ข้อตกลงการเรียน: จดได้ ถามได้

หาที่จดไว้ ไม่ลืมแน่นอน

ระหว่างเรียน ถามคำถามได้ตลอดเวลาใน slido.com

ตอนจบแต่ละ Section จะมีเวลาให้ถามคำถาม

slido

Join at

slido.com

#3557 928

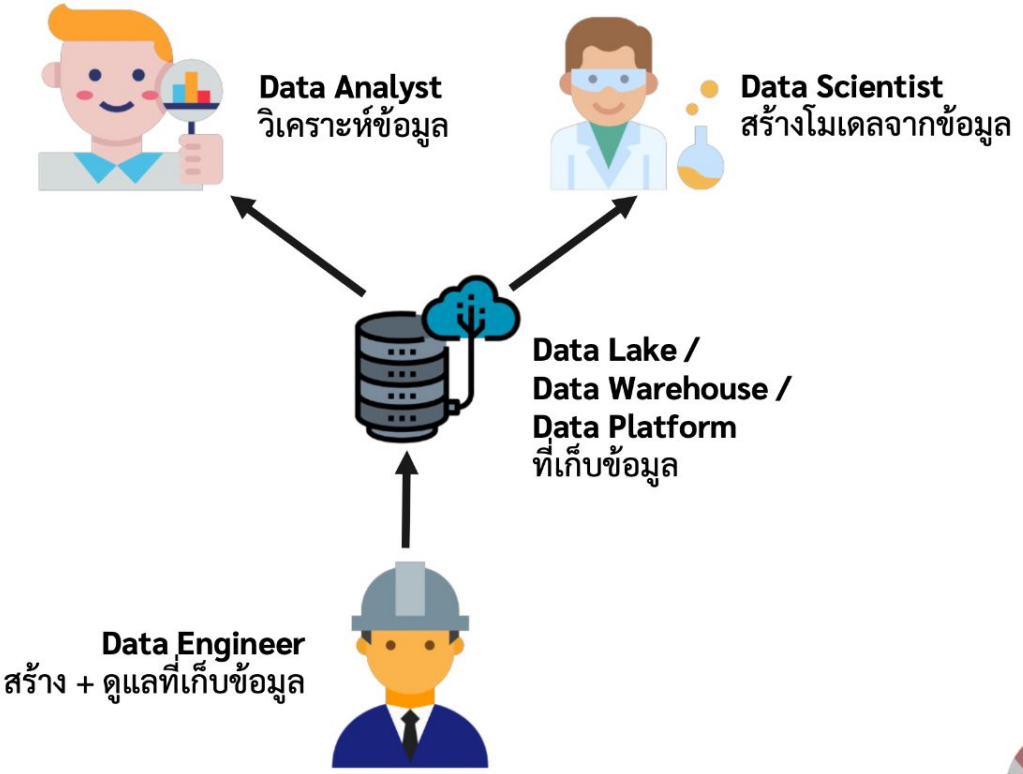
Topic

1. Introduction to Data Engineer
2. Data Engineer Challenge
3. Programming Language (as Data Engineer)
4. Data Pipeline and ETL
5. Example of Google Cloud Platform products

The Role of a Data Engineer

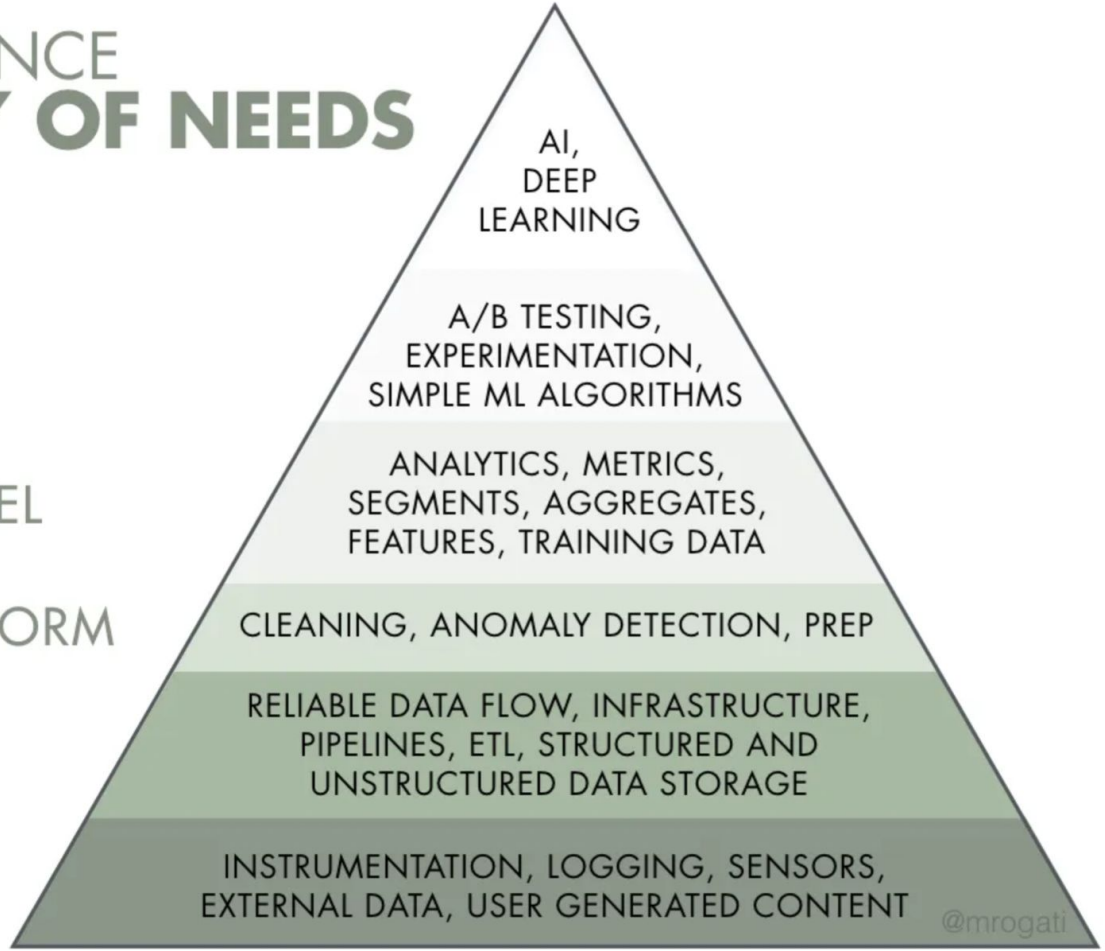
World of Data Science

โลก Data Science มีคน
อยู่ 3 ประเภท



DataTH

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

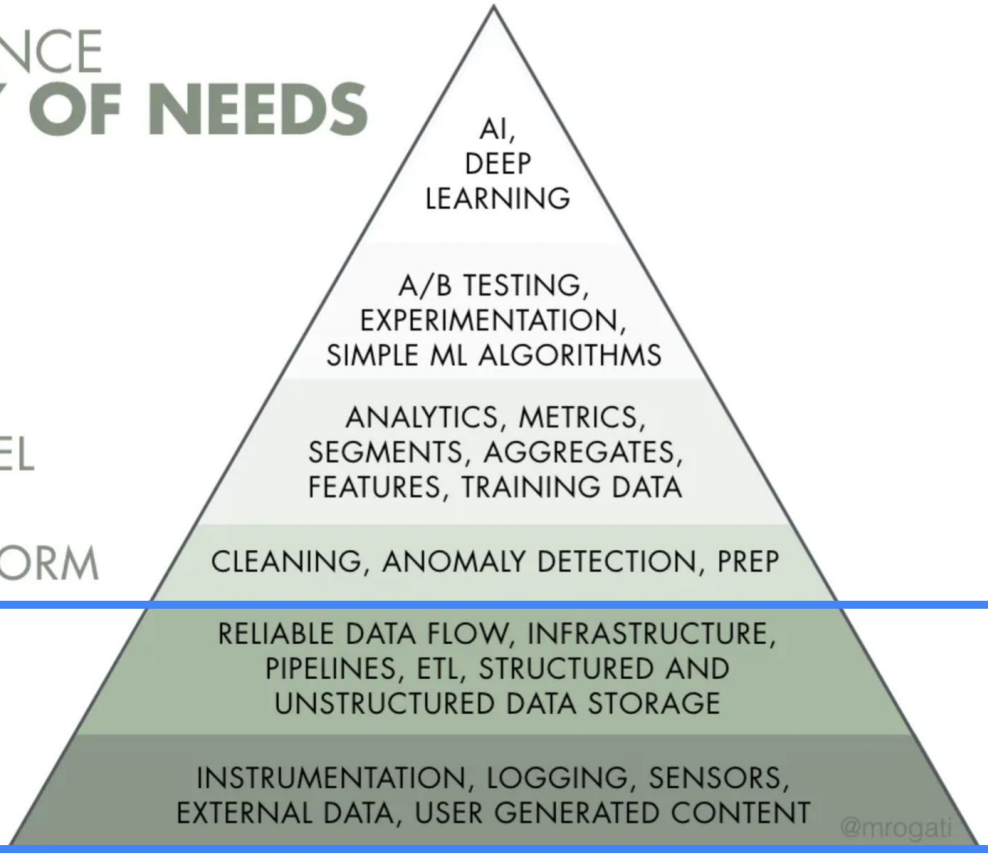
AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

AGGREGATE/LABEL

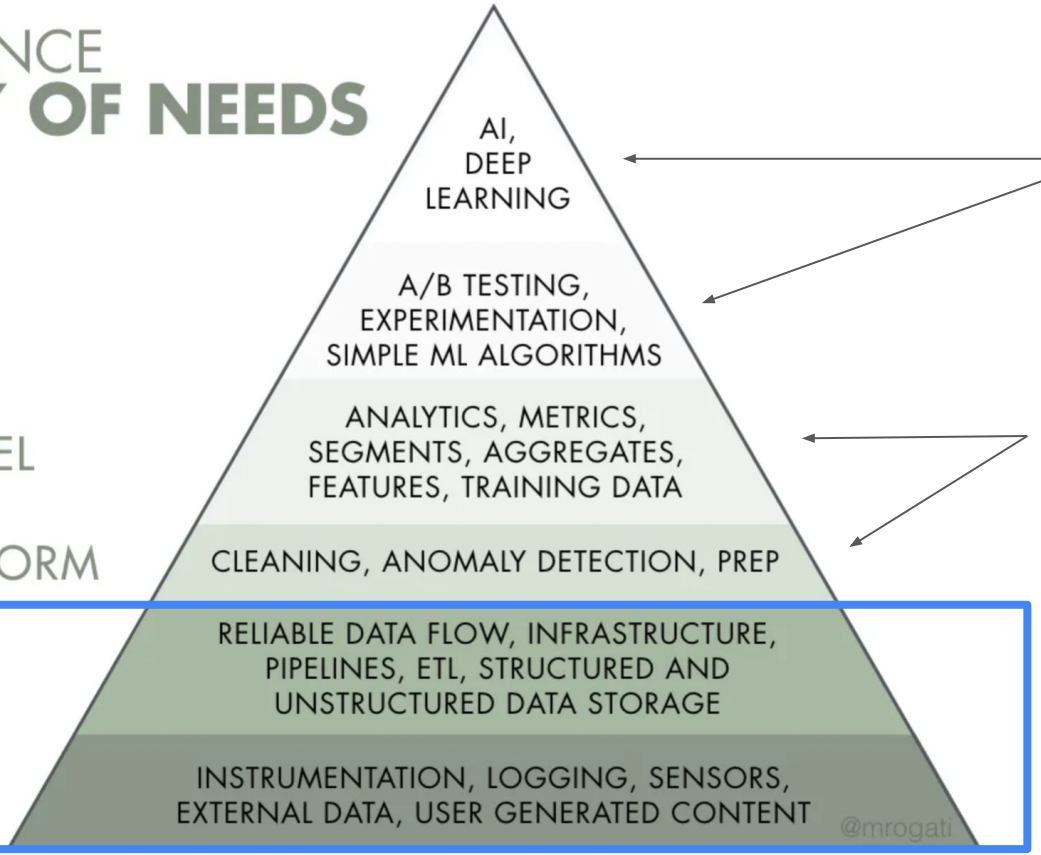
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

Data Engineer

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

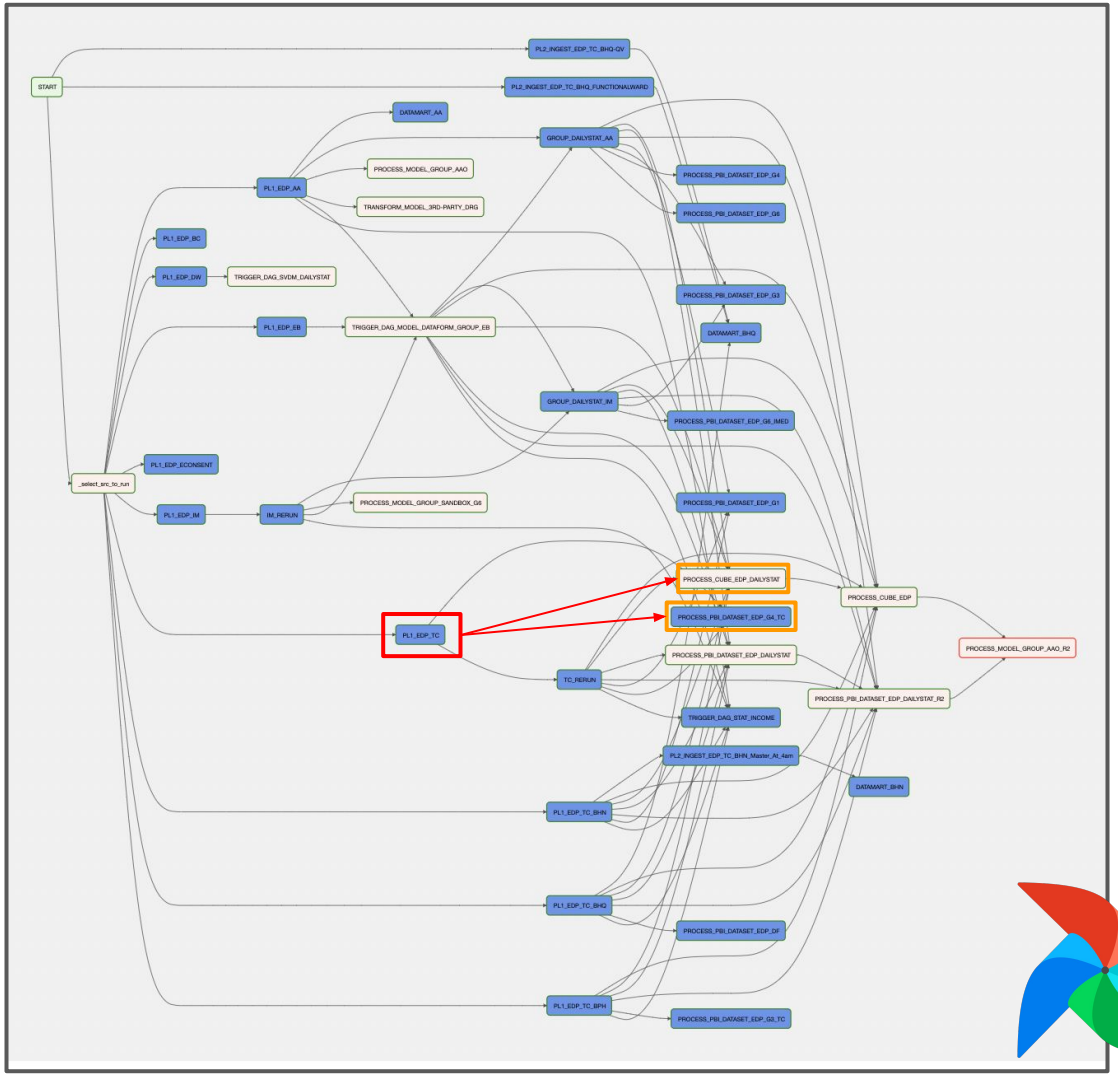
MOVE/STORE

COLLECT

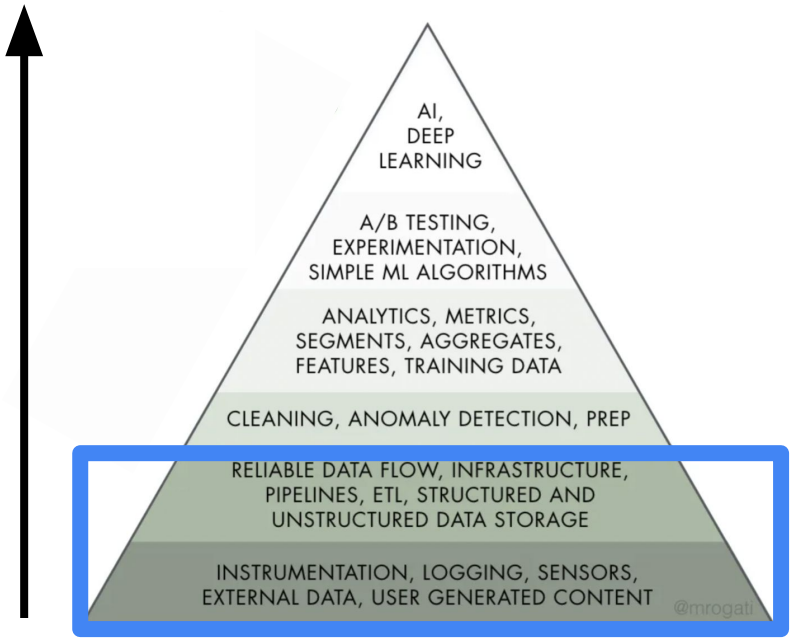
Data Scientist

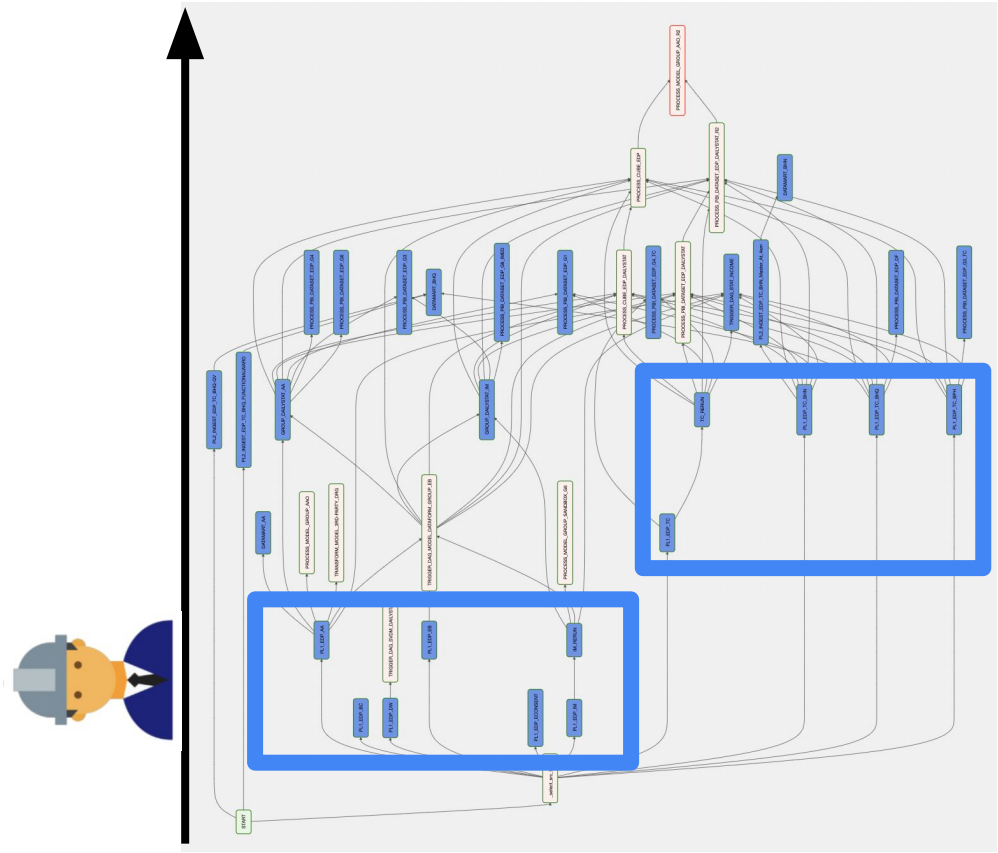
Data Analyst

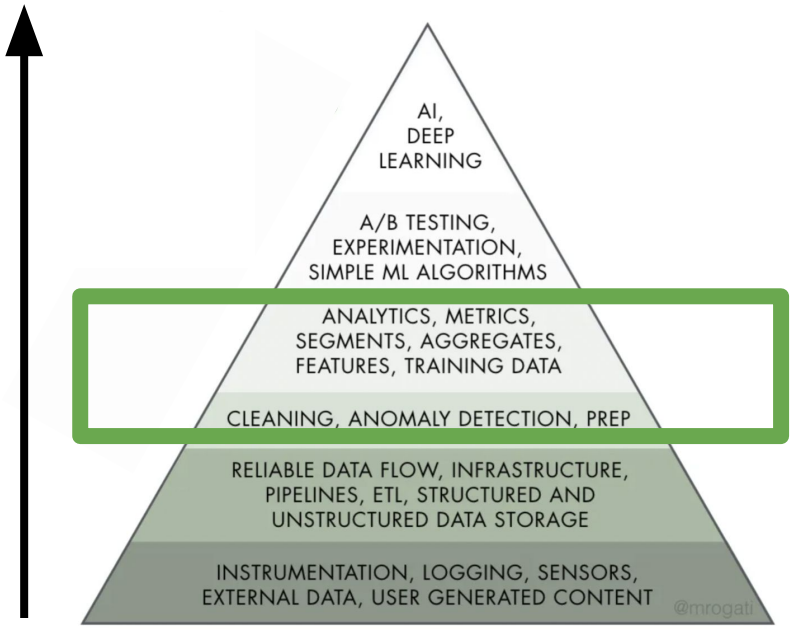
Data Engineer

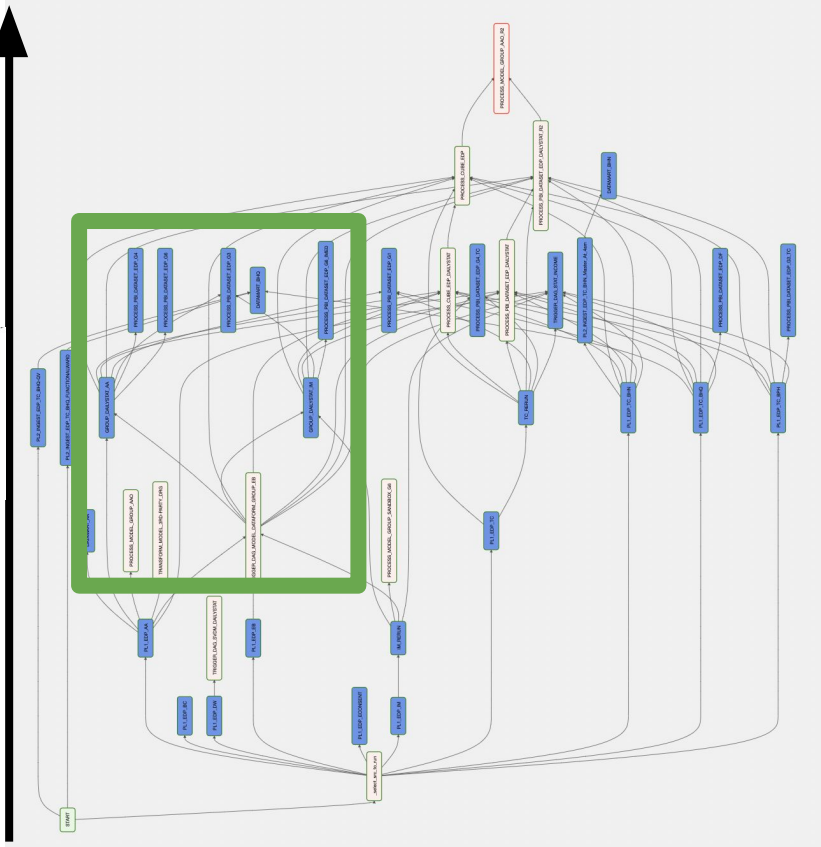


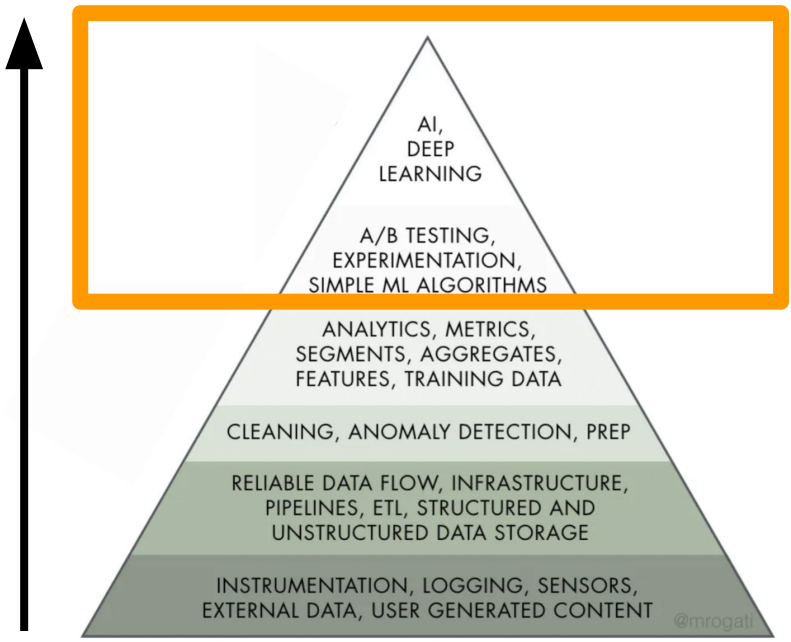
Apache Airflow


AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

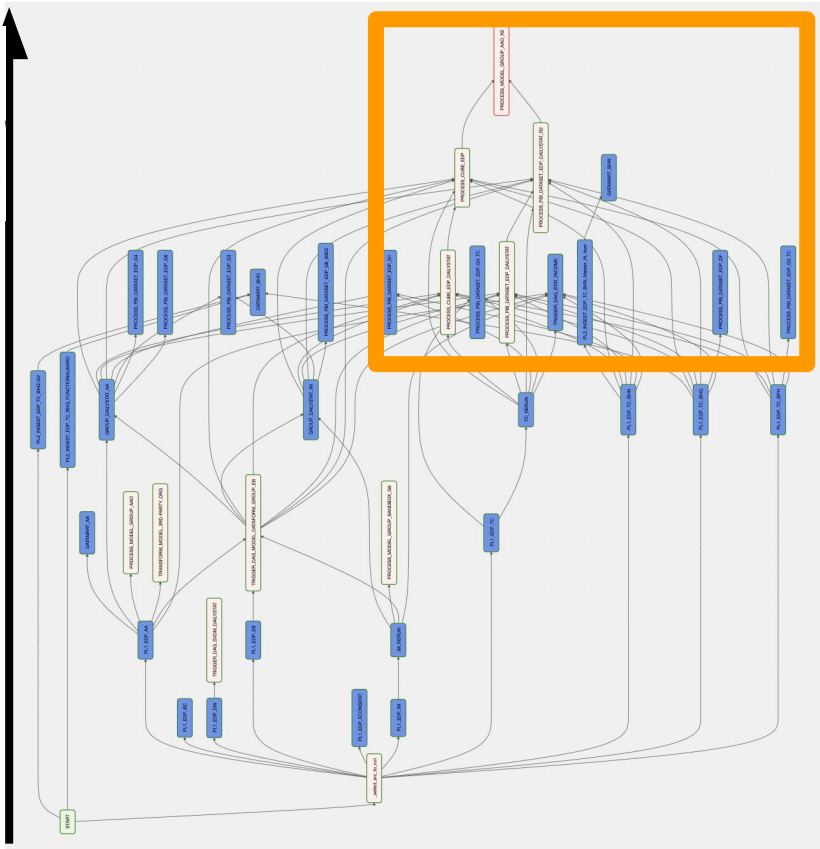



AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT
@mrogati



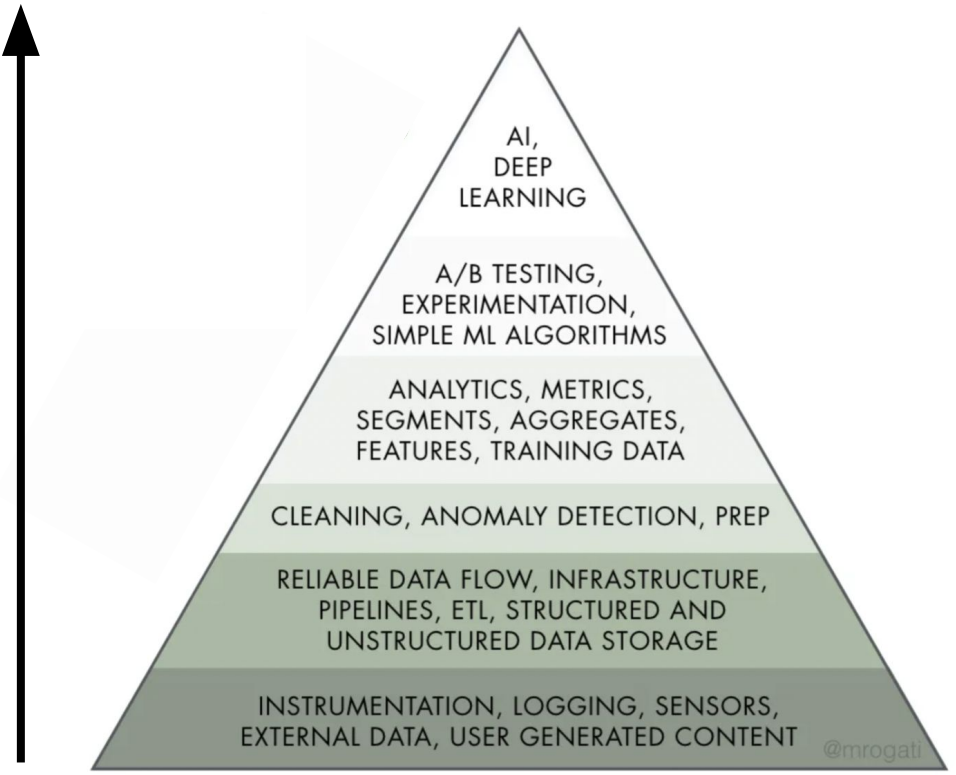

AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

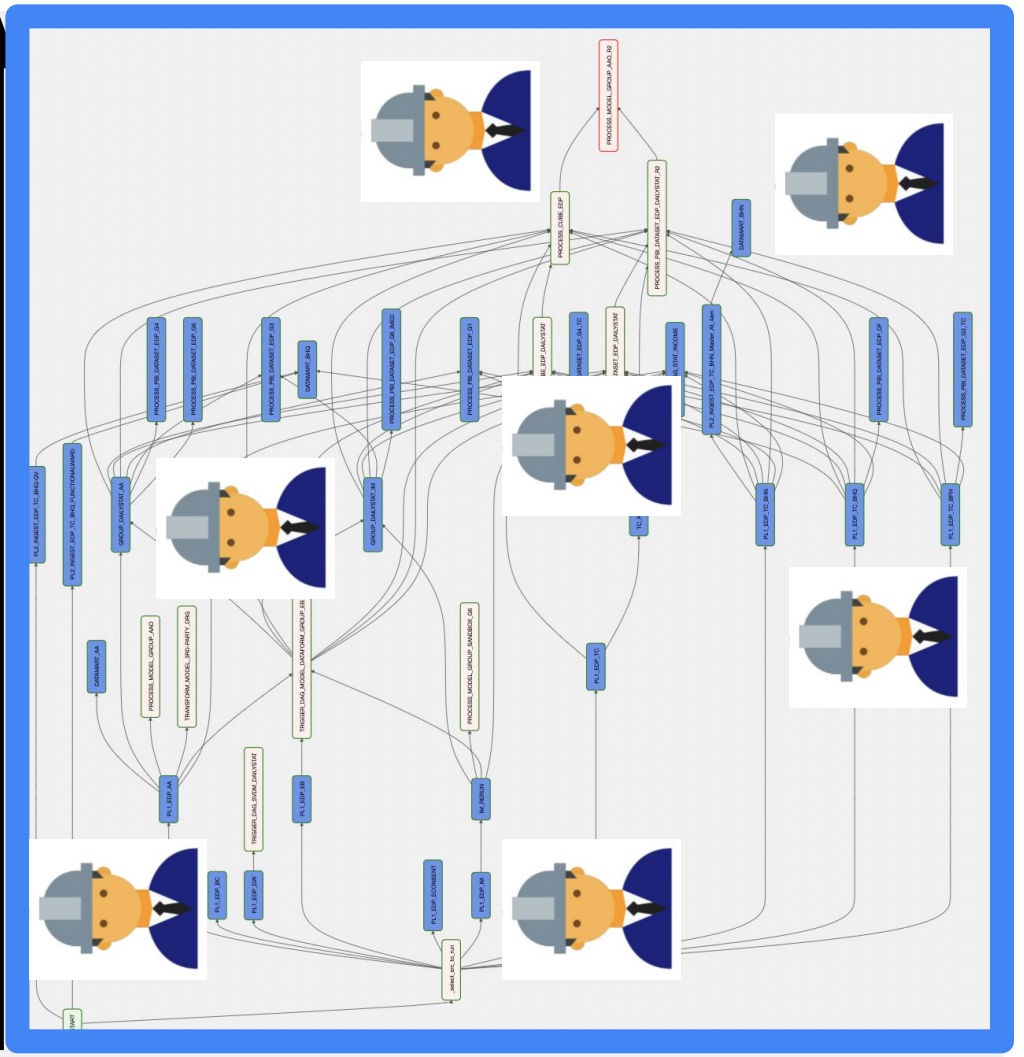

Focus



Focus

AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT
RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE
@mrogati


AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT
@mrogati



งาน Data Engineer ในไทย

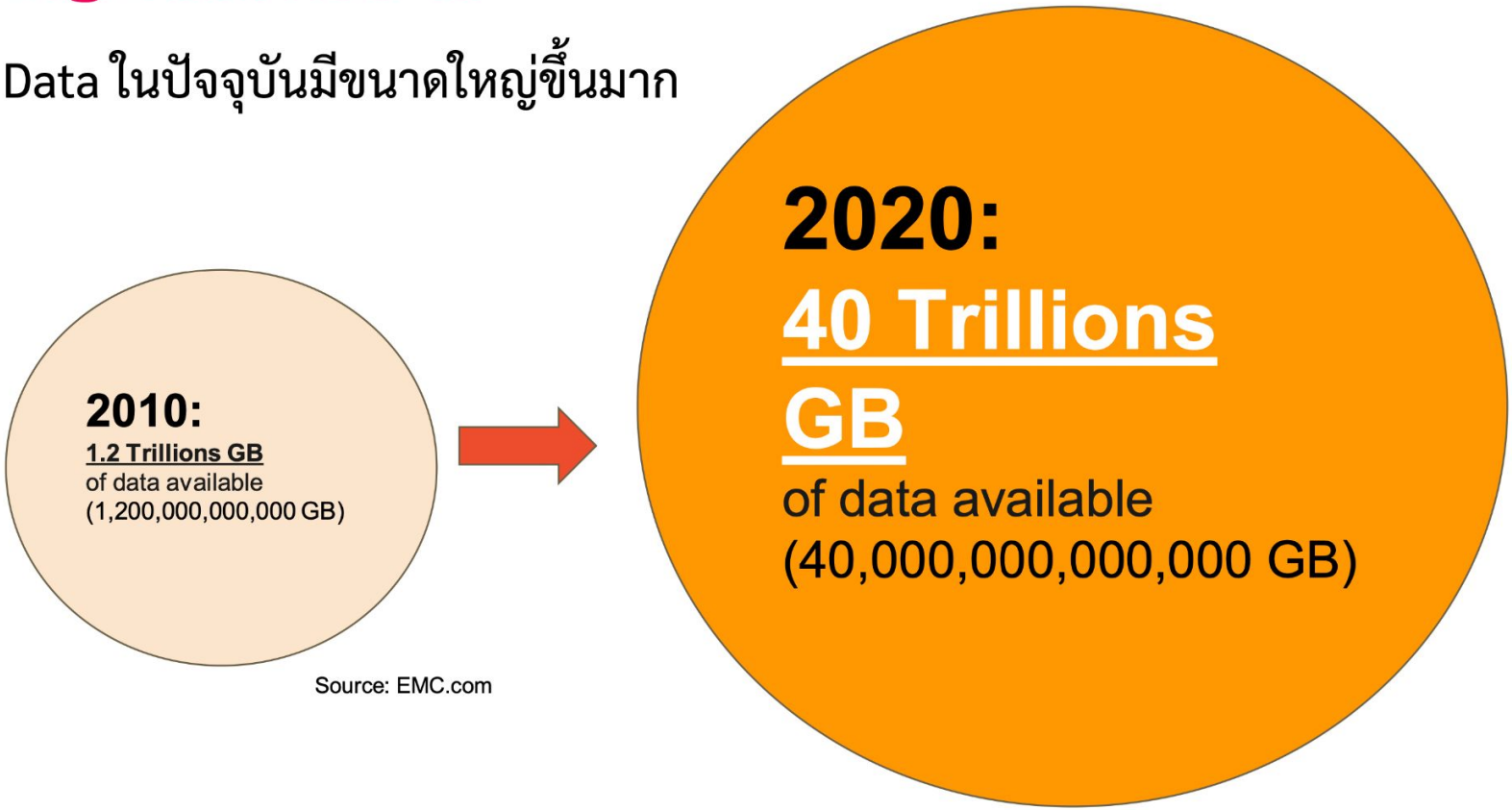
ชื่อตำแหน่งงานใกล้เคียง: Cloud Engineer, DevOps Engineer, Data Architect, System Engineer ฯลฯ

องค์กรใหญ่ ๆ ที่มีข้อมูลเยอะ จำเป็นต้องมีคนที่ดูแลงานด้านนี้
เช่น บริษัทในตลาดหุ้น, การเงิน ธนาคาร, บริษัทประกัน, E-Commerce

Data Engineering Challenges

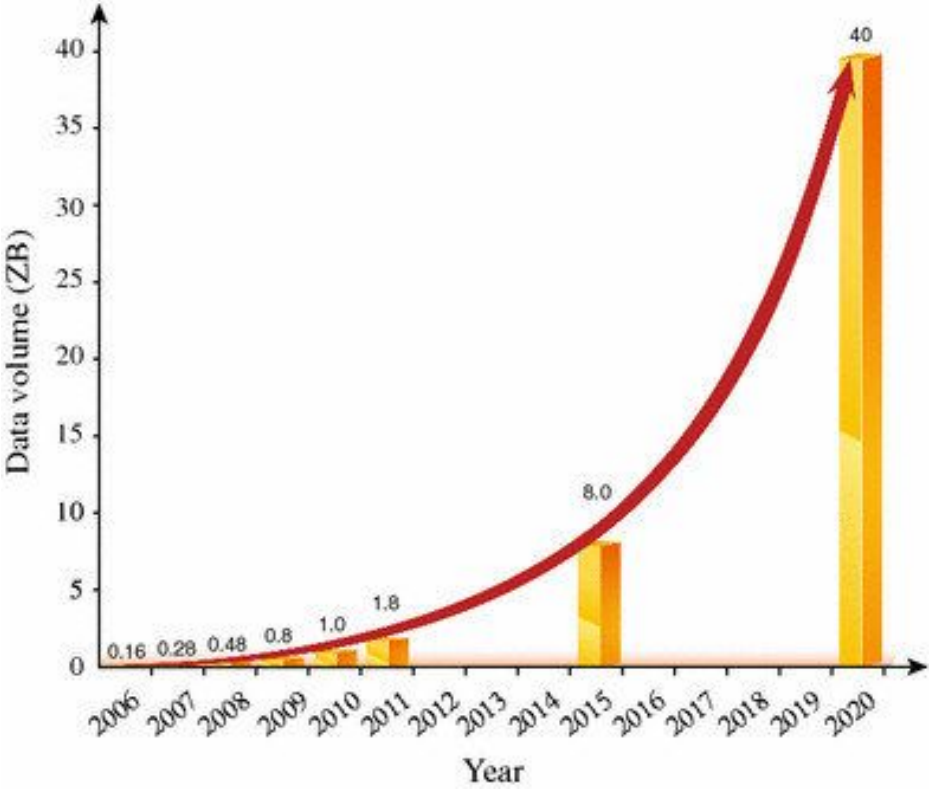
Big Data คืออะไร

Data ในปัจจุบันมีขนาดใหญ่ขึ้นมาก



Source: EMC.com

Big Data คืออะไร



A zettabyte is 10^{21} (1,000,000,000,000,000,000,000) bytes,

ทำไม Big Data ถึงใหญ่ขึ้นอย่างรวดเร็ว

ข้อมูลเติบโตขึ้น **ทุกนาที**

ใน 1 นาที:

- ผู้ใช้บน Youtube ดูวิดีโอ 4.3 ล้านครั้ง
- ผู้ใช้บน Google ทำการค้นหา 3.8 ล้านครั้ง
- ผู้ใช้ Twitter ส่งทวีต 470K ทวีต
- ผู้ใช้ Spotify เล่นเพลง 750,000 เพลง
- ผู้ใช้ Uber ใช้เวลาเดินทาง 1.3K

(Reference: Data Never Sleep - https://www.domo.com/learn/data-never-sleeps-6)

DataTH

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

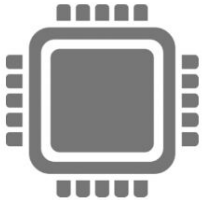
Common challenges encountered by data engineers

Access to data

Variety

Data accuracy and quality

Veracity



Availability of computational resources

Volume

Query performance

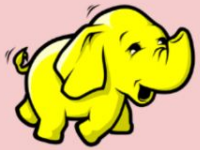
Volume

Velocity

Big Data Platform

ระบบที่สร้างขึ้นมาเพื่อจะช่วยให้เราควบคุม Big Data ได้

แบ่งเป็น 2 ประเภทหลัก ๆ



On-premise

ติดตั้งบนเซิร์ฟเวอร์ของบริษัท

Cloud Computing

ใช้บริการเซิร์ฟเวอร์ผ่านระบบอินเตอร์เน็ต
ส่วนใหญ่มักจะเป็นเซิร์ฟเวอร์ของบริษัทอื่น

DataTH

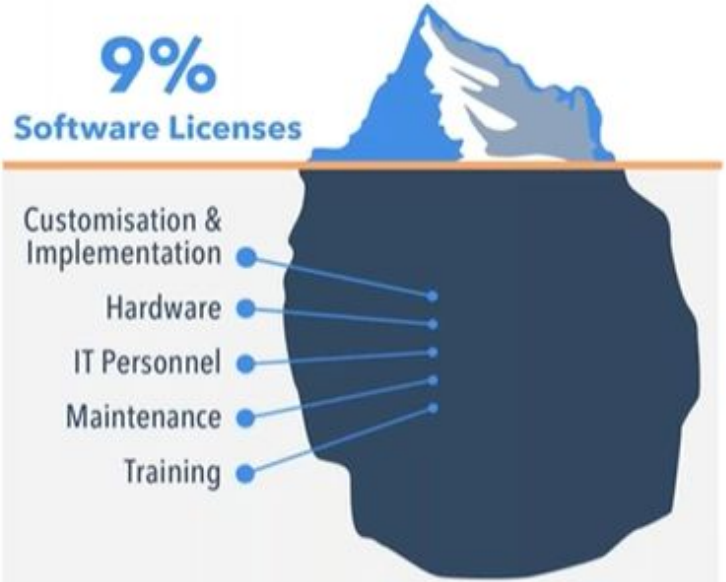
TROOP MESSENGER

ON PREMISE

Vs

CLOUD

On-Premises



Ongoing Costs

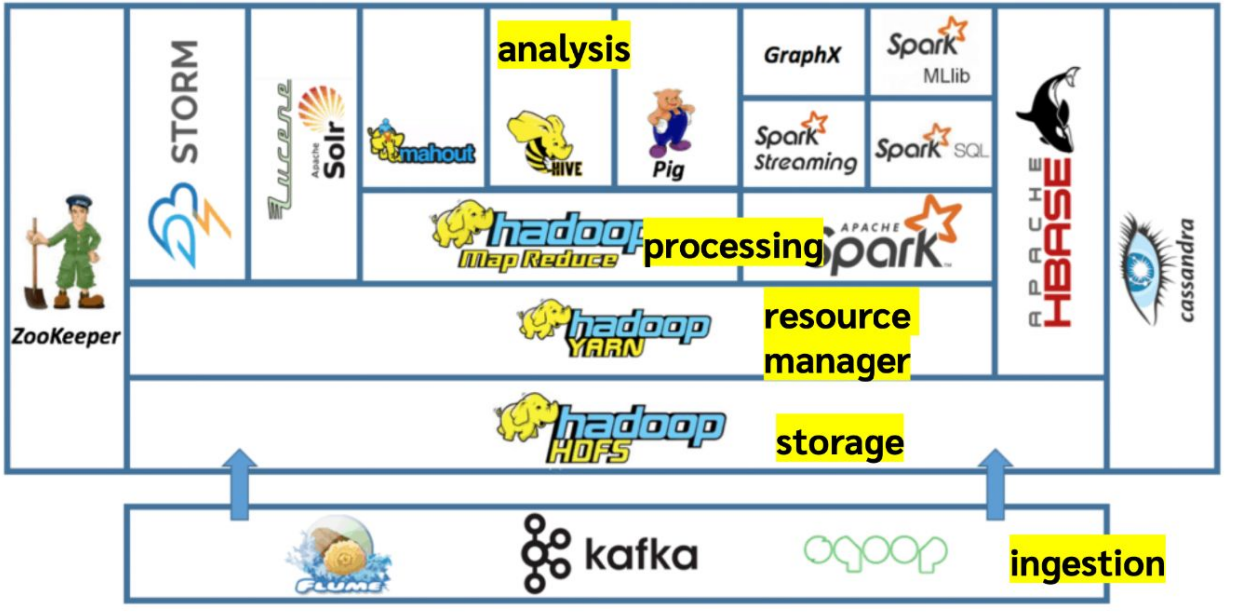
- Apply Fixes, Patches, Upgrade
- Downtime
- Performance tuning
- Rewrite customizations
- Rewrite integrations
- Upgrade dependent applications
- Ongoing burden on IT
- Maintain/upgrade hardware
- Maintain/upgrade network
- Maintain/upgrade security
- Maintain/upgrade database

Cloud Computing

Ongoing Costs

- Subscription fee

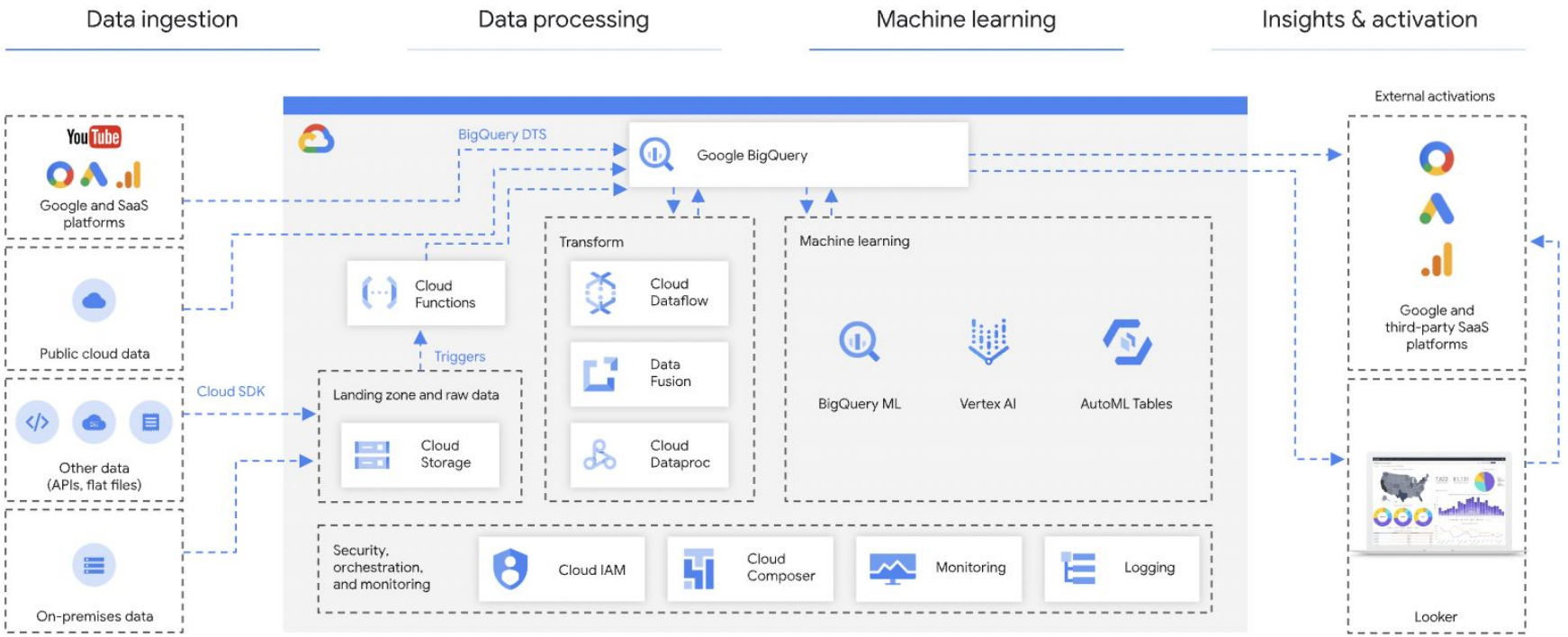
Big Data Platform (On-Premise)



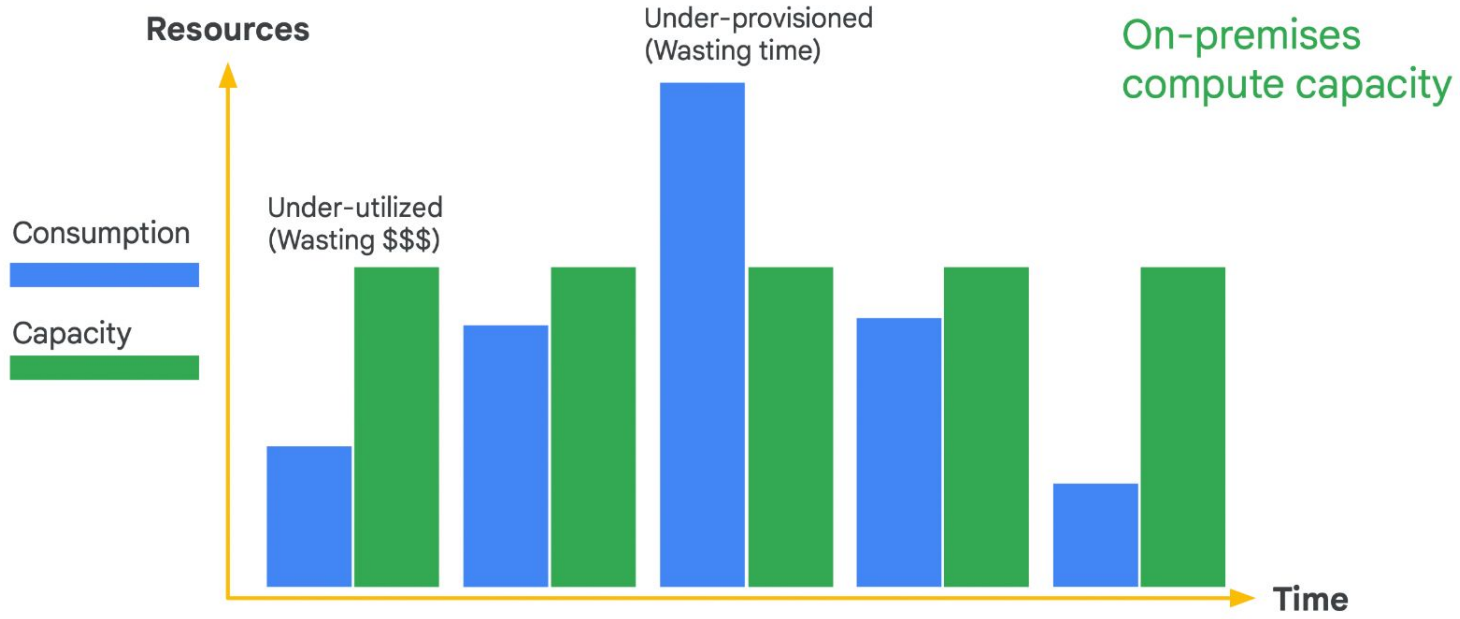
http://blog.newtechways.com/2017/10/apache-hadoop-ecosystem.html

DataTH

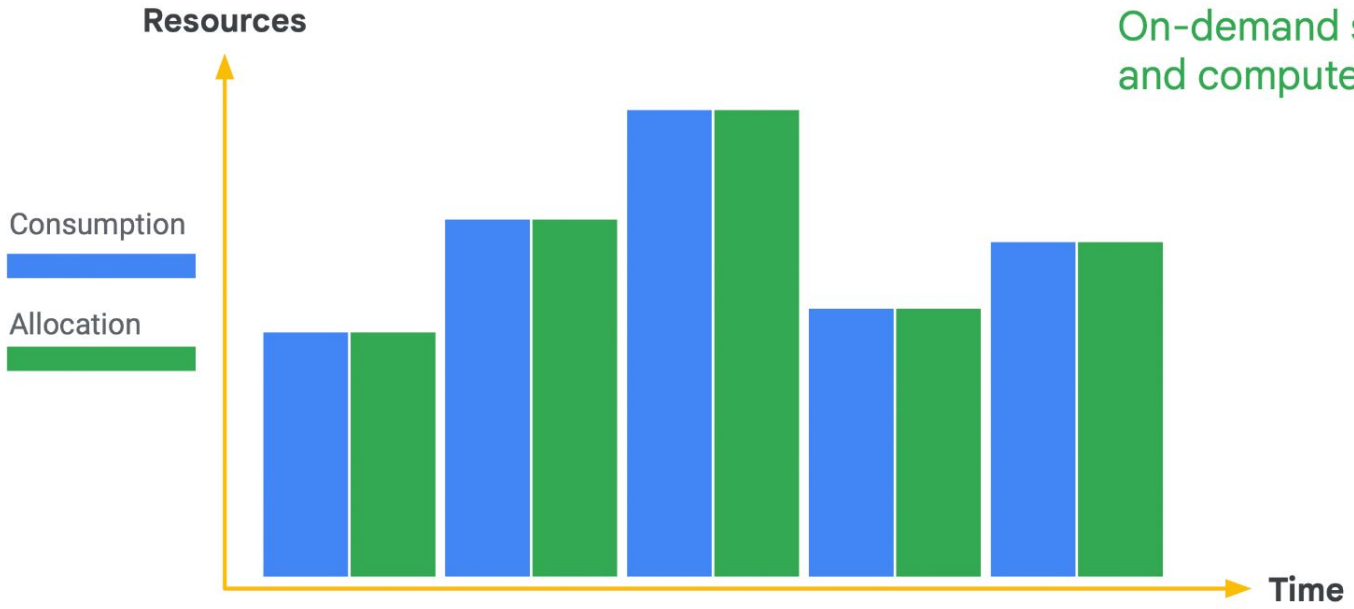
Big Data Platform (Cloud computing)



Challenge: Data Engineers need to manage server and cluster capacity if using on-premise



You don't need to provision resources before using BigQuery



Cloud computing: 3 top public cloud providers

Programming Language

(as Data Engineer)

Programming Language

Framework

Library

Tools

Programming Language

- A coding language, also known as a programming language, is a formal language with a **set of syntax rules and semantics** used to write instructions that a computer can execute.
- Coding languages allow developers to write programs, scripts, or applications by specifying the logic and behavior of the software.
- Examples of coding languages include Python, JavaScript, Java, C++, and Ruby

Framework

- A framework is a **pre-built structure or set of tools and libraries** that provide a foundation for building software applications.
- Frameworks typically enforce a specific structure or architecture, guiding developers on how to organize and implement their code.
- Frameworks often provide a standardized way to handle common tasks such as database interaction, user authentication, and routing.
- Examples of frameworks include Django and Flask for web development in Python, React and Angular for frontend development, and TensorFlow and PyTorch for machine learning.

Library

- A library is a **collection of pre-written code** modules provide reusable code for common operations, such as mathematical calculations, file manipulation, or network communication.
- Examples of libraries include NumPy and pandas for data manipulation in Python, jQuery for DOM manipulation in JavaScript, and React for building user interfaces

Tools

- **Program or utility designed to assist developers** in performing specific tasks. Tools can range from simple utilities for code editing, debugging, and version control to more specialized tools for performance optimization, testing, and documentation generation.
- Examples of tools include text editors like Visual Studio Code and Sublime Text, version control systems like Git

Programming Language

Framework

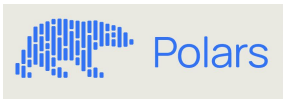
Library

Tools

Programming Language

Framework

Library



Tools

Language



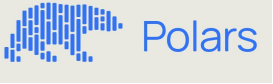


Framework

Library



Koalas




Tools


Visual Studio Code

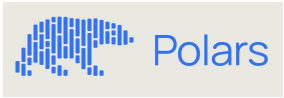
GitLab

Language

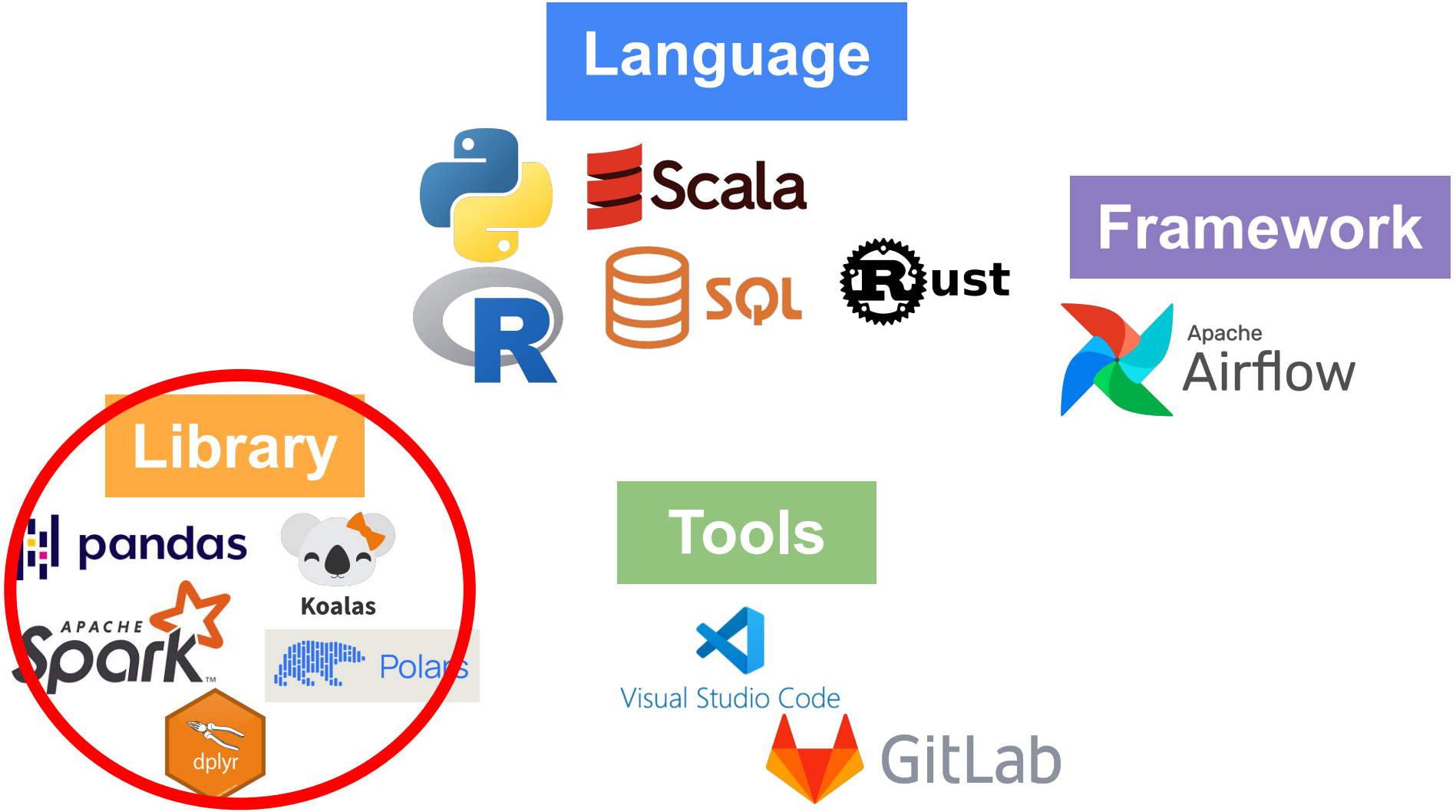


Framework

Library



Tools


Language
Scala
SQL
Rust
R
Framework
Apache
Airflow
Library
pandas
Koalas
APACHE
Spark
Polars
dplyr
Tools
Visual Studio Code
GitLab

Library

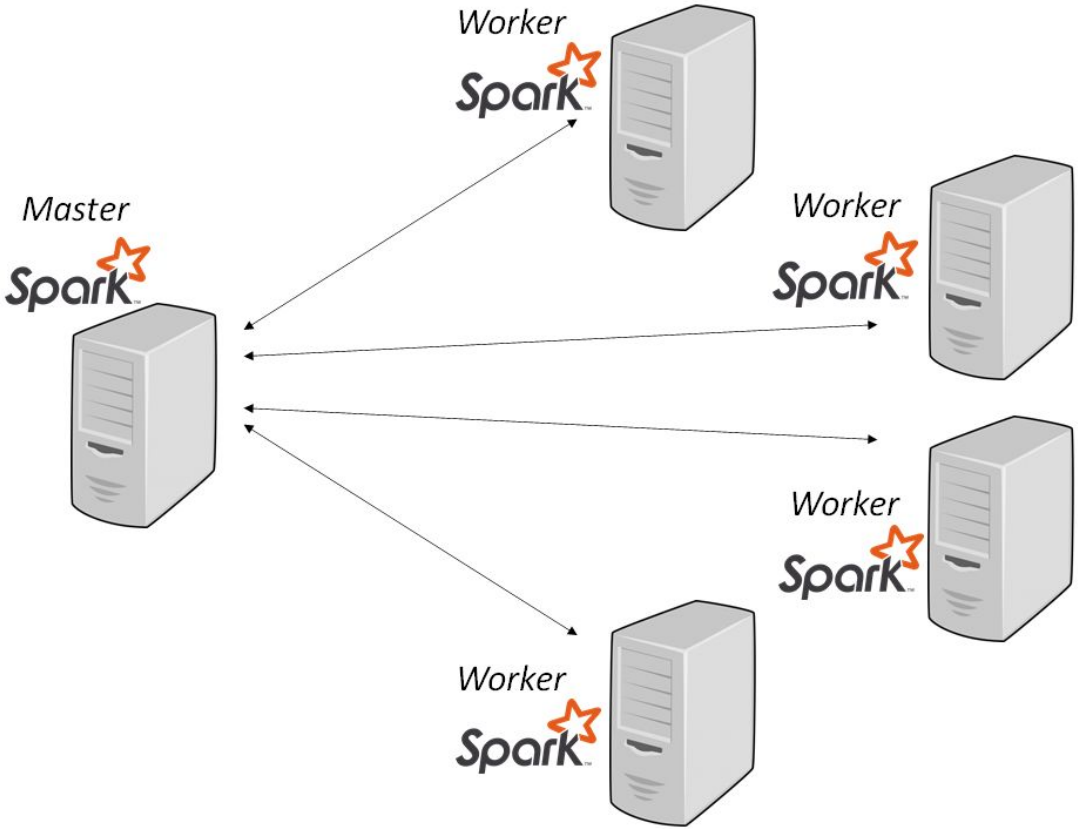
	pandas	Apache Spark	Polars	Koalas	dplyr
Programming Language	Python	Scala, Python, Java, R, Spark SQL	Rust, Python	Spark SQL	R
Dataset size	Small-Medium	Large dataset	Large dataset	Large dataset	
Operate	Single Node	Multiple Node	Multiple Node	Multiple Node	
Difficulty	Easy	Medium-Hard	Easy	Easy	

pandas

Polars

APACHE Spark

Koalas

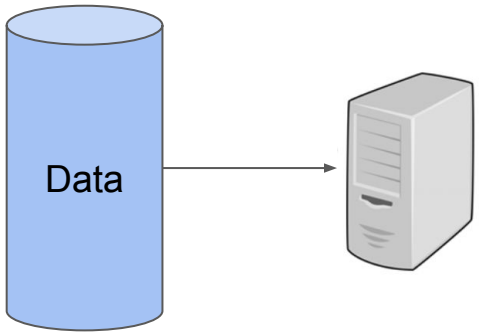


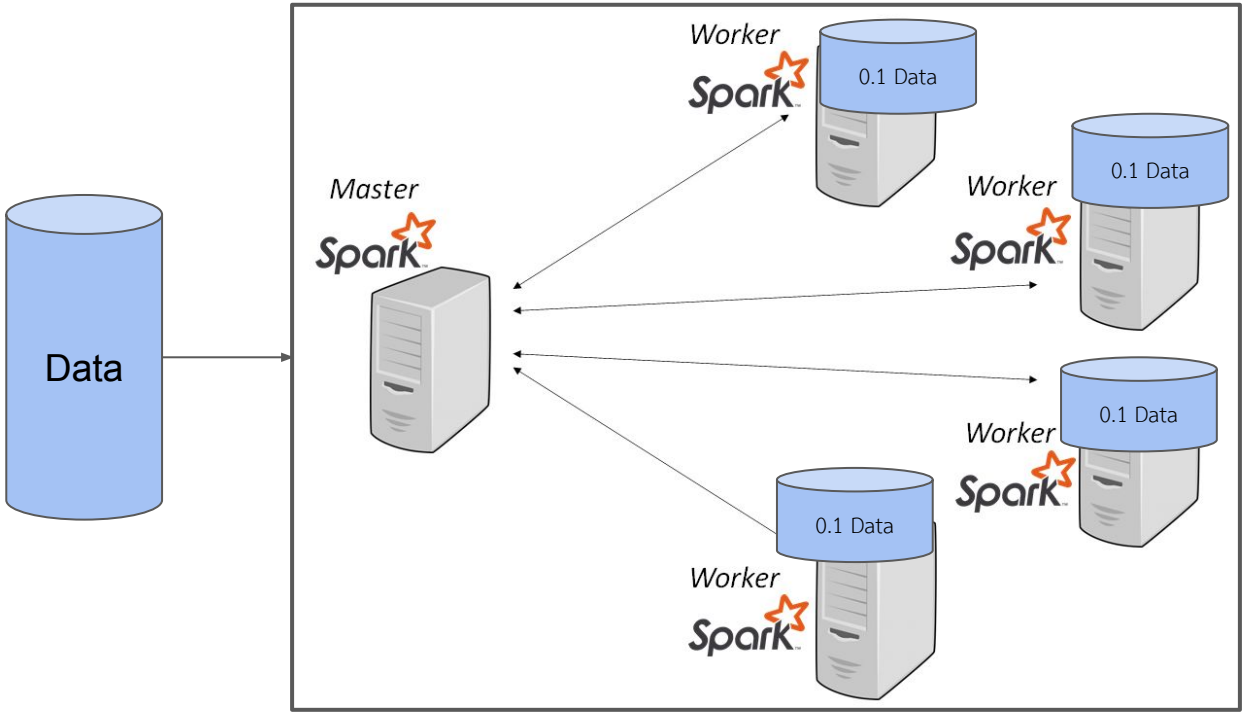
pandas

Polars

APACHE Spark

Koalas



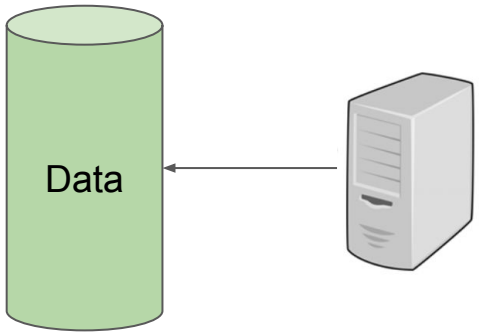


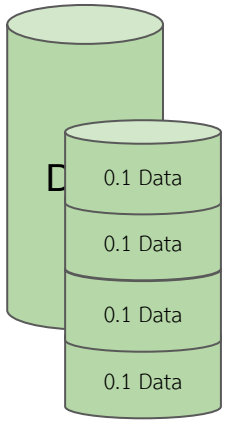
pandas

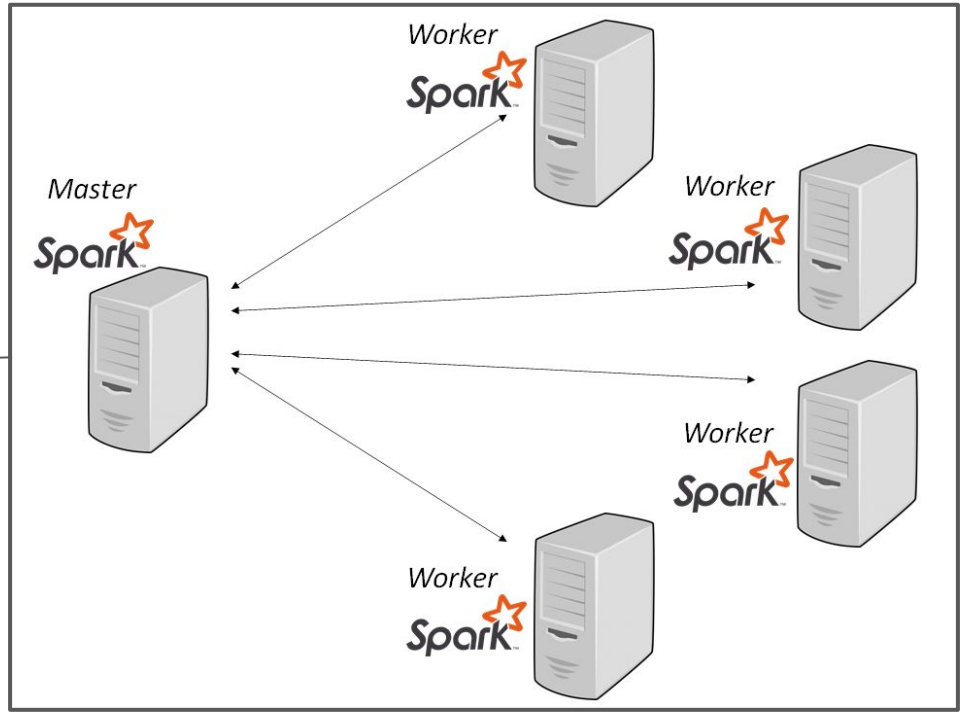
Polars

APACHE Spark

Koalas

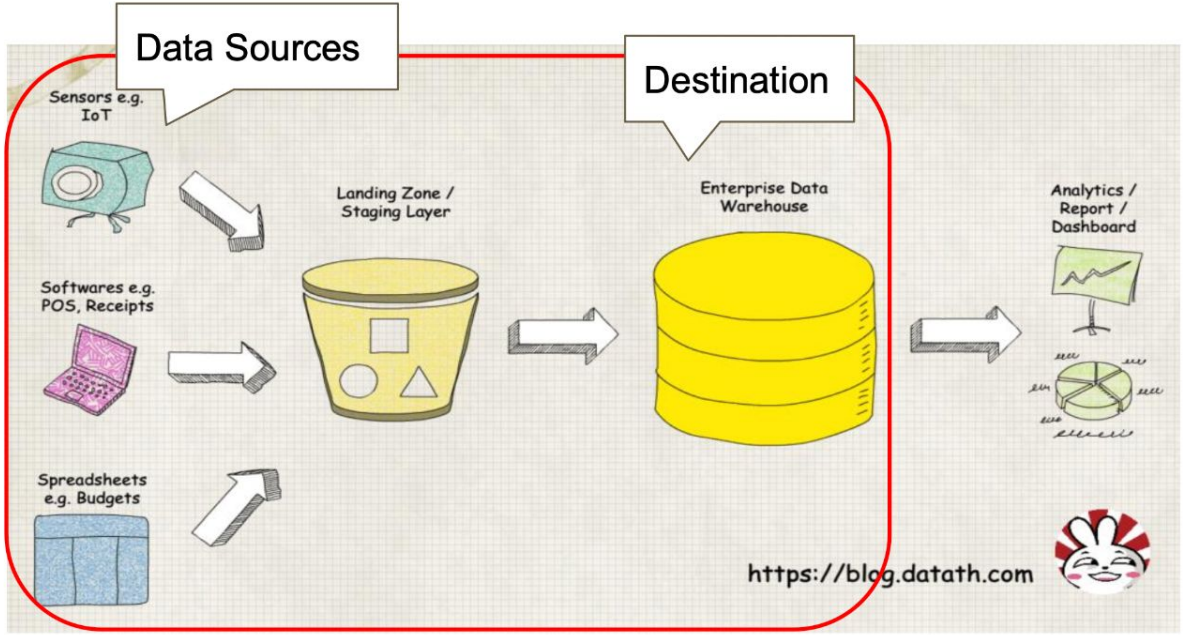






Data Pipelines and ETL

What is Data Pipeline?



ท่อในการลำเลียงข้อมูล จาก

แหล่งข้อมูล
(Data Source)

ไปยัง

จุดหมาย
(Destination)

DataTH

องค์ประกอบของ Data Pipeline

Monitoring

Automation

Data Sources

Business Data

Customer Data

Marketing Data

Extract

Transformation

Load

Destination

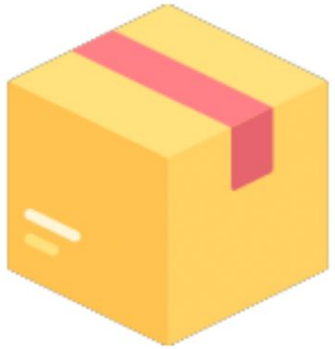
Data Warehouse

Database

Data Lake

Processing

ประเภทของ Data Pipeline



Initial Load / Historical load / Full load

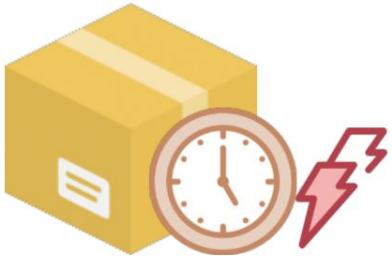
ดึงข้อมูลทั้งหมดจากแหล่งข้อมูล

Incremental Load / Change Data Capture (CDC) Load

ดึงข้อมูลใหม่ และข้อมูลที่เปลี่ยนแปลงจากครั้งล่าสุด

DataTH

ประเภทของการประมวลผลข้อมูล (Processing)



Batch

Scheduled: ดึงตามช่วงเวลาที่กำหนด เช่น วันละครั้ง, ชั่วโมงละครั้ง

Event-Driven: ประมวลผลข้อมูลตอนที่ได้รับสัญญาณให้ประมวลผล

Stream

ข้อมูลจะถูกส่งเข้ามาทันที และเราประมวลผลทันที

(อีกวิธีที่นิยม คือ ทำเป็น Mini-batch ประมวลผลทุก 5 วินาที - 5 นาที)

Tip: เลือกตามความถี่ของแหล่งข้อมูลเป็นหลัก ข้อมูลบางแหล่งสามารถ Stream ได้ บางแหล่งไม่สามารถทำได้

DataTH

E = Extract

T = Transform

L = Load

E = Extract

Business Data

Customer Data

Marketing Data

การดึงข้อมูลออกมาจากแหล่งข้อมูล (Data Source) ต่าง ๆ มีข้อที่ควรคำนึงถึงดังนี้:

- **ประเภทข้อมูล**
  เช่น CSV, JSON, API, Database, Data Warehouse / Mart ฯลฯ
- **หน้าตาข้อมูล**
  เช่น จำนวนคอลัมน์, ชื่อคอลัมน์, Format ของข้อมูลที่เก็บ (เช่น ชื่อ นามสกุล อาจเก็บรวมหรือแยกกัน)
- **ความถี่ในการอัพเดท**
  เช่น อัพเดทข้อมูลทุกชั่วโมง หรืออาทิตย์ละครั้ง

DataTH

T = Transform

การเปลี่ยนแปลงข้อมูลสามารถทำได้หลากหลายรูปแบบ เช่น

- **ปรับให้รูปแบบเหมาะสมกับระบบปลายทาง**
  เช่น ต้นทางใช้วันที่แบบ DD/MM/YYYY ส่วนปลายทางใช้ YYYY-MM-DD

- **สรุปข้อมูล (Aggregation)**
  เช่น คำนวณค่าเฉลี่ย, ผลรวม

- **เพิ่มคุณค่าของข้อมูล (Enrichment)**
  เช่น รวมข้อมูลยอดขายของแต่ละวัน กับข้อมูลสภาพอากาศในวันนั้น ๆ

DataTH

L = Load

Database

Data Warehouse

Data Lake

การนำข้อมูลเข้าไปในระบบปลายทาง

Database, Data Warehouse, Data Lake ก็สามารถเป็นระบบปลายทางได้ทั้งหมด

หากทำการ Transform มาก่อนหน้านี้ การ Load จะเป็นการส่งข้อมูลจากที่เก็บข้อมูลชั่วคราว (Staging Layer / Area) เข้าไปเก็บในระบบปลายทาง

DataTH

ETL vs ELT

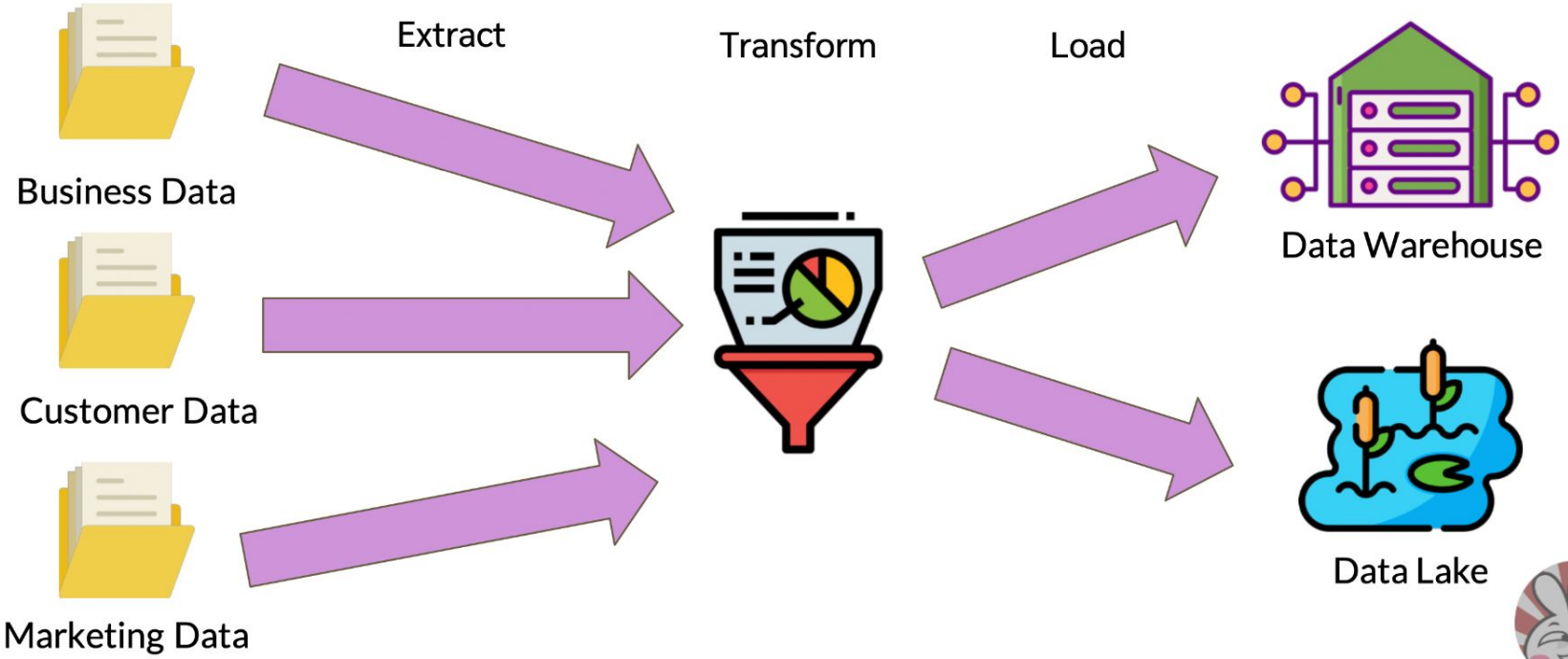
ETL - Extract - Transform - Load

- วิธีปกติในการย้ายข้อมูล เป็นที่นิยมในการย้ายข้อมูลไปที่ต่าง ๆ
- ระบบปลายทางไม่จำเป็นต้องประมวลผล (Transform) ข้อมูลเยอะ
- ต้องมี Staging Area แยก เพื่อประมวลผลข้อมูล
- Data Analyst ต้องรอ ETL เสร็จถึงจะได้ข้อมูล

ELT - Extract - Load - Transform

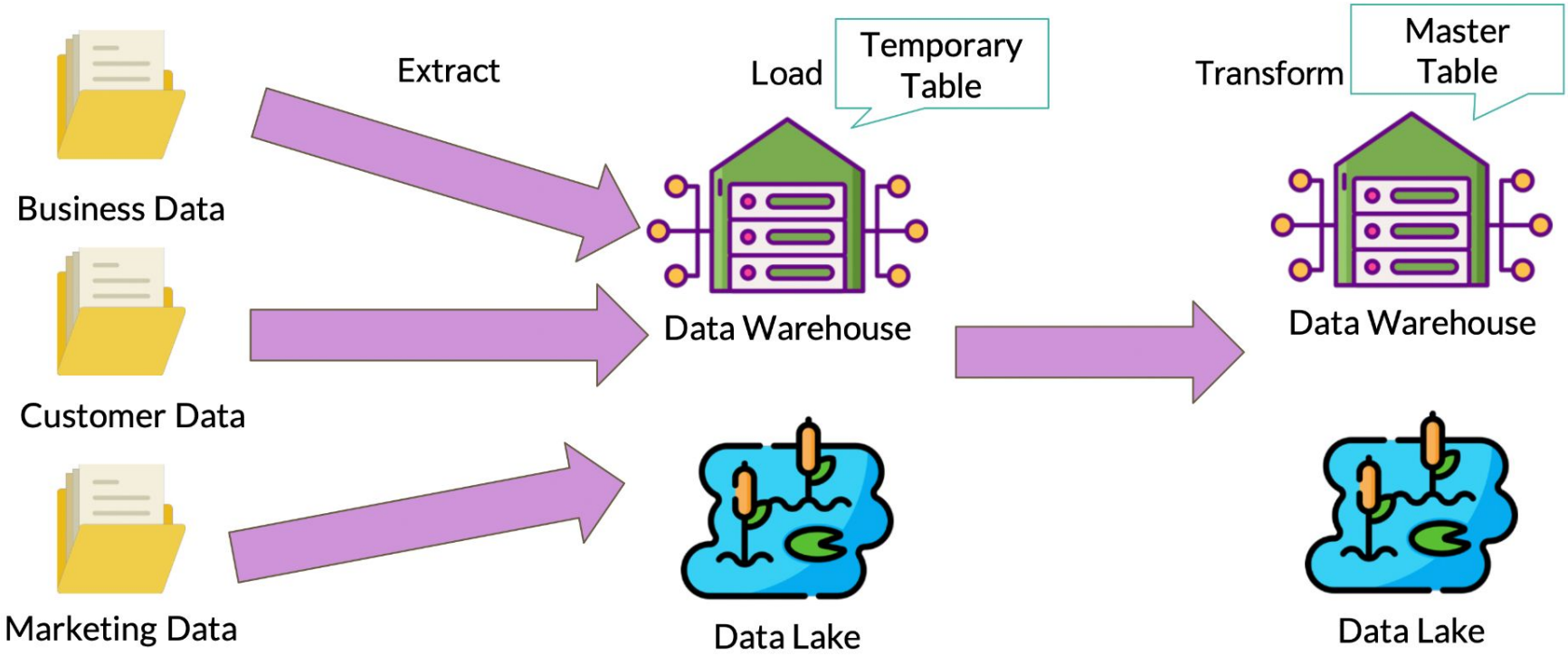
- วิธีการย้ายข้อมูลสมัยใหม่ ระบบใหม่ ๆ จะสามารถทำได้ เช่น Redshift, Snowflake
- ระบบปลายทางจะต้องประมวลผลข้อมูลเยอะ
- ใช้ระบบปลายทางเป็น Staging Area
- Data Analyst เข้าถึงข้อมูลได้เร็วกว่า ไม่ต้องรอ ELT เสร็จ สามารถ Transform ข้อมูลดิบตอนดึงข้อมูลได้เลย

ETL – Extract, Transform, Load



DataTH

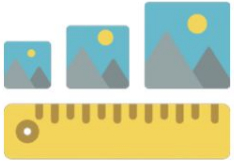
ELT – Extract, Load, Transform



Key Considerations / Trade-offs

Accuracy
ความถูกต้องของข้อมูล

Speed
ความเร็วในการย้ายข้อมูล



Scalability
ความสามารถในการรับข้อมูลปริมาณมาก

Security
ความปลอดภัยของข้อมูลระหว่างส่ง

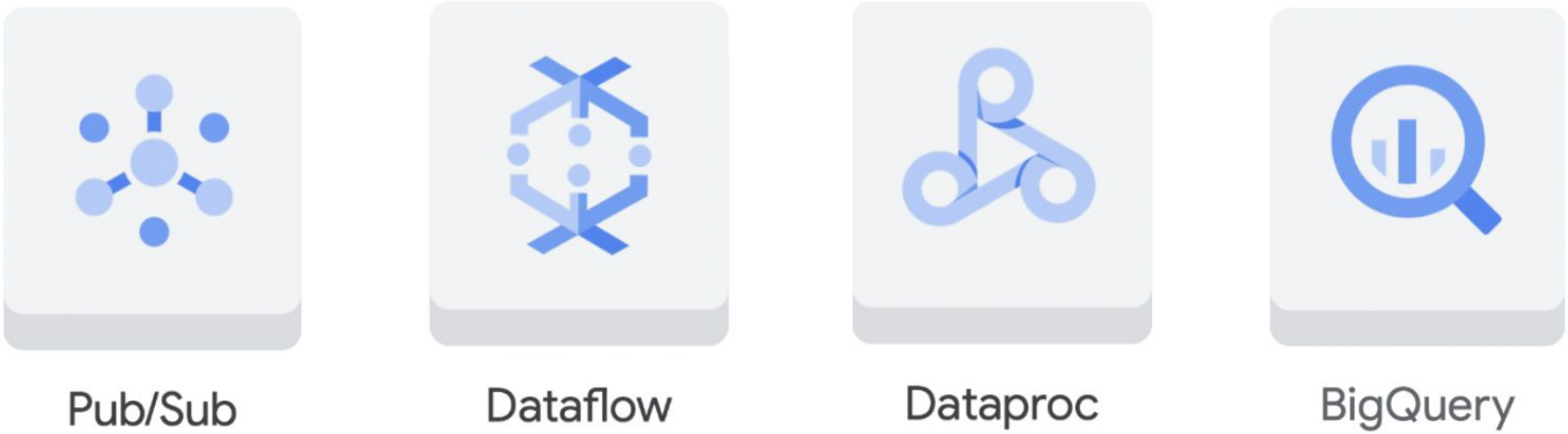
Tip: การเพิ่มเรื่องหนึ่ง อาจจะไปลดอีกเรื่อง เช่น เพิ่ม Speed แล้ว Accuracy ลดลง
เราต้องหาบาลานซ์ให้เหมาะสมกับความต้องการของโปรเจค

DataTH

Cloud computing: 3 top public cloud providers

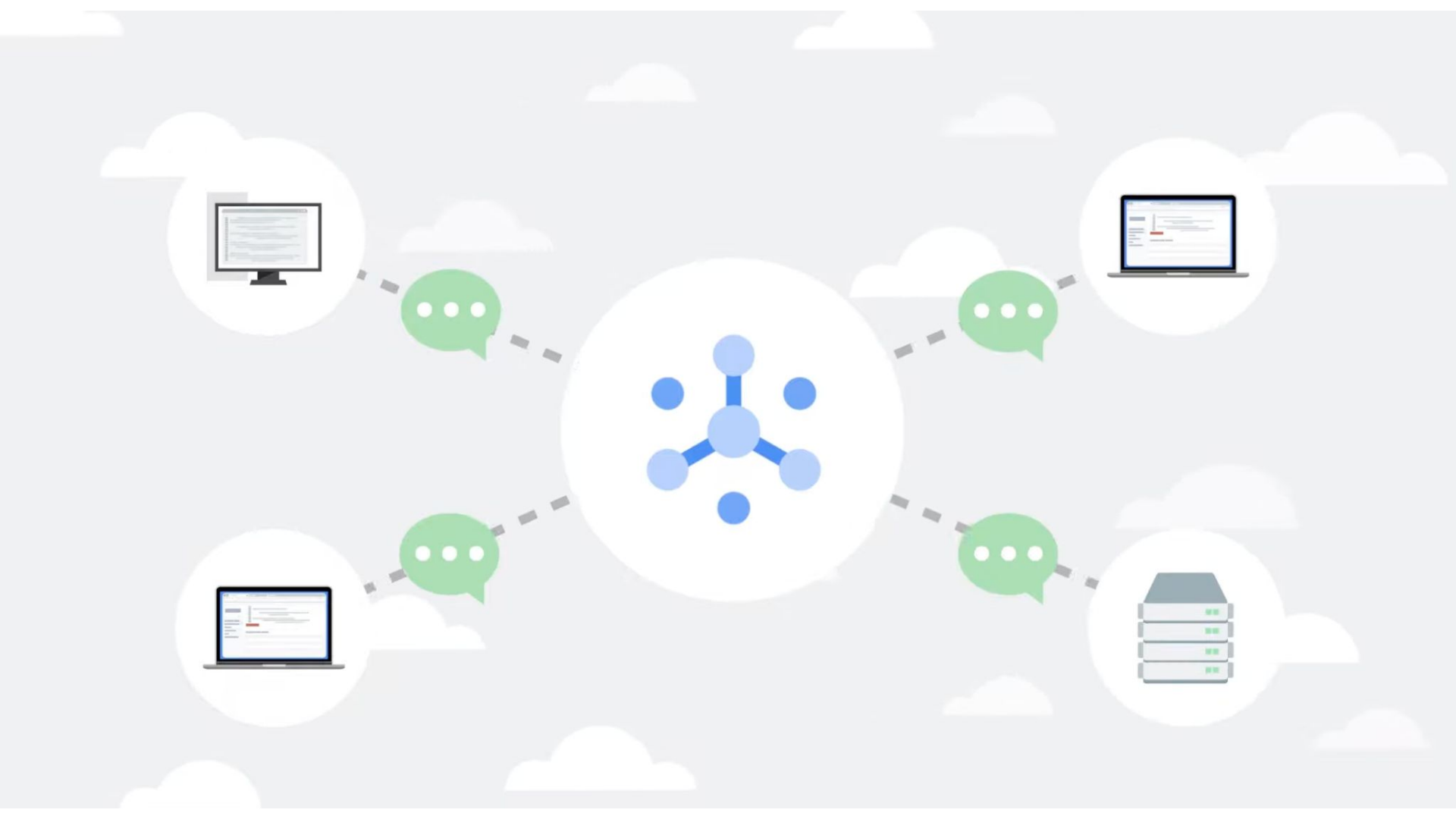
DataTH

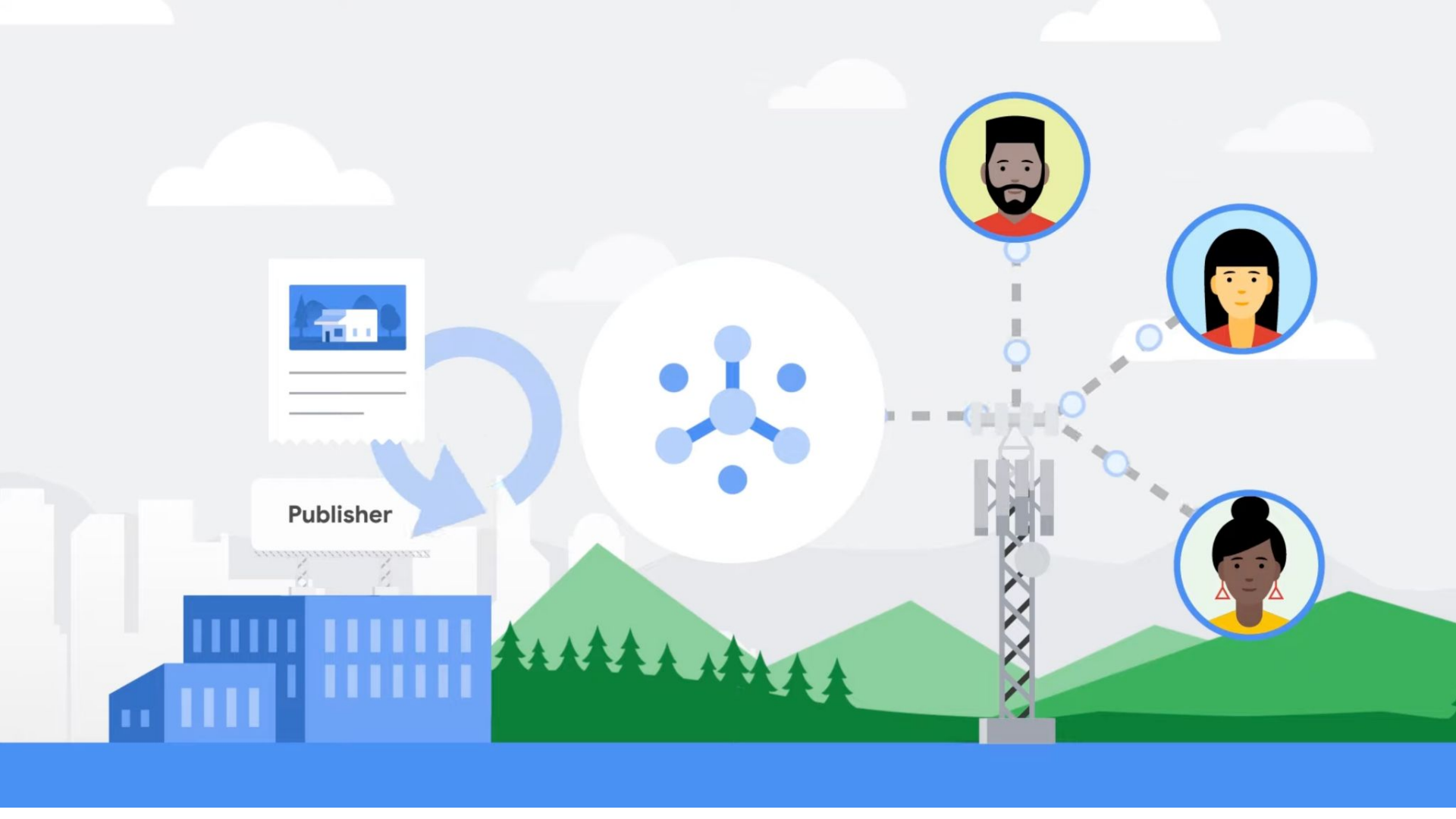
Example of Google Cloud Platform (GCP) products

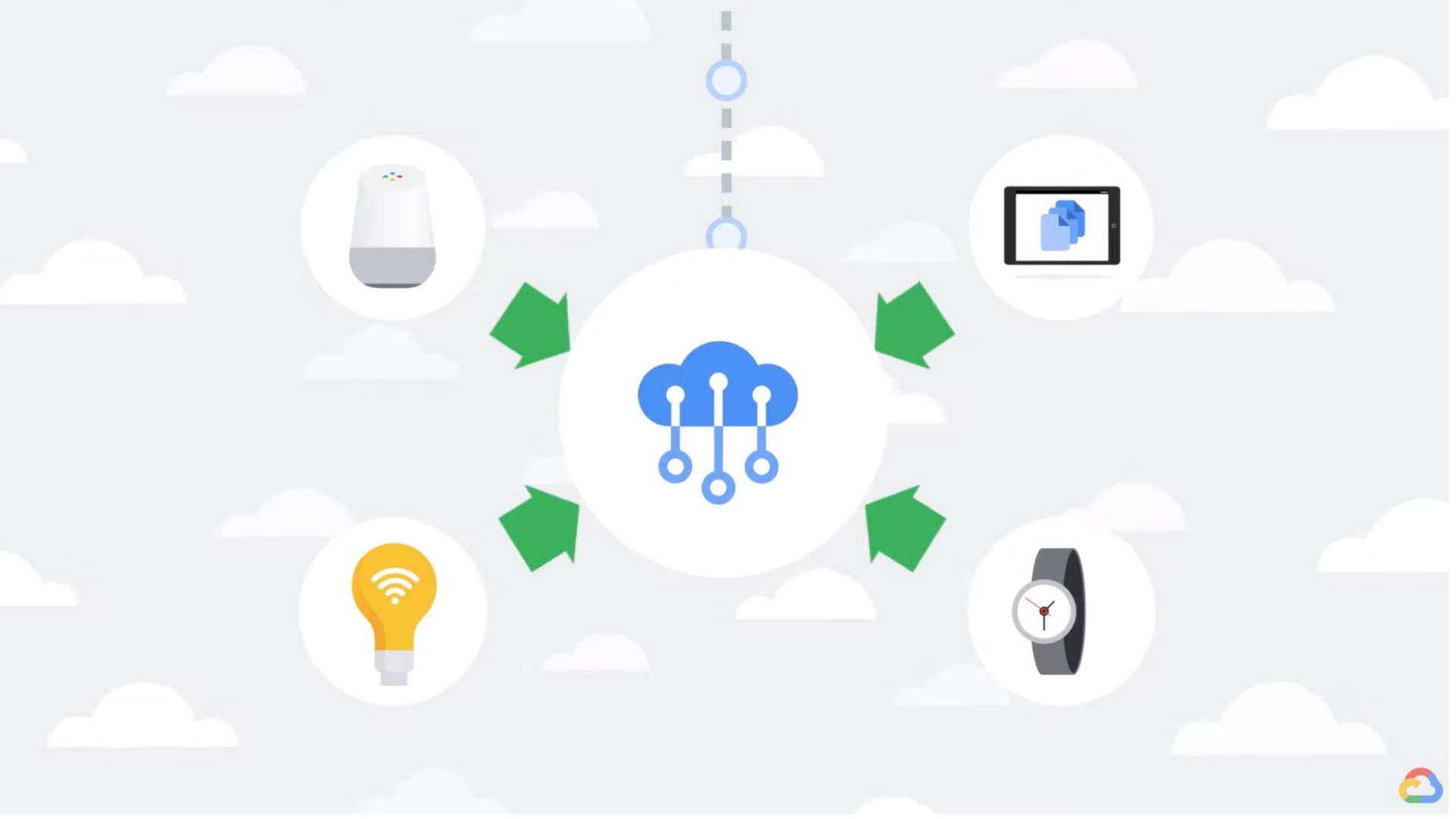


Pub/Sub

aka → Publisher / Subscriber

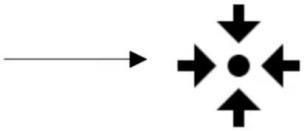



Publisher



Dataflow

Data sources

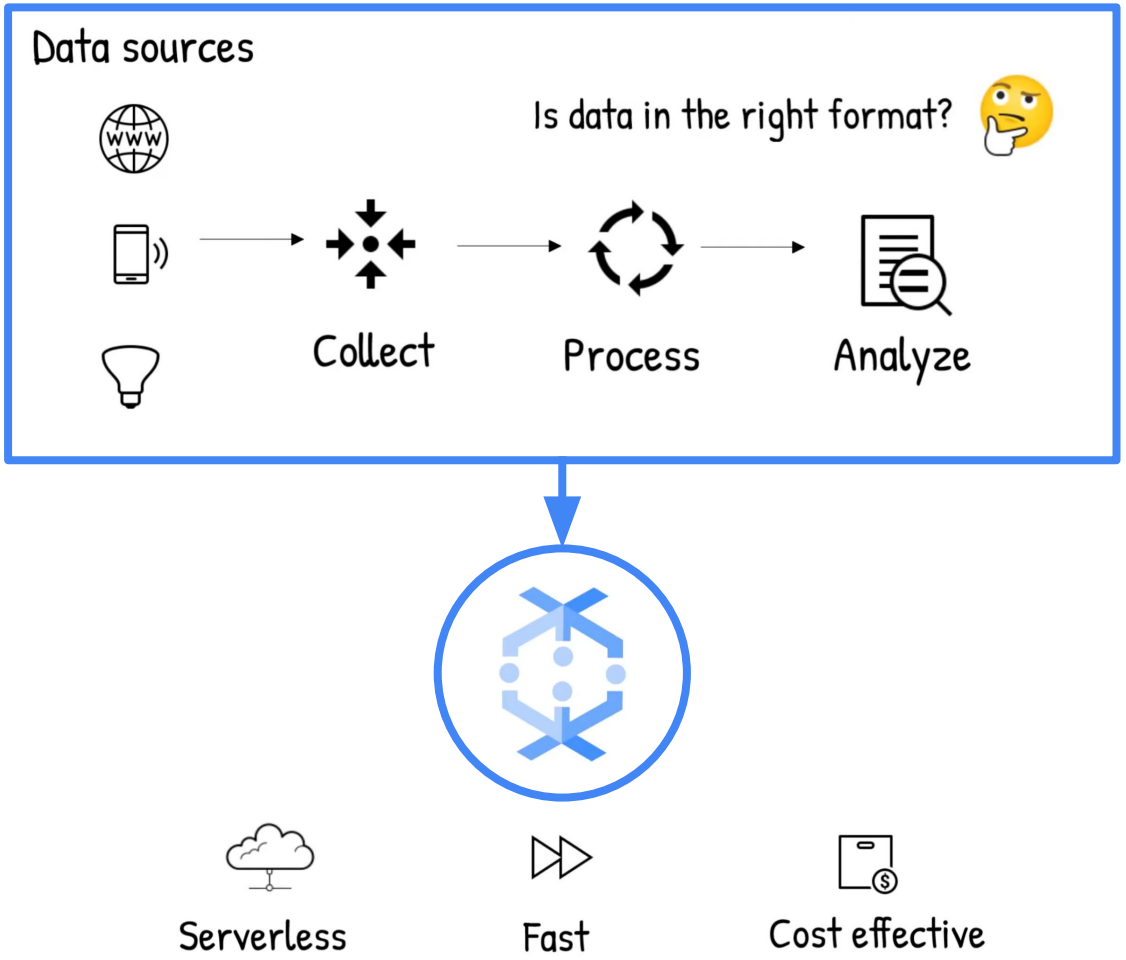


Collect

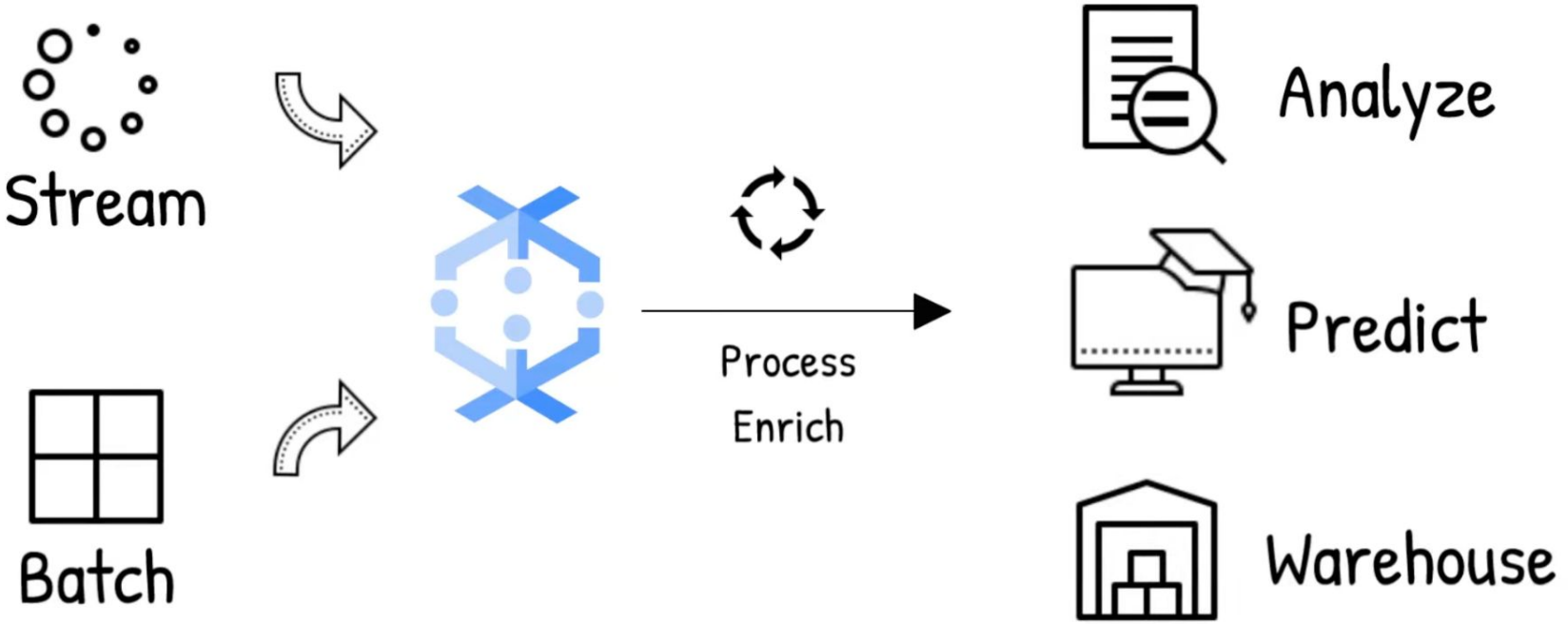
Process

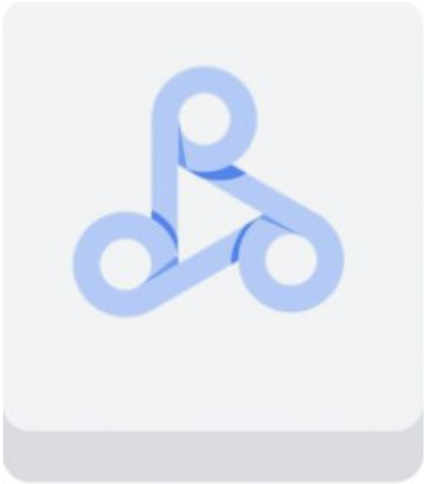
Analyze

Is data in the right format?

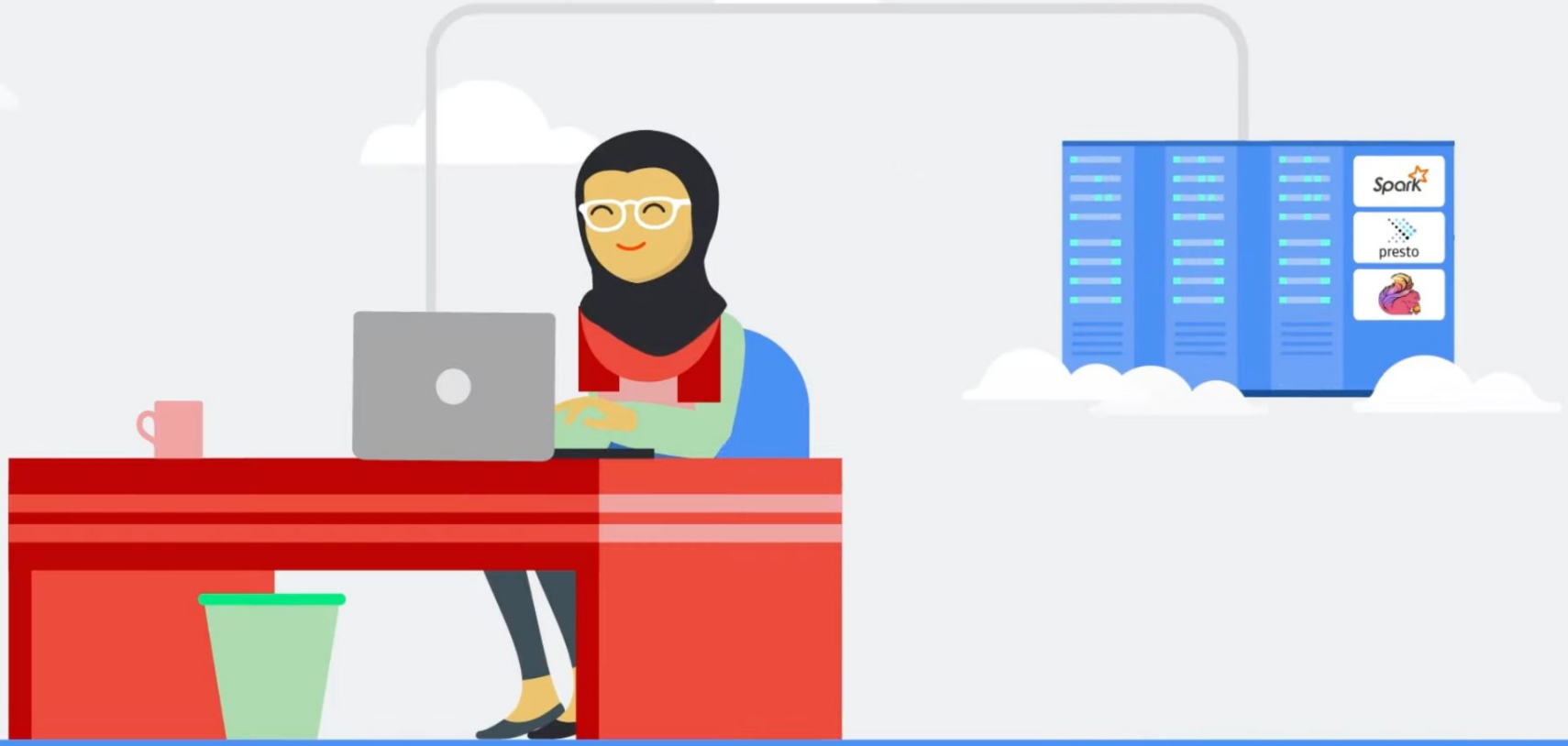

Data sources
Is data in the right format?
Collect
Process
Analyze
Serverless
Fast
Cost effective

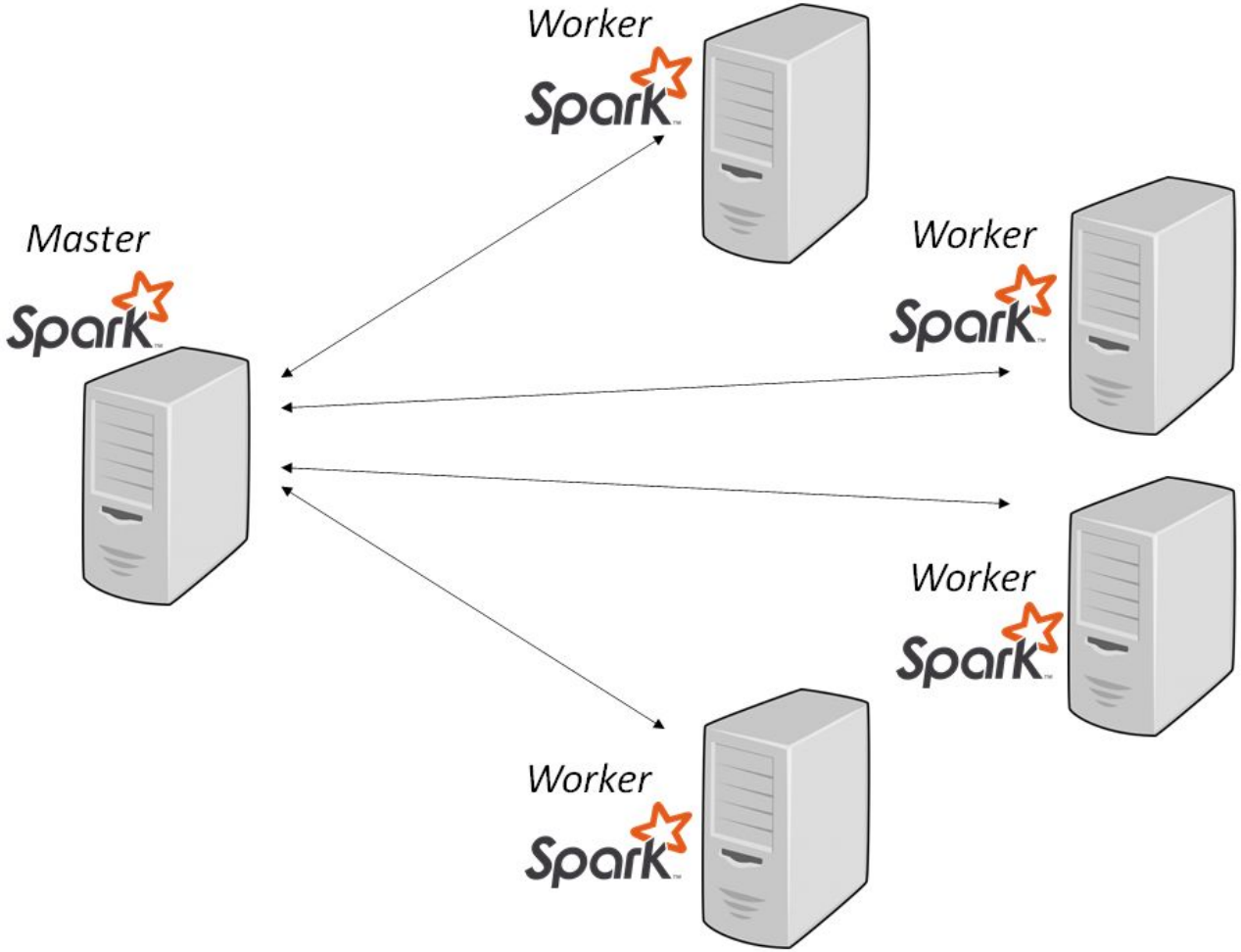
Where to use Dataflow?

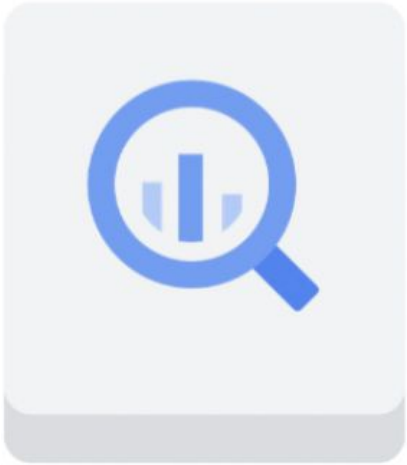




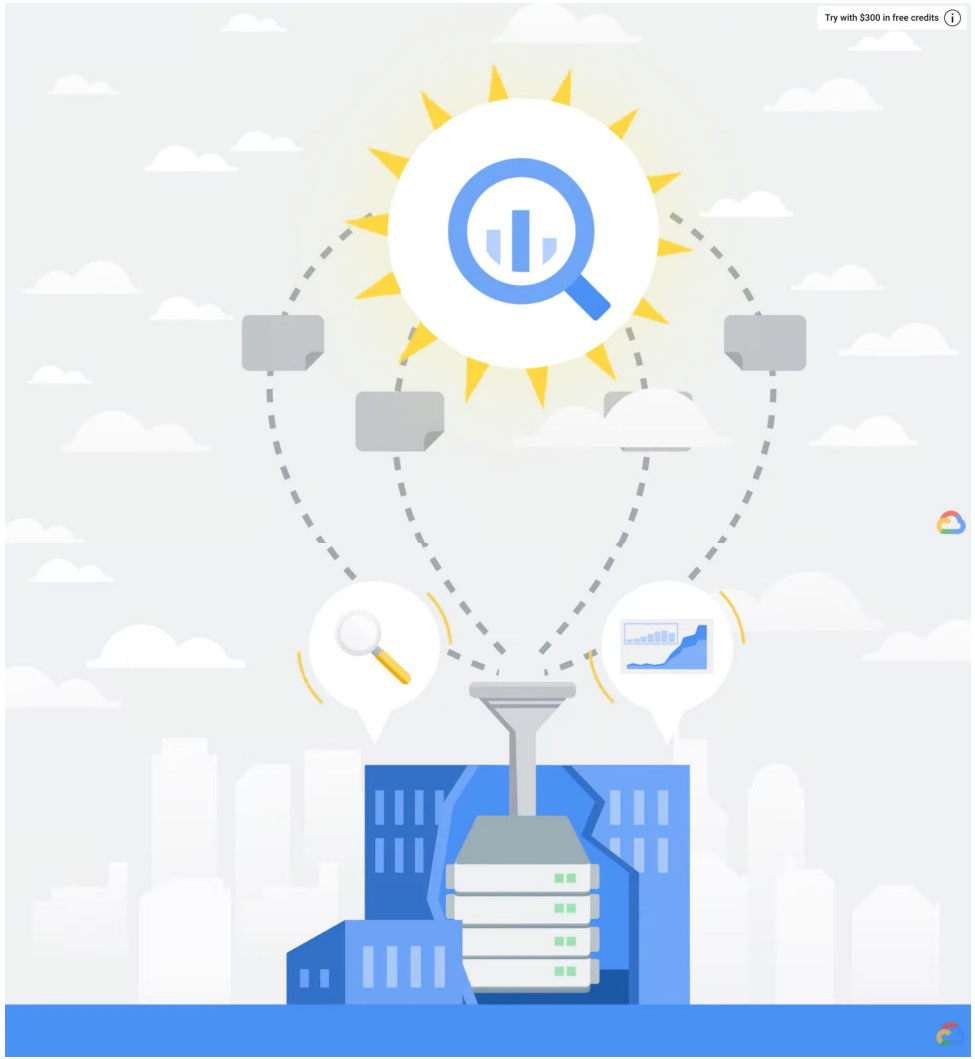
Dataproc

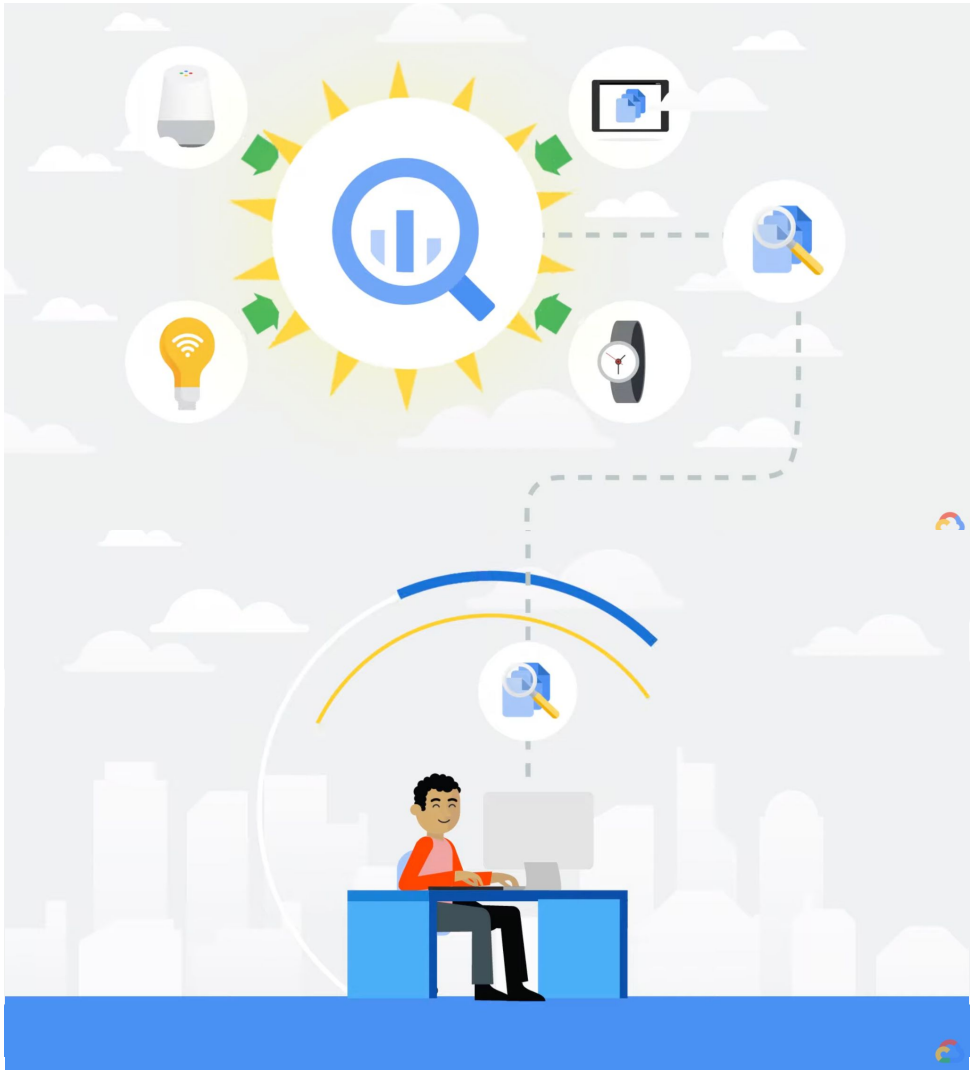

Spark
presto

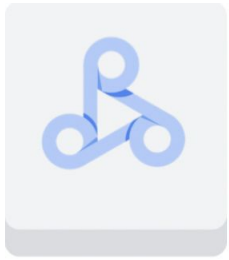
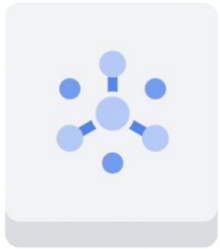
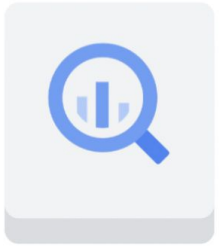

Worker
Spark
Master
Spark
Worker
Spark
Worker
Spark
Worker
Spark



BigQuery


Try with $300 in free credits



	 Dataproc	 Dataflow	 Pub/Sub	 BigQuery
Programing Language	Python, Spark	Beam	User Interface	SQL
Built-in	Spark, Hadoop, Hive, Pig		Kafka	Dremel, Colossus, Borg
Described	Compute Process Managed	Pipeline orchestrate	Publisher, Subscriber	Analytic Fully Managed
Service Type	Depend on **resource** and **duration**	Fully Managed	Fully Managed	High
Price		High	High	

What if your data is not usable in its original form?

Extract, **T**ransform, and **L**oad

Data processing

Dataproc

Dataflow

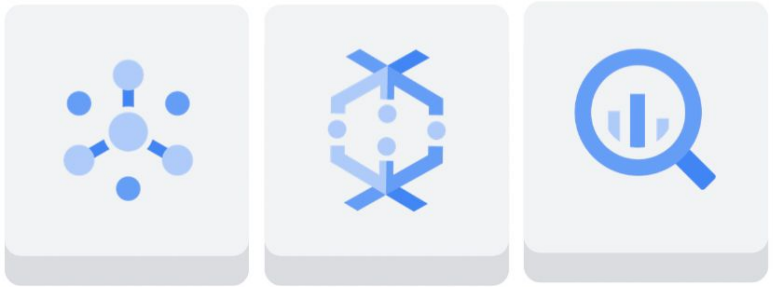
What if your data is not usable in its original form?

Extract, **T**ransform, and **L**oad

What if your data arrives continuously and endlessly?

What if your data arrives continuously and endlessly?

Streaming data processing



Pub/Sub Dataflow BigQuery

Q & A