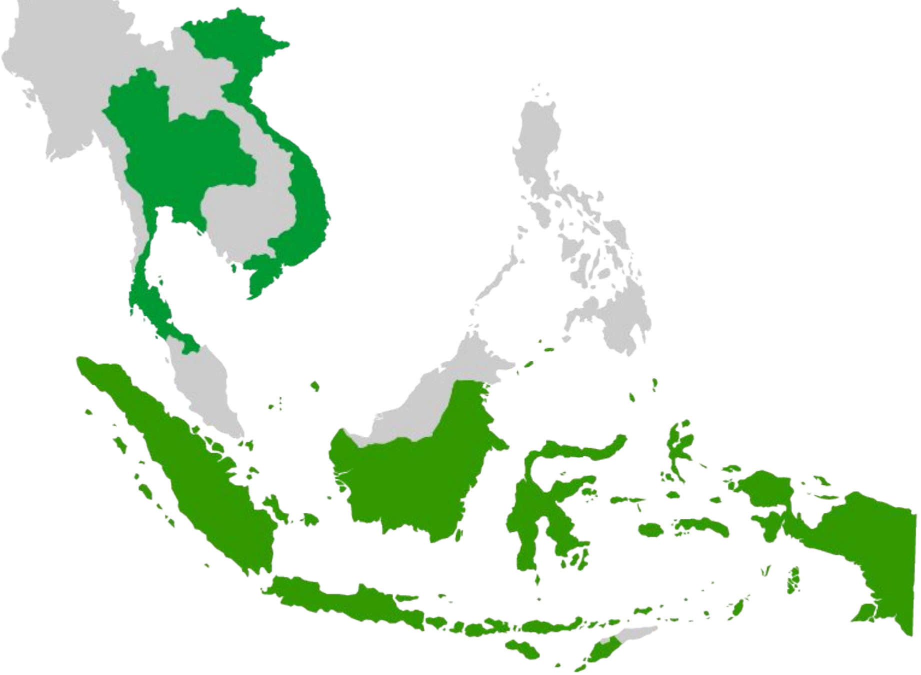
in/mitbal

/@mitbal

# Outline

- Introduction to Gojek, and its DS team
- Intro to DS workflow and challenges faced
- How proper data pipeline can help alleviate them
- How BigQuery fit into all this
- Pros, cons, and other consideration

# Gojek in Southeast Asia



Operates in **207 cities** in **Southeast Asia**

**>155 millions**
**App Downloads**

**+400 thousands**
**Merchant Partners (96% SMEs)**

**+2 millions**
**Driver Partners**

gojek

# Range of products and solutions

goride

gocar

gosend

gobox

gobluebird

gofood

gobuy

goshop

gomed

gotix

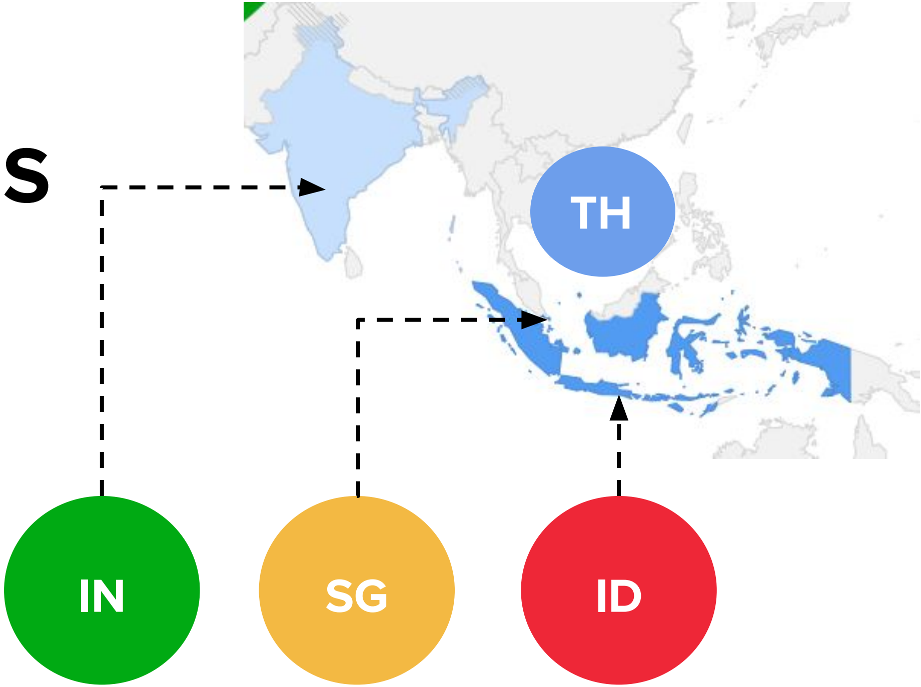
gopay

gopoints

gopulsa

gobills

# Geo @ DS



# Streams @ DS

Pricing

Matchmaking

Supply

Demand

Search

Recs

KYC

C Fraud

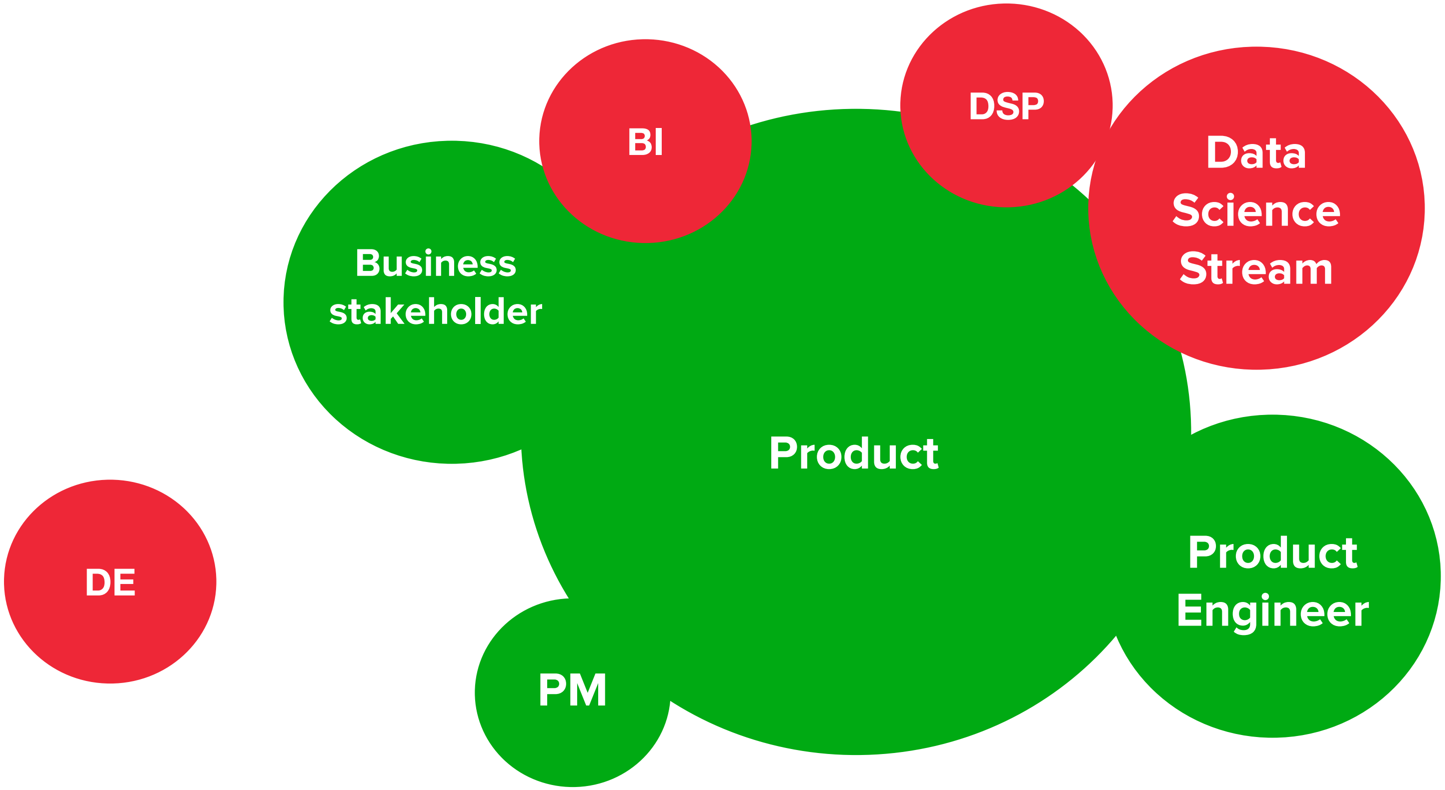
D Fraud

Merchant

Logistic

etc

# DS, Embedded to Product



BI

DSP

Data Science Stream

Business stakeholder

Product

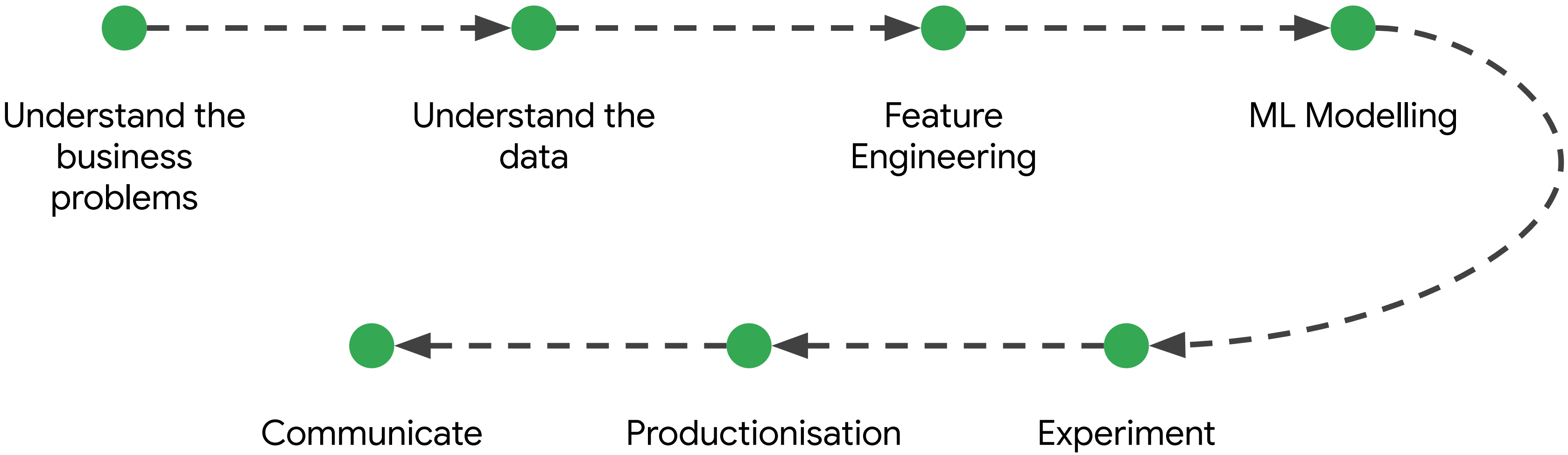
DE

Product Engineer

PM

gojek

# (Ideal) Data Science Timeline and Workflows



Understand the business problems

Understand the data

Feature Engineering

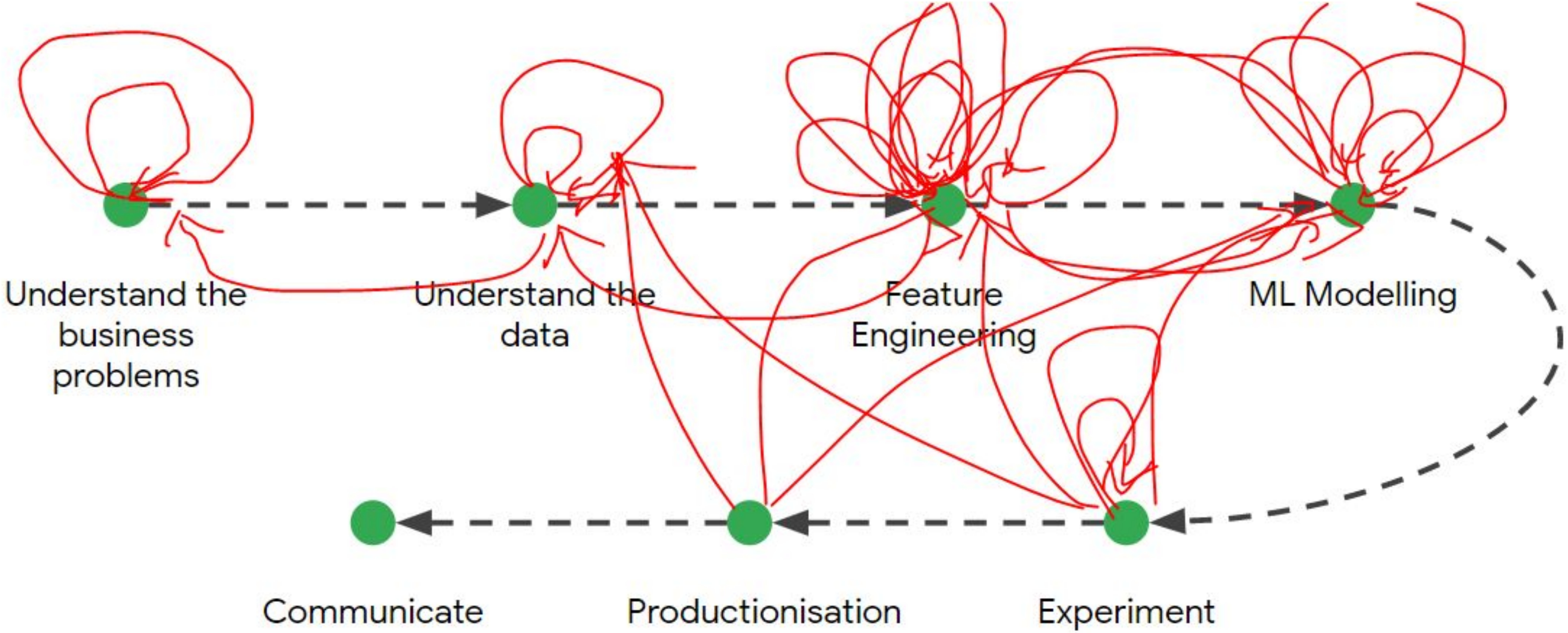
ML Modelling

Experiment

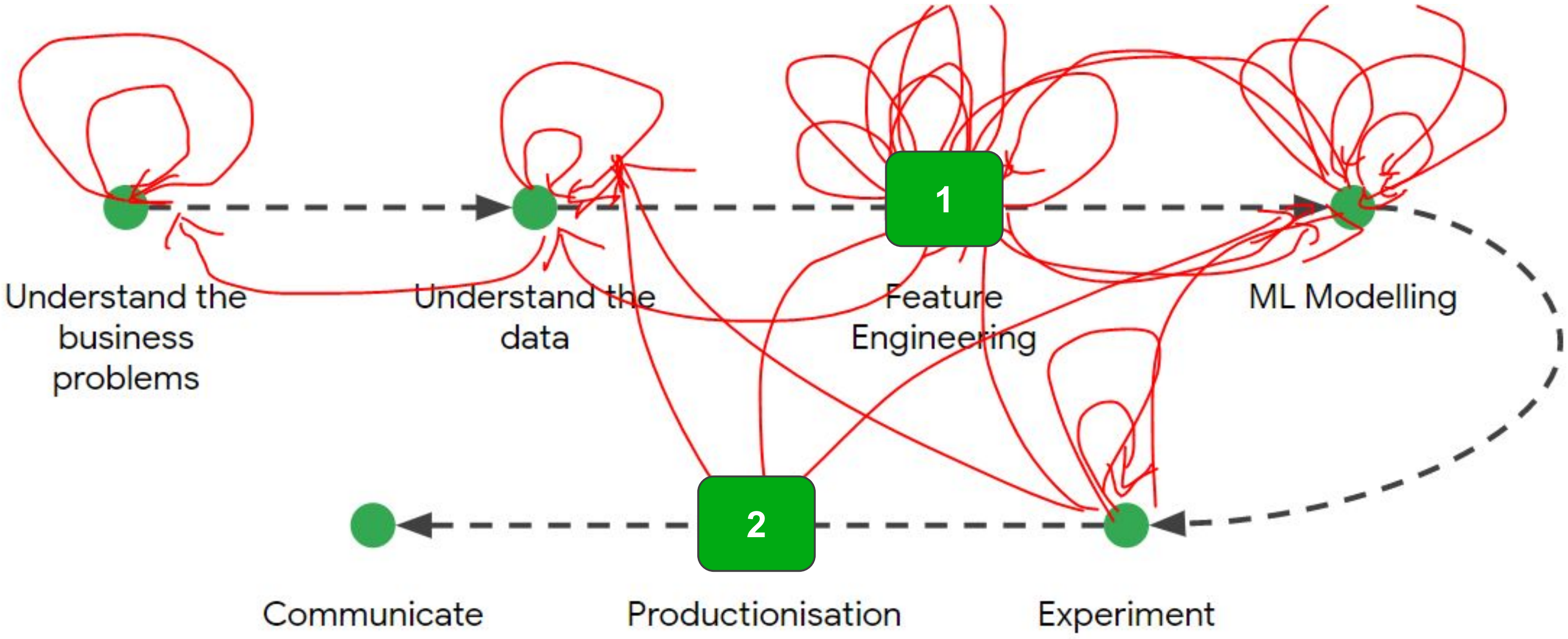
Productionisation

Communicate

# Actual Data Science Workflow!!?!?!

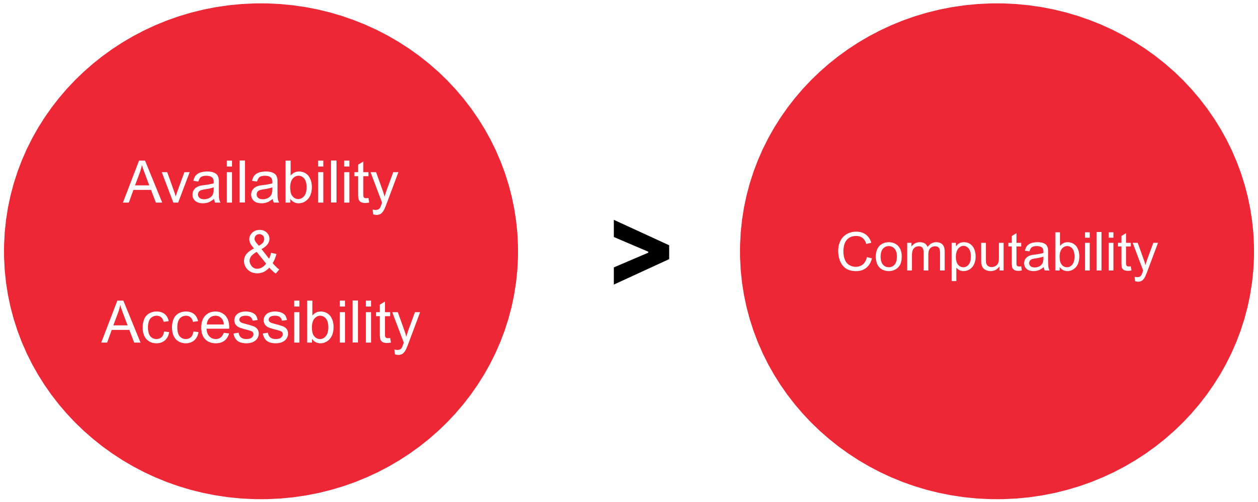


# Where proper DS pipeline can help most



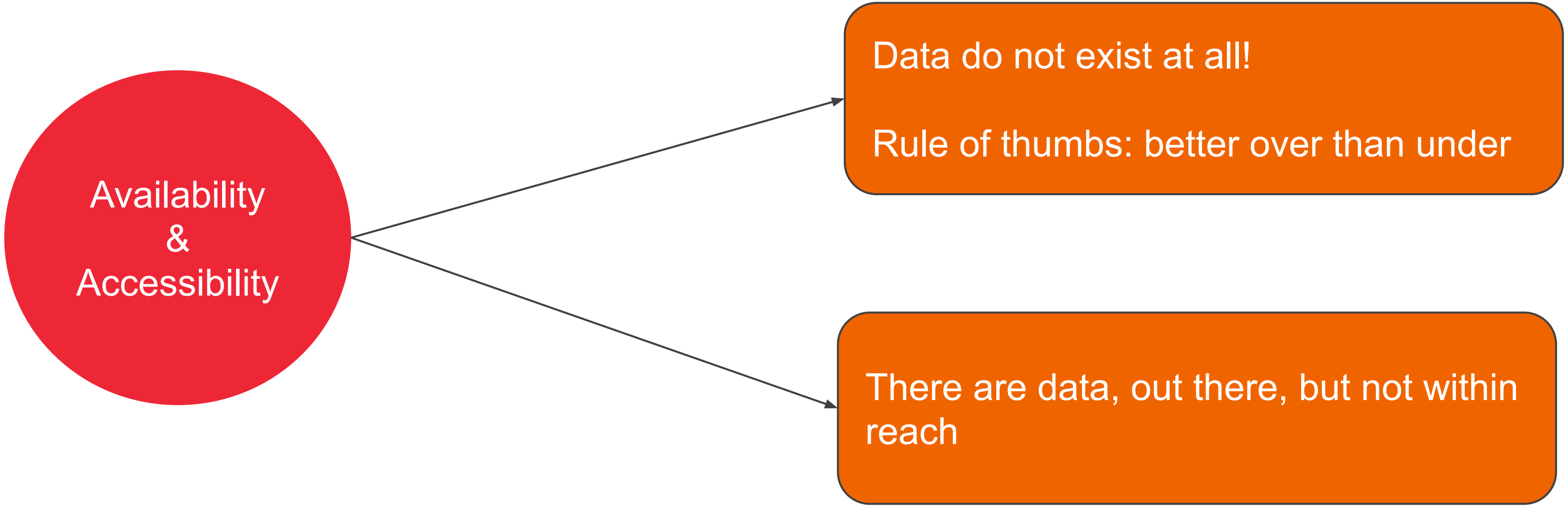
gojek

# 1st part: 2 core challenges



Availability & Accessibility > Computability

# Availability & Accessibility Problem



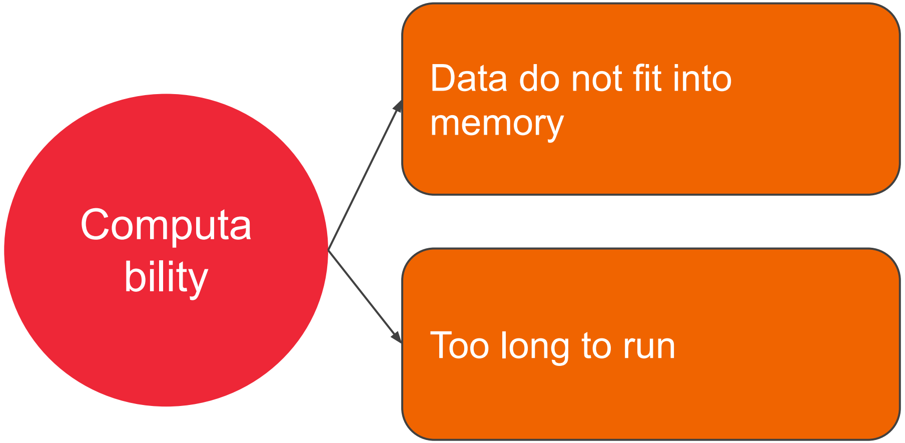
Availability
&
Accessibility

Data do not exist at all!

Rule of thumbs: better over than under

There are data, out there, but not within reach

# Computability Problem

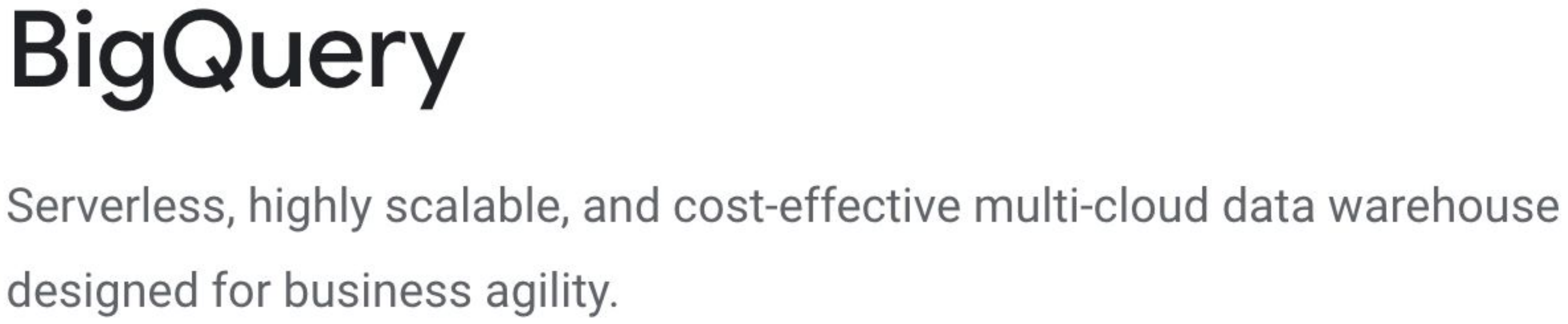


Computability

Data do not fit into memory

Too long to run

# BigQuery come into Picture



BigQuery

Serverless, highly scalable, and cost-effective multi-cloud data warehouse designed for business agility.

# Not the only player in town

amazon REDSHIFT

TERADATA

ORACLE DATA WAREHOUSE

Data warehouse alternative

Processing engine alternative

APACHE Spark™

hadoop

presto

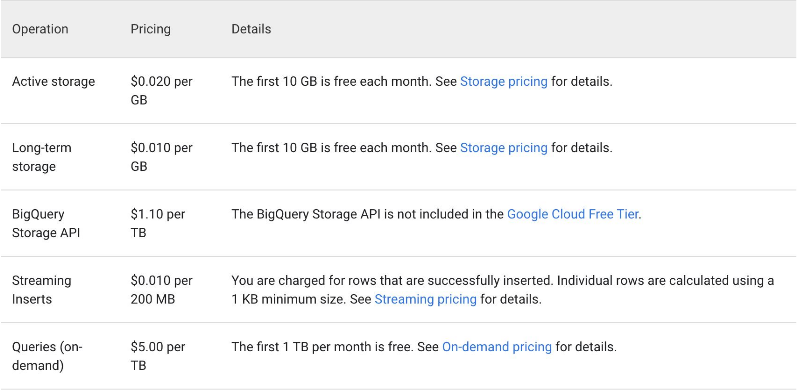
# For your consideration

## Pros

- combine storage & processing engine in one package
- no need to manage physical or virtual instance
- computation can scale with data size automatically
- using SQL as the interface language to interact

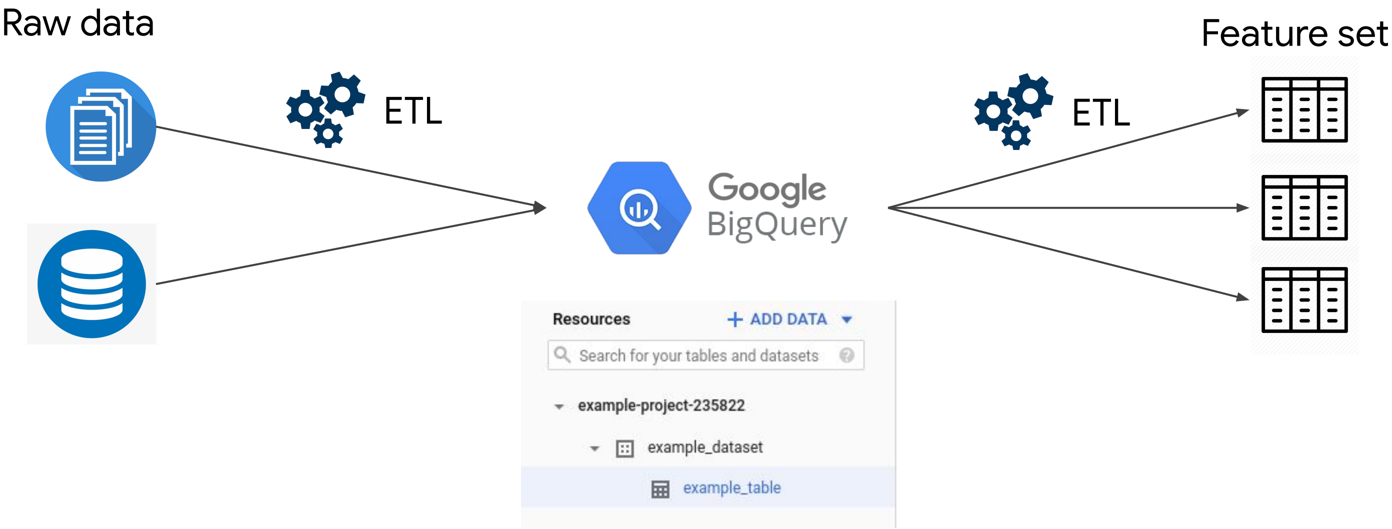
## Cons

- can be costly if usage is not monitored properly
- might not suited for certain types of operation



| Operation | Pricing | Details |
|---|---|---|
| Active storage | $0.020 per GB | The first 10 GB is free each month. See Storage pricing for details. |
| Long-term storage | $0.010 per GB | The first 10 GB is free each month. See Storage pricing for details. |
| BigQuery Storage API | $1.10 per TB | The BigQuery Storage API is not included in the Google Cloud Free Tier. |
| Streaming Inserts | $0.010 per 200 MB | You are charged for rows that are successfully inserted. Individual rows are calculated using a 1 KB minimum size. See Streaming pricing for details. |
| Queries (on-demand) | $5.00 per TB | The first 1 TB per month is free. See On-demand pricing for details. |

# Accessibility solved?: 1 place to put them all



Resources + ADD DATA

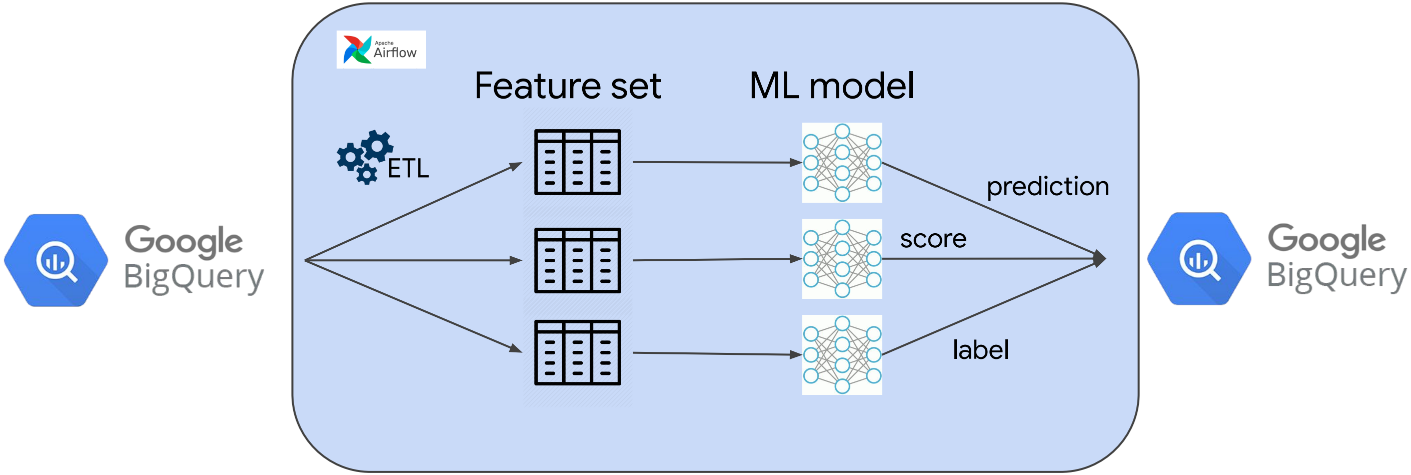
Search for your tables and datasets

example-project-235822

example_dataset

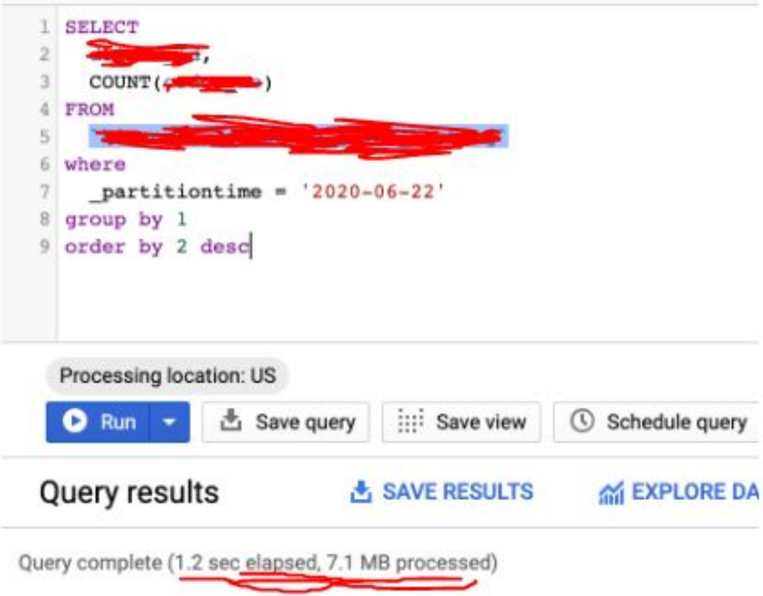
example_table

# Batch deployment option



# Computability solved?: Compute at any scale, up and down

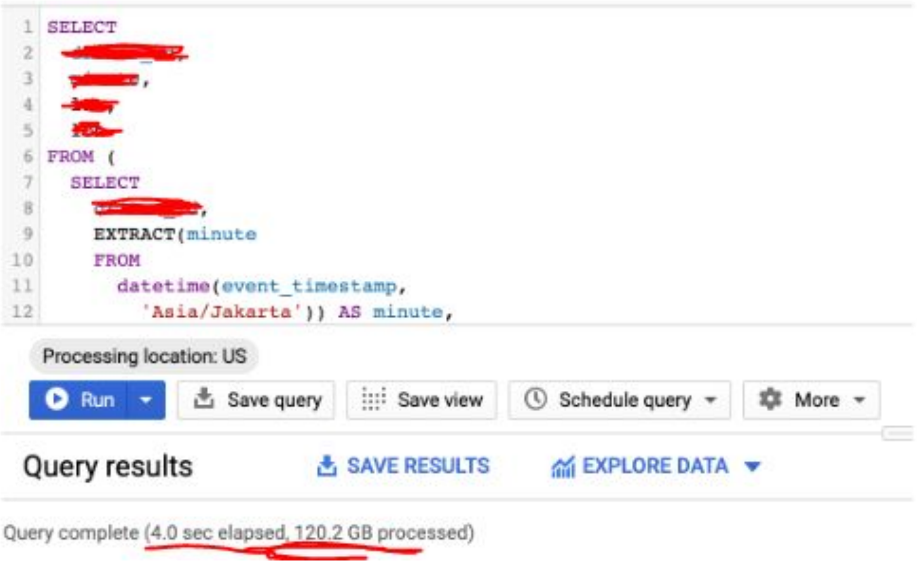
● Small OK



```
SELECT
  [redacted],
  COUNT([redacted])
FROM
  [redacted]
where
  _partitiontime = '2020-06-22'
group by 1
order by 2 desc
```

Query complete (1.2 sec elapsed, 7.1 MB processed)

● Big No problem



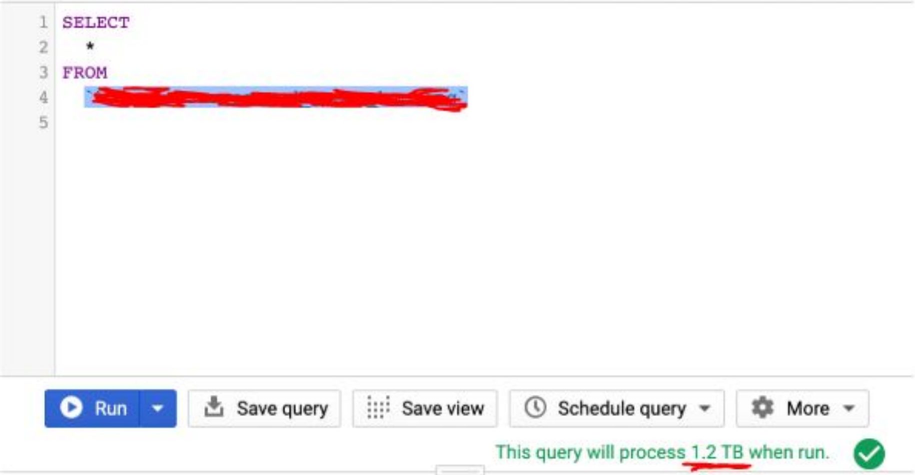
```
SELECT
  [redacted]
  [redacted],
  [redacted]
  [redacted]
FROM (
  SELECT
    [redacted],
    EXTRACT(minute
    FROM
      datetime(event_timestamp,
        'Asia/Jakarta')) AS minute,
```

Query complete (4.0 sec elapsed, 120.2 GB processed)

# Best Practice: Do and Don't (1)

- Never select *, explicitly choose field and attribute to be included into processing

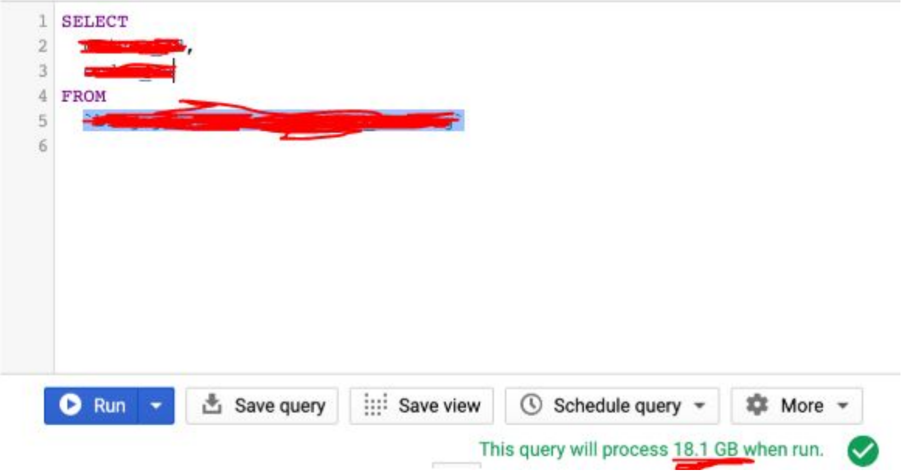
● Before



```
SELECT
    *
FROM
  [redacted]
```

This query will process 1.2 TB when run. ✓

● After



```
SELECT
  [redacted],
  [redacted]
FROM
  [redacted]
```

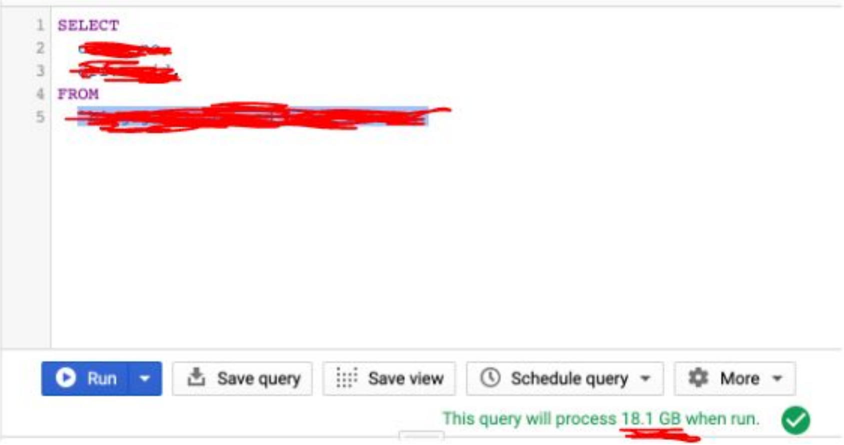
This query will process 18.1 GB when run. ✓

gojek

# Best Practice: Do and Don’t (2)

- Always filter by partition date

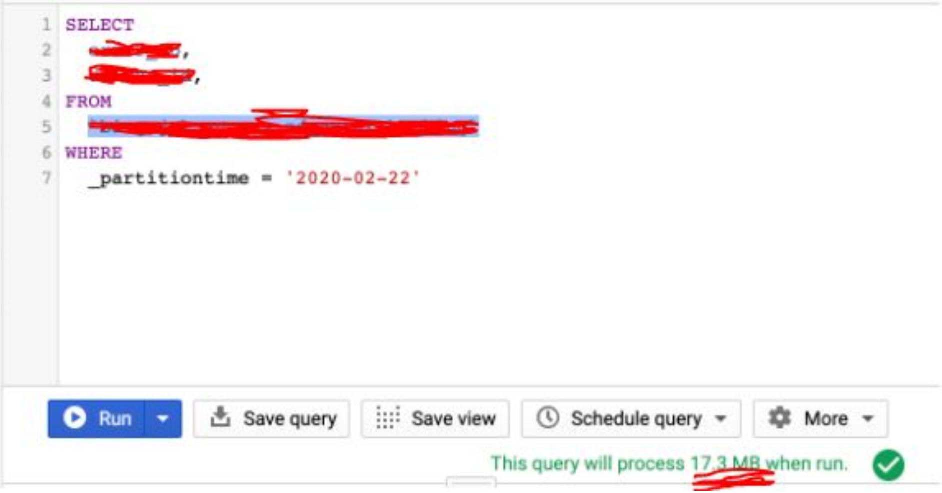
● Before



```
SELECT
  [redacted]
FROM
  [redacted]
```

This query will process 18.1 GB when run.

● After



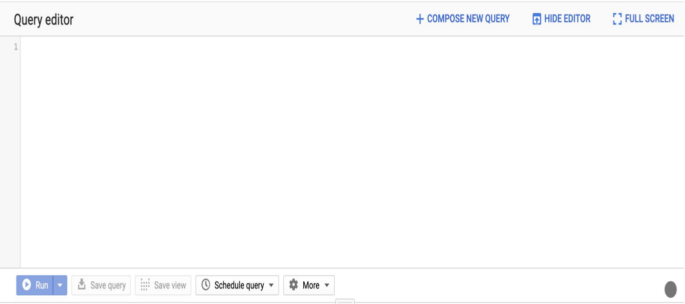
```
SELECT
  [redacted]
FROM
  [redacted]
WHERE
  _partitiontime = '2020-02-22'
```

This query will process 17.3 MB when run.

gojek

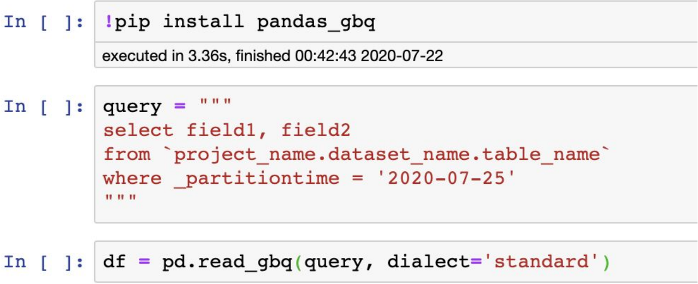
# How to use as data scientist

## Web UI



- Autoformat
- Syntax highlighting
- Save result

## Programmatically



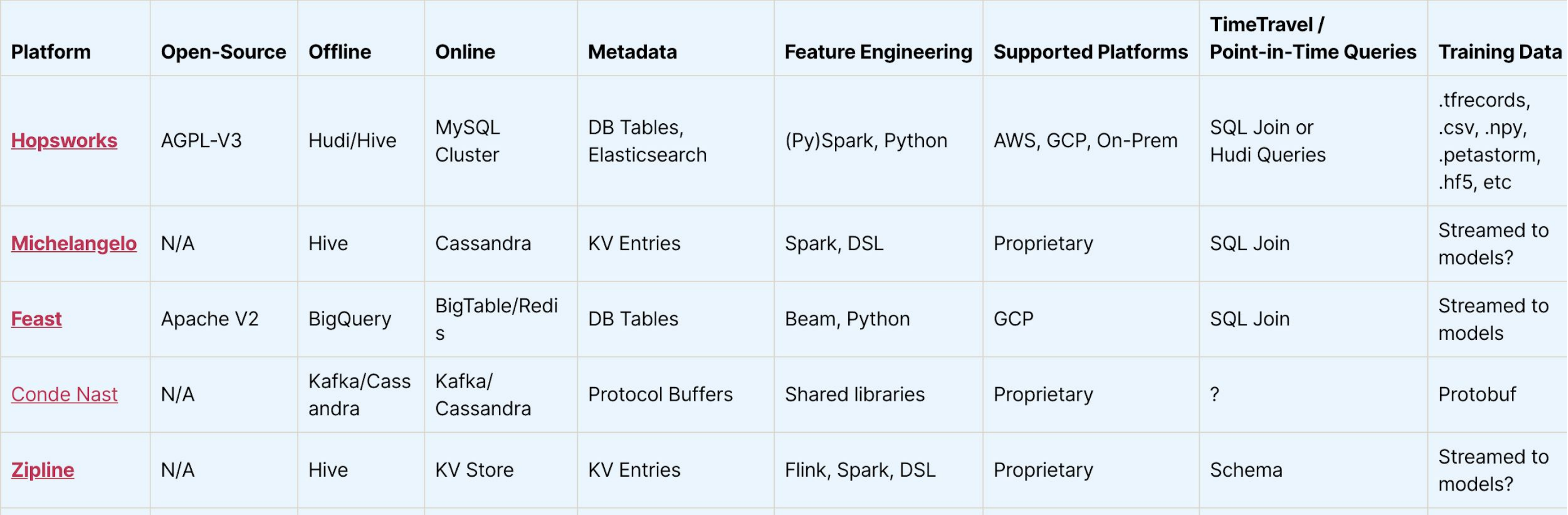
```
In [ ]: !pip install pandas_gbq
        executed in 3.36s, finished 00:42:43 2020-07-22

In [ ]: query = """
        select field1, field2
        from `project_name.dataset_name.table_name`
        where _partitiontime = '2020-07-25'
        """

In [ ]: df = pd.read_gbq(query, dialect='standard')
```

- Need to setup authentication first
- Seamlessly used in the next processing steps

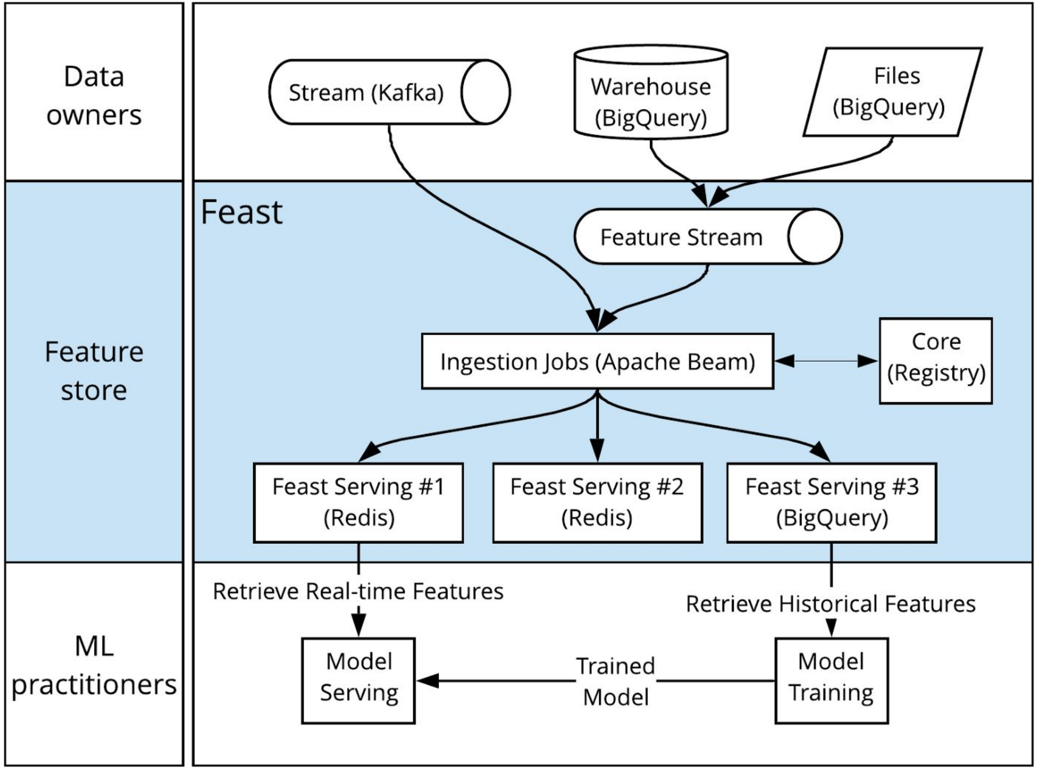
# Going Beyond: Feature engineering reusability and serving with Feature Storage



| Platform | Open-Source | Offline | Online | Metadata | Feature Engineering | Supported Platforms | TimeTravel / Point-in-Time Queries | Training Data |
|---|---|---|---|---|---|---|---|---|
| Hopsworks | AGPL-V3 | Hudi/Hive | MySQL Cluster | DB Tables, Elasticsearch | (Py)Spark, Python | AWS, GCP, On-Prem | SQL Join or Hudi Queries | .tfrecords, .csv, .npy, .petastorm, .hf5, etc |
| Michelangelo | N/A | Hive | Cassandra | KV Entries | Spark, DSL | Proprietary | SQL Join | Streamed to models? |
| Feast | Apache V2 | BigQuery | BigTable/Redis | DB Tables | Beam, Python | GCP | SQL Join | Streamed to models |
| Conde Nast | N/A | Kafka/Cassandra | Kafka/ Cassandra | Protocol Buffers | Shared libraries | Proprietary | ? | Protobuf |
| Zipline | N/A | Hive | KV Store | KV Entries | Flink, Spark, DSL | Proprietary | Schema | Streamed to models? |

source http://featurestore.org/

# FEAST: Gojek own Feature Storage solution



https://github.com/feast-dev/feast

gojek

# Closing Remark

1 Data Science and Machine Learning project workflow rarely work in straightforward manner. Good data pipeline is almost become necessity for it to be successful

2 BigQuery can be used to fill this shoes. However, with big power comes big responsibility. Is it the only final solution? No, but it’s a good start

3 Need strong data foundation pipeline ready first

4 Data science is a team sport. All of these cannot happen without, the biz with their infinite wisdom, DSP with their cool tools, product engineering for their system, and BI for maintaining original data pipeline

# Thank You!

gojek