Road to Data Engineer Workshop

มาลองเป็น Data Engineer กัน 2 วัน

DataTH.com

ข้อตกลงการเรียน: จดได้ ถามได้

หาที่จดไว้ ไม่ลืมแน่นอน

ระหว่างเรียน ถามคำถามได้ตลอดเวลาใน slido.com

ตอนจบแต่ละ Section จะมีเวลาให้ถามคำถาม

Our Focus

Practical
คุยกันแบบ Engineer
ประสบการณ์ > ทฤษฎี

Up-to-date
เน้นเทคโนโลยีที่กำลังมาแรง
และมีแววว่าจะอยู่ไปอีกนาน

Chapter 0:
Welcome to the world of Data Engineer

Welcome to **Sarah Gift World**

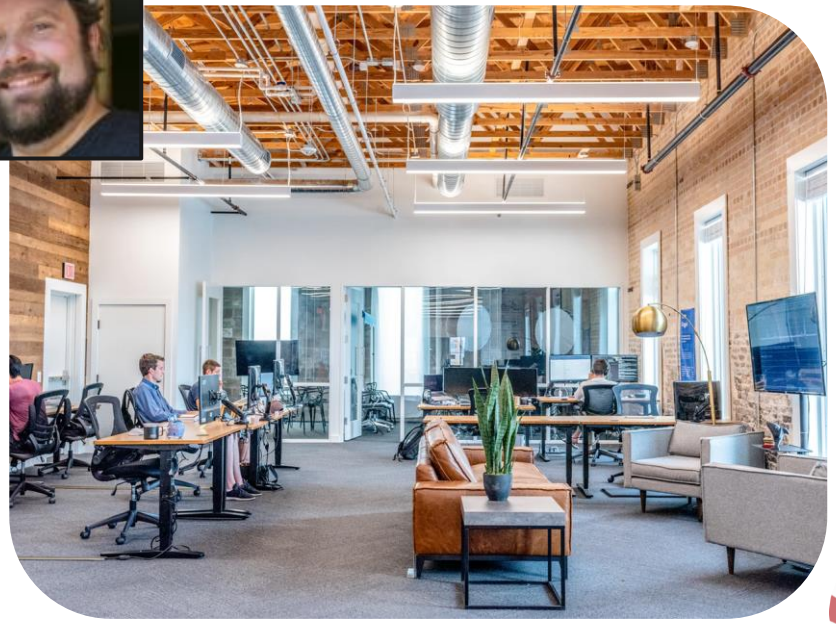
ยินดีด้วย!

คุณได้รับเข้าทำงานเป็น Data Engineer คนใหม่ ที่บริษัทร้านค้าออนไลน์ชื่อดังจากอังกฤษ **Sarah Gift World**

Your Mentor: Nick

นิคเป็น Senior Data Engineer ที่ Sarah Gift World

นิค: *“ผมเป็น Data Engineer ที่นี่มา 3 ปี มีอะไรให้ช่วยก็บอกได้เลยนะ”*



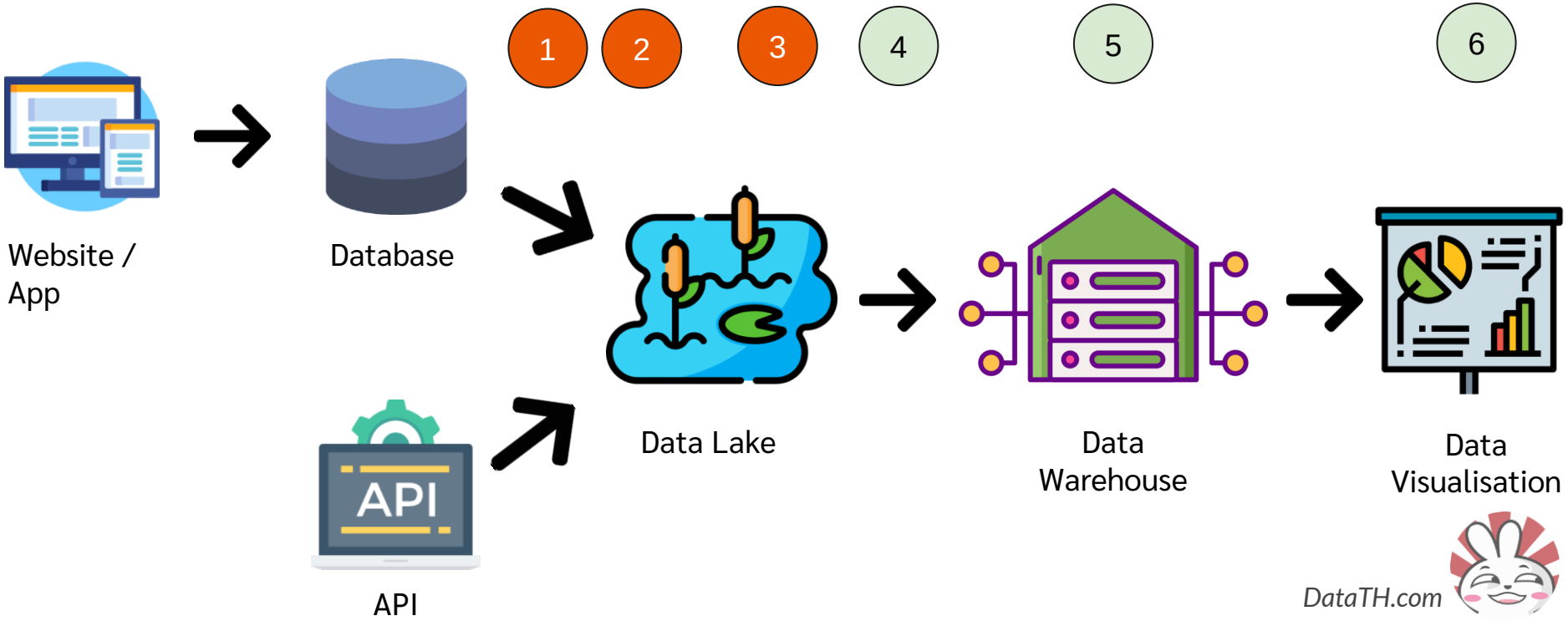
Your first project at Sarah Gift World

Requirement ทางธุรกิจ:
ทีม Marketing และ Supplier Relationship อยากรู้ว่าสินค้าไหนขายดี และวันไหนขายดี เพื่อจะได้วางแผนจัดโปรโมชั่น รวมถึงนำข้อมูลนี้ไปต่อรองกับ Supplier

Requirement ทาง Tech:
บริษัทเก็บข้อมูลยอดขายส่วนหนึ่งไว้ใน Database และสามารถต่อกับ API ที่แปลงค่าเงิน GBP เป็นเงินบาทได้

เราอยากให้ Data Analyst เอาข้อมูลนี้มาทำ Report และ Dashboard เสนอทีมอื่น

Your first project at Sarah Gift World



Workshop Day 1

- Introduction to Data Engineering
- Data Pipeline & ETL
- **Workshop 1: Data Collection with Python**
- Data Quality & Wrangling
- **Workshop 2: Data Wrangling with Spark**
- Basic Cloud - Google Cloud Platform
- **Workshop 3: Data Storage with GCS**

DataTH.com

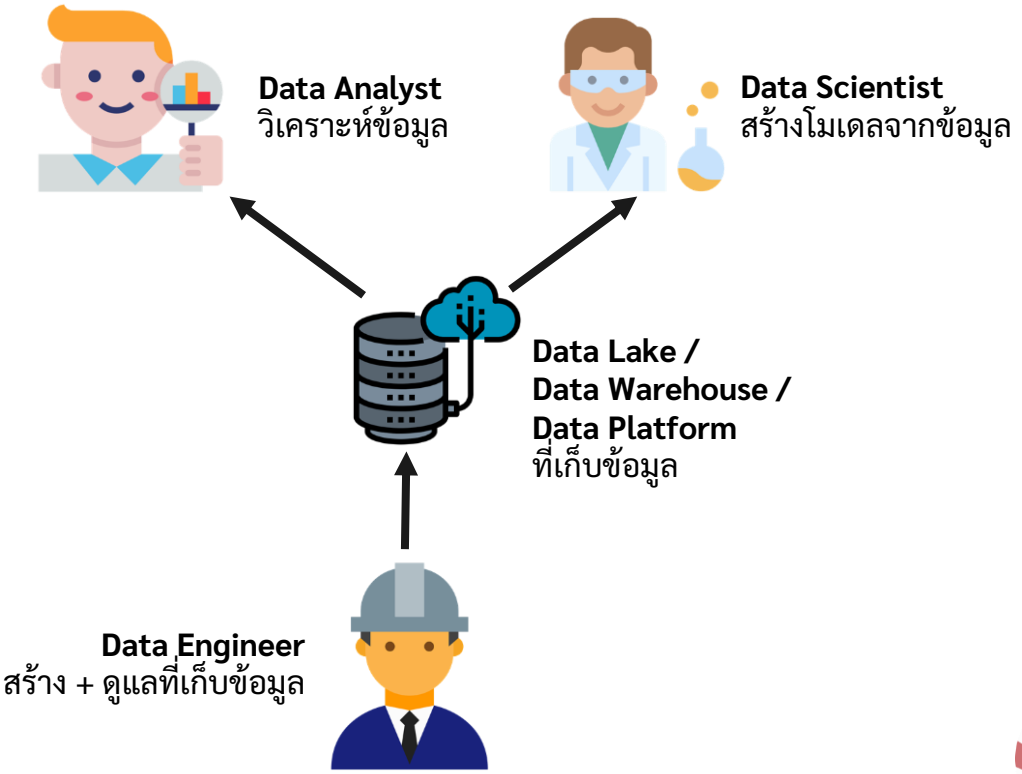
Workshop Day 2

- Introduction to Airflow
- **Workshop 4: Automated Data Pipeline with Airflow**
- Introduction to BigQuery
- **Workshop 5: Building Data Warehouse with BigQuery**
- Introduction to Google Data Studio
- **Workshop 6: Building dashboard with Google Data Studio**
- Advanced Data Engineering

Who is Data Engineer

World of Data Science

โลก Data Science มีคน
อยู่ 3 ประเภท

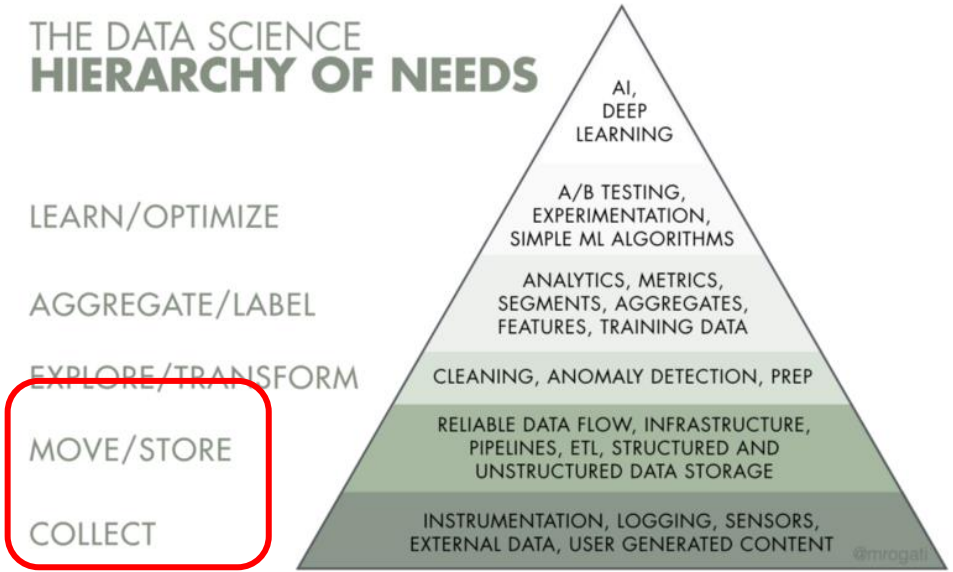


อาชีพที่ขาดไม่ได้ในองค์กรที่ต้องการทำ Data Science

“สำหรับองค์กรที่มีข้อมูลไม่ซับซ้อน ควรมี Data Engineer 2-3 คน ต่อ Data Scientist 1 คน

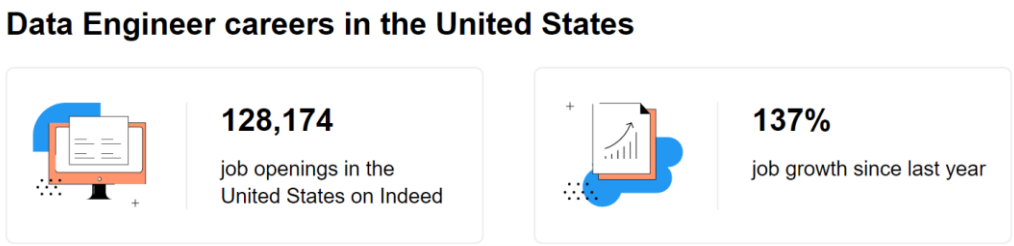
สำหรับองค์กรที่มีข้อมูลซับซ้อน ควรมี Data Engineer 4-5 คน ต่อ Data Scientist 1 คน”

Data Engineer



https://www.oreilly.com/radar/data-engineers-vs-data-scientists/
https://towardsdatascience.com/who-is-a-data-engineer-how-to-become-a-data-engineer-1167ddc12811

Data Engineer เป็นที่ต้องการของตลาด



https://www.indeed.com/career/data-engineer

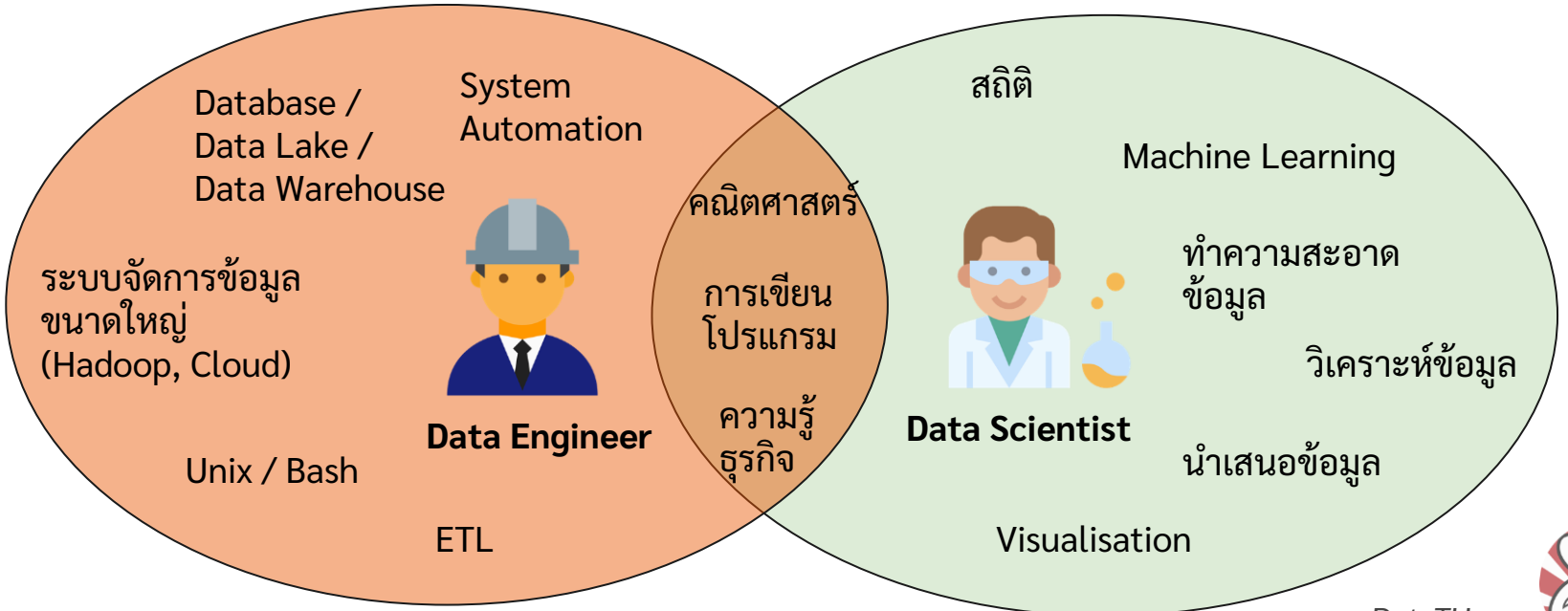
DataTH.com

งาน Data Engineer ในไทย

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

องค์กรใหญ่ ๆ ที่มีข้อมูลเยอะ ยิ่งจำเป็นต้องมีคนที่ดูแลงานด้านนี้
เช่น บริษัทในตลาดหุ้น, ธนาคาร, บริษัทประกัน, E-Commerce

จ้างคนเดียวเป็นทั้ง Data Scientist และ Data Engineer ได้มั้ย?

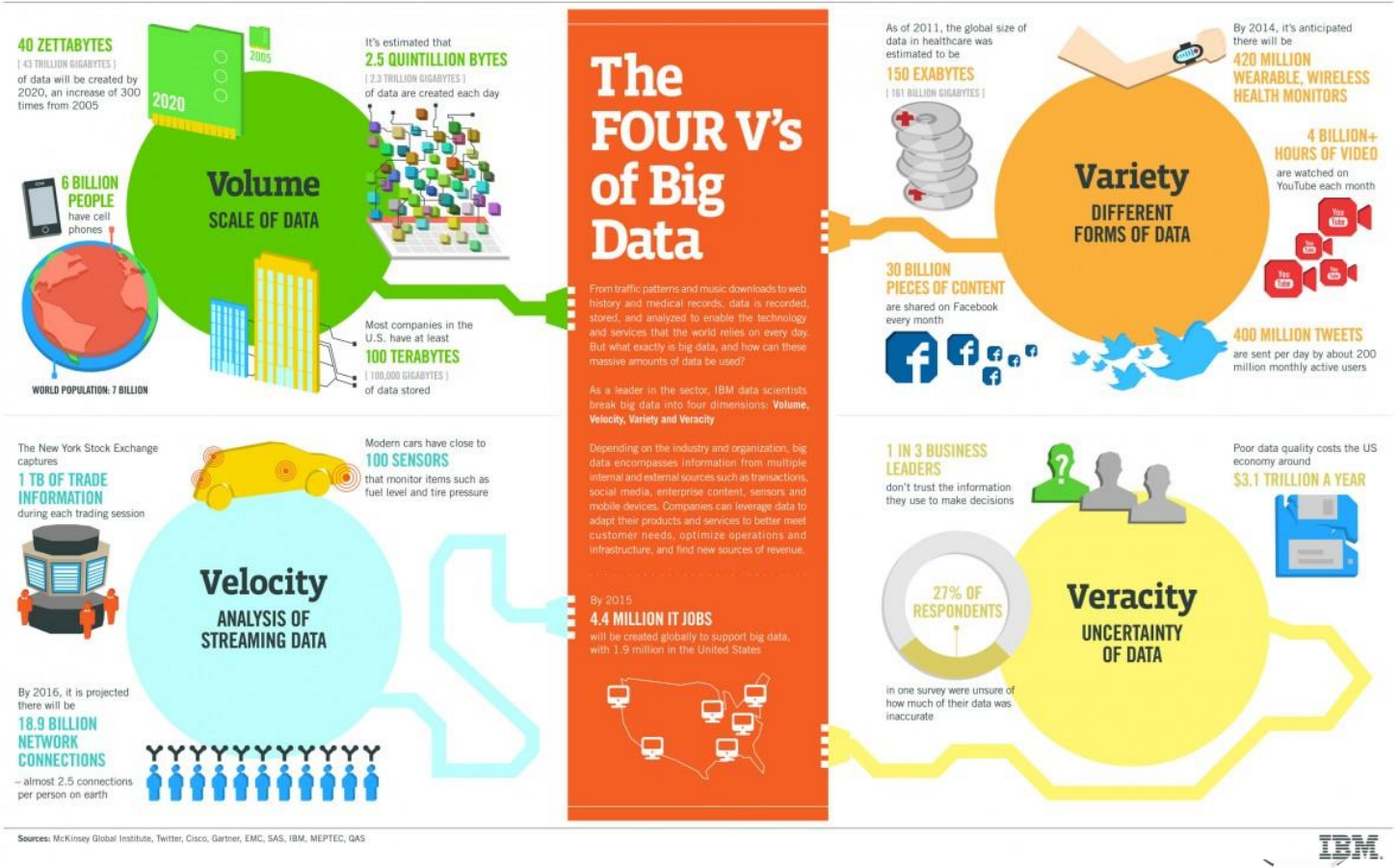


Data Engineer ต้องรู้อะไรบ้าง

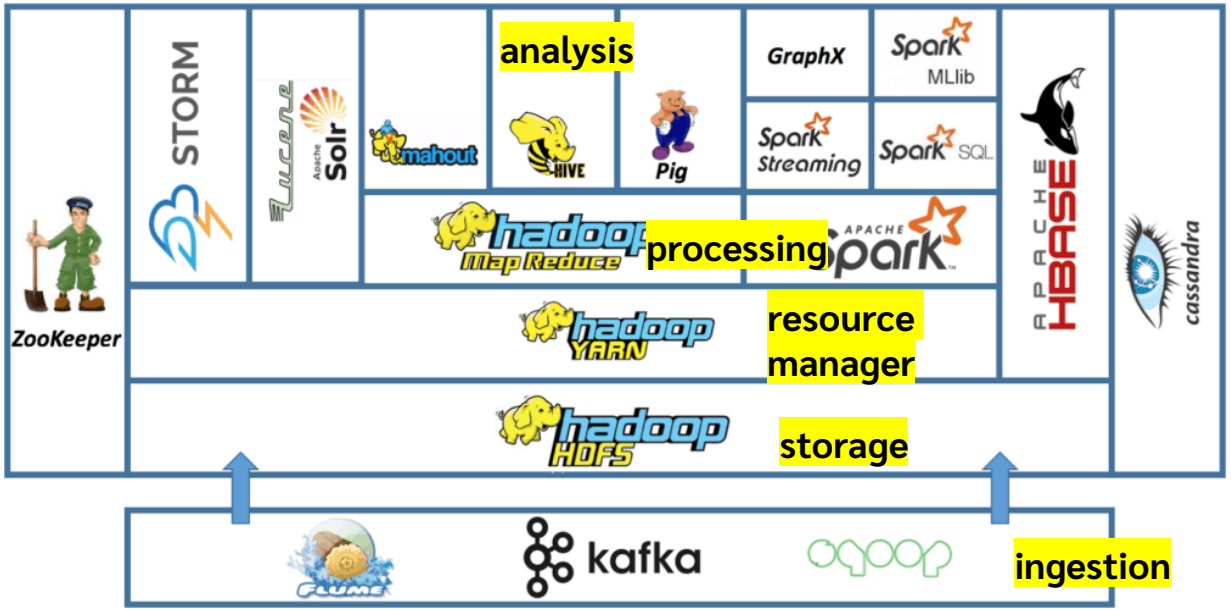
Big Data

Most popular definition:
4Vs by IBM

https://www.ibmbigdatahub.com/infographic/four-vs-big-data



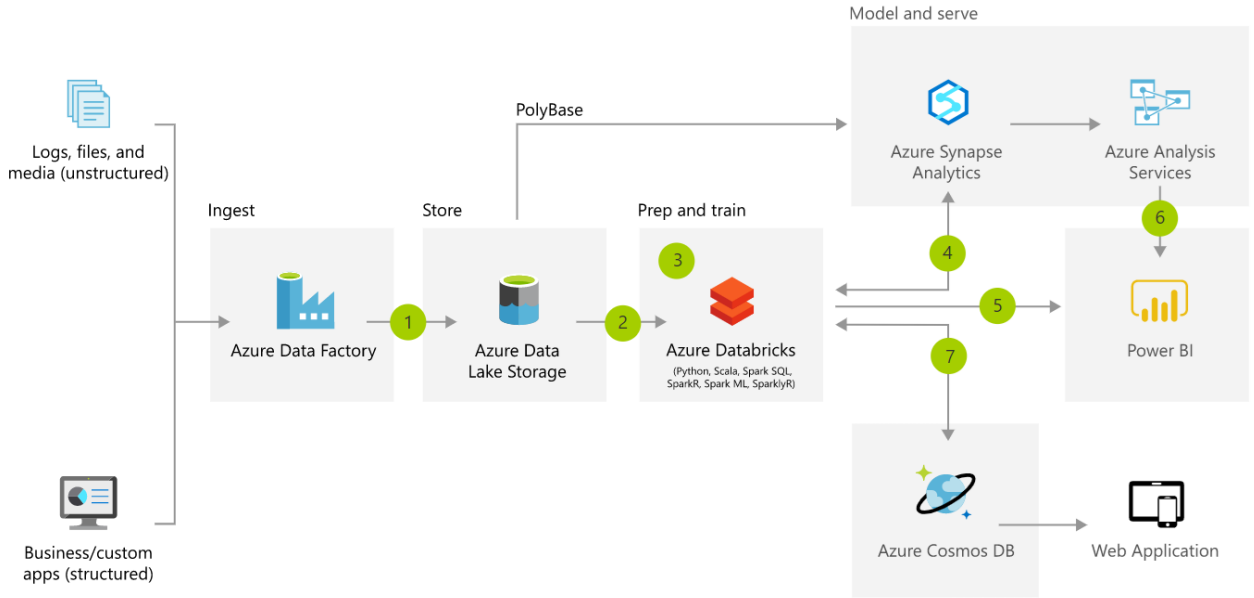
Big Data Platform (Hadoop)



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

DataTH.com

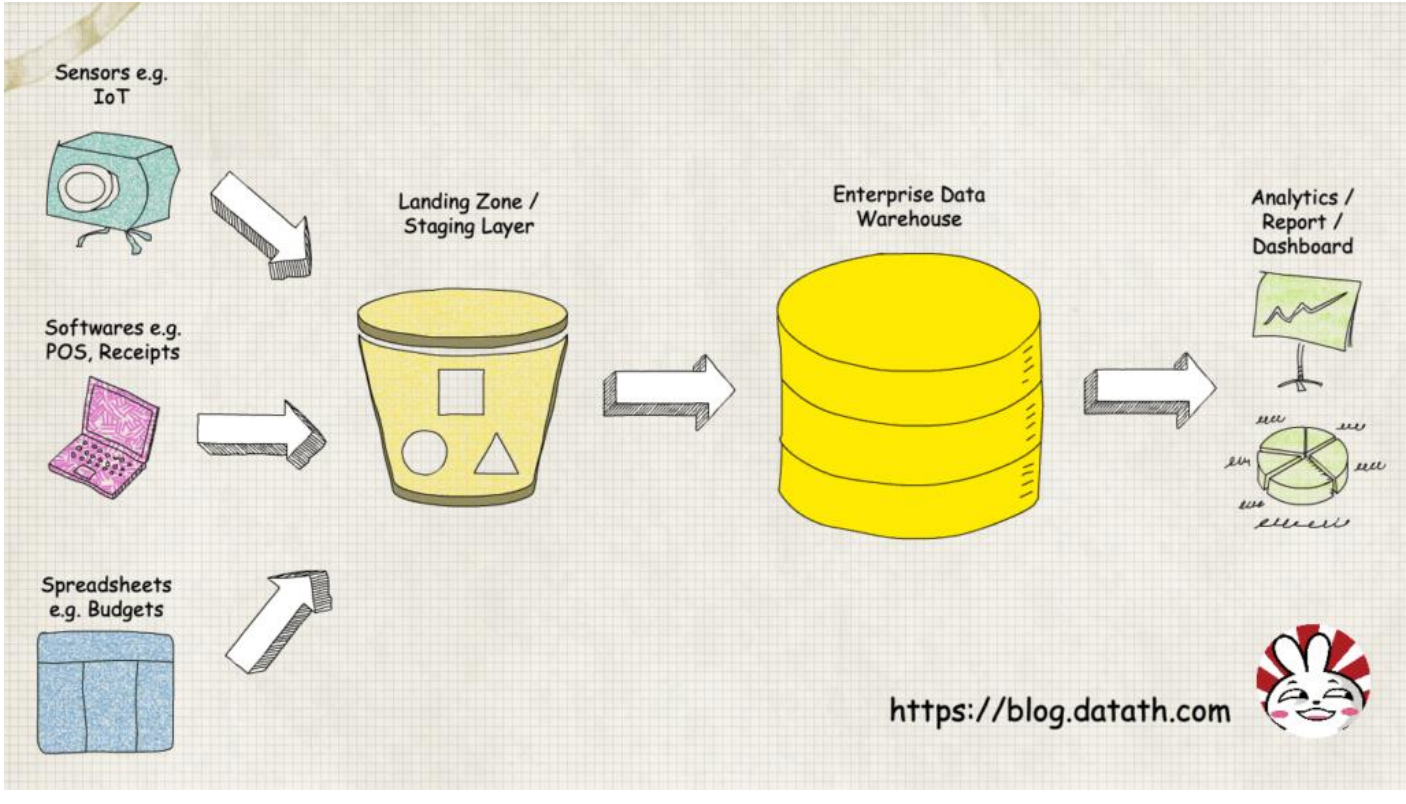
Big Data Platform (Cloud)



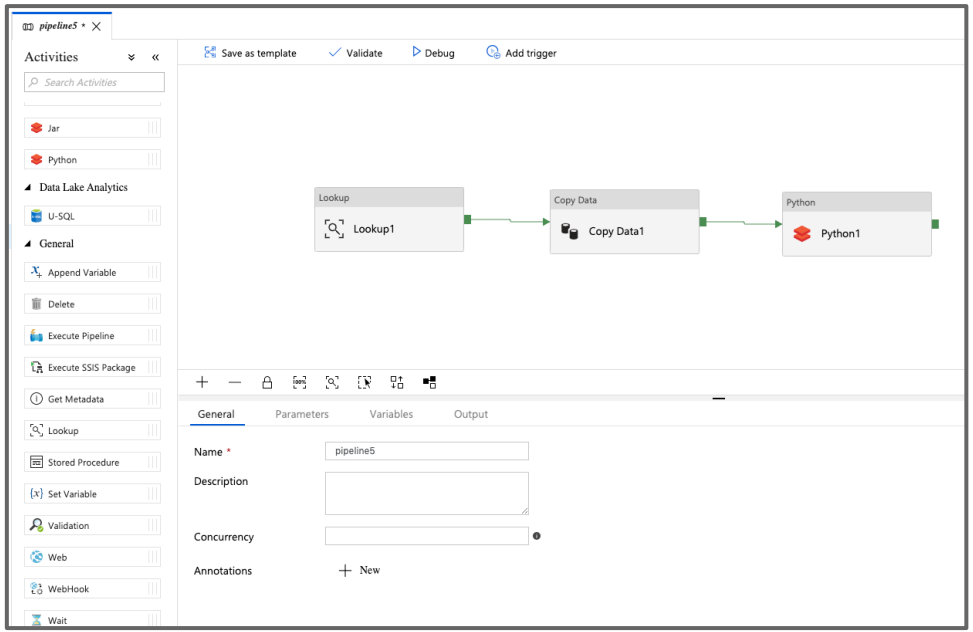
DataTH.com

Cloud (AWS, Azure, GCP)

Data Pipeline & Data Warehouse



Data Pipeline & Data Warehouse



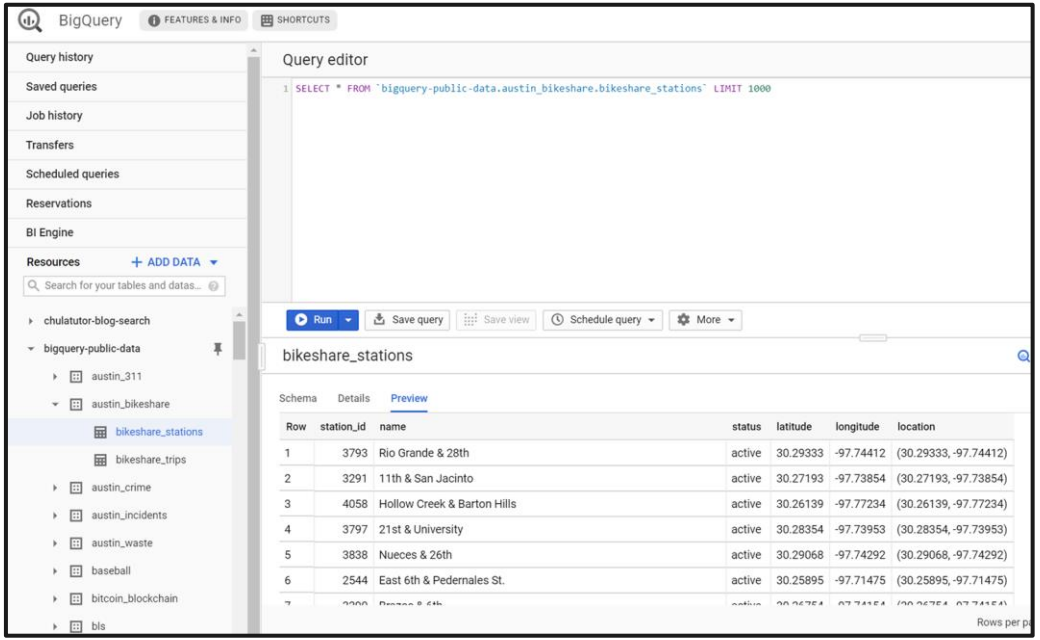
Azure Data Factory

Apache Airflow

Amazon Glue

DataTH.com

Data Pipeline & **Data Warehouse**



Amazon Redshift

Google BigQuery

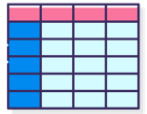
Azure Synapse
(former Azure DW SQL)

Snowflake

DataTH.com

Type of Data

1. Structured Data
2. Semi-Structured Data
3. Unstructured Data



Structured Data คืออะไร

	A	B	C	D	E	F	G	H	I
6	term	site	Popularity	Competition	Competition_paid	Traffic_score	Paid_impression_score	Share_of_voice	Percent_of_traffic
7	woocommerce ???	designil.com	27	40	0	67		15.6731876	8.17067682
8	adobe xd	designil.com	59	66	9	61		0.03968074	2.73210345
9	bootstrap 4	designil.com	60	32	0	61		0.03243026	2.52159489
10	??? woocommerce	designil.com	28	23	1	60		3.40286743	2.33023584
11	font thai	designil.com	23	47	0	60		9.81897263	2.09183006
12	lorem ipsum	designil.com	68	63	0	60		0.012835	1.99960353
13	font ???	designil.com	22	47	0	58		8.73875751	1.60369875
14	infographic	designil.com	55	34	6	58		0.07675551	1.35195664
15	bounce rate ???	designil.com	9	45	0	57		45.14197988	1.32611949
16	google font thai	designil.com	20	24	0	57		5.43834797	1.25140507
17	thai font	designil.com	25	57	0	57		4.65907223	1.13315291
18	woocommerce	designil.com	62	72	28	56		0.02017068	0.94929497
19	??????	designil.com	22	22	5	56		4.33455945	0.91912392
20	bootstrap	designil.com	68	17	0	55		0.00602208	0.84748578
21	getty image ????	designil.com	6	11	0	55		26.7051048	0.77074522
22	jquery	designil.com	62	52	0	55		0.01405276	0.75528394
23	responsive web design	designil.com	42	64	22	54		0.13143857	0.70343553
24	lorem ipsum thai	designil.com	14	44	0	54		8.1528485	0.68265439
25	landing page	designil.com	54	63	42	54		0.02484704	0.62524488
26	thai font for mac	designil.com	11	51	0	54		22.87633838	0.61262878
27	???????????????? html	designil.com	6	20	0	54		31.97742702	0.60416615
28	sukhumvit font	designil.com	12	47	0	53		9.87686444	0.59991033
29	infographic ???	designil.com	10	43	0	53		13.679723	0.58999238
30	???????? wordpress	designil.com	14	46	5	53		10.23742658	0.57643415
31	font thai download free	designil.com	9	51	0	53		27.21822853	0.5711661
32	responsive	designil.com	45	68	1	53		0.11200085	0.56716538
33	wireframe	designil.com	50	59	14	53		0.07768544	0.55139784
34	sukhumvit set font	designil.com	14	55	0	53		5.43973756	0.55101134
35	page speed	designil.com	54	34	1	53		0.01378355	0.54127025
36	???????? ???	designil.com	12	47	0	53		10.61349979	0.53552906

DataTH.com

Semi-Structured Data คืออะไร

ข้อมูลแบบที่มีโครงสร้างข้อมูลที่มีความยืดหยุ่น สามารถขยายโครงสร้างข้อมูลได้ในอนาคต และเรียกใช้ได้รวดเร็ว

เหมาะสำหรับ: ข้อมูลที่สามารถปรับเปลี่ยนโครงสร้างได้ตลอดเวลา

เช่น Key-value, Document, JSON, XML, Graph

NoSQL

```
{
  "users": [
    {
      "userId": 1,
      "firstName": "Adam",
      "lastName": "Lee",
      "emailAddress": "adam@gmail.com"
    },
    {
      "userId": 2,
      "firstName": "John",
      "lastName": "Doe",
      "phoneNumber": "0406817201",
      "emailAddress": "john.doe@gmail.com"
    }
    ]
}
```

Unstructured Data คืออะไร

ไฟล์

รูปภาพ

วีดิโอ

เสียง

เก็บใน
Data Lake

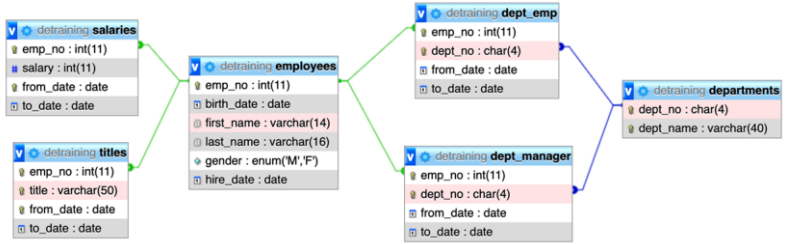
SQL Database

หรือที่รู้จักกันในนาม **RDBMS** (Relational Database Management System) คือ ฐานข้อมูลสำหรับเก็บ **Structured Data** สามารถ Query ด้วย **SQL** ประกอบด้วย

- **Table** (ตารางข้อมูล)
- **Schema** (โครงสร้างตาราง)
- **Relation** (ความสัมพันธ์ระหว่างตาราง)
- **Primary Key** (Unique key ที่ใช้อ้างอิงในแต่ละแถว)

Name	Type	Collation	Attributes	Null	Default
emp_no	int(11)			No	*None*
birth_date	date			No	*None*
first_name	varchar(14)	utf8mb4_0900_ai_ci		No	*None*
last_name	varchar(16)	utf8mb4_0900_ai_ci		No	*None*
gender	enum('M', 'F')	utf8mb4_0900_ai_ci		No	*None*
hire_date	date			No	*None*

Schema



Relation

[DB] [DW]

Structured Data Storage แบบ OLTP VS OLAP

OLTP (On-Line Transaction Processing)

- ออกแบบมาสำหรับการเขียนข้อมูล
- เหมาะกับการเก็บข้อมูลของเว็บไซต์หรือระบบ ที่มีคนใช้งานตลอดเวลา (hot data)
- ควรทำการ normalize เพื่อลดการเก็บข้อมูลซ้ำซ้อน และจัดเก็บข้อมูลอย่างมีประสิทธิภาพ

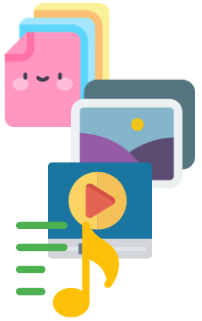
OLAP (On-Line Analytical Processing)

- ออกแบบมาสำหรับการอ่านข้อมูล
- เหมาะกับการวิเคราะห์ข้อมูลที่ซับซ้อน โดย Data Analyst และ Data Scientist
- ควรทำการ de-normalize เพื่อความรวดเร็ว ในการ query ข้อมูล และลดเวลาการประมวลผล

Database VS Data Lake VS Data Warehouse

Database
Semi-Structured Data
&
Structured data

Data Lake



Unstructured Data

ข้อมูล Structured (SQL)
- เก็บใน SQL Database, Data Lake, Data Warehouse

ข้อมูล Semi-structured (NoSQL), Unstructured
- เก็บใน NoSQL Database หรือ Data Lake
- Data Warehouse บางตัวเก็บ Semi-structured ได้แล้ว

Data Warehouse, Data Mart

NoSQL Database

ฐานข้อมูลอีกประเภท ที่เกิดขึ้นมาสำหรับ Semi-Structured Data

NoSQL = Not Only SQL หมายถึงว่า Database แบบ NoSQL บางตัวก็อ่าน SQL ได้

เช่น MongoDB, Neo4J

Document stores

Key-value stores

Wide column stores

Graph DBMS

Database Models ยังมีอีกมากมาย

- Relational DBMS
- Key-value stores
- Document stores
- Graph DBMS
- Time Series DBMS
- Object oriented DBMS
- Search engines
- RDF stores (triplestore)
- Multivalue DBMS
- Wide column stores
- Native XML DBMS
- Event Stores
- Content stores
- Navigational DBMS

Source: https://db-engines.com/en/ranking

DB Ranking

DB-ENGINES

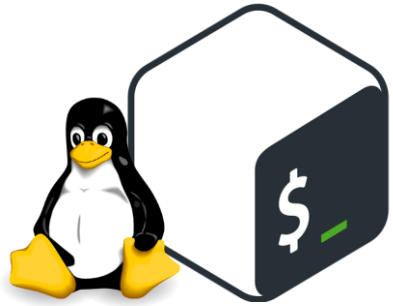
350 systems in ranking, February 2020

Rank Feb 2020	Rank Jan 2020	Rank Feb 2019	DBMS	Database Model	Score Feb 2020	Score Jan 2020	Score Feb 2019
1.	1.	1.	Oracle	Relational, Multi-model	1344.75	-1.93	+80.73
2.	2.	2.	MySQL	Relational, Multi-model	1267.65	-7.00	+100.36
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	1093.75	-4.80	+53.69
4.	4.	4.	PostgreSQL	Relational, Multi-model	506.94	-0.25	+33.38
5.	5.	5.	MongoDB	Document, Multi-model	433.33	+6.37	+38.24
6.	6.	6.	IBM Db2	Relational, Multi-model	165.55	-3.15	-13.87
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model	152.16	+0.72	+6.91
8.	8.	↓ 7.	Redis	Key-value, Multi-model	151.42	+2.67	+1.97
9.	9.	9.	Microsoft Access	Relational	128.06	-0.52	-15.96
10.	10.	10.	SQLite	Relational	123.36	+1.22	-2.81
11.	11.	11.	Cassandra	Wide column	120.36	-0.31	-3.02
12.	12.	↑ 13.	Splunk	Search engine	88.77	+0.10	+5.96
13.	13.	↓ 12.	MariaDB	Relational, Multi-model	87.34	-0.11	+3.91
14.	14.	↑ 15.	Hive	Relational	83.53	-0.71	+11.25
15.	15.	↓ 14.	Teradata	Relational, Multi-model	76.81	-1.48	+0.84

DataTH.com

Programming

Python



Bash Script

Java or Scala

Software Engineer & Automation

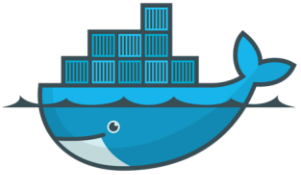
Software Engineering
Git, Container (Docker, K8s)

Automation
Airflow, Terraform

Apache Airflow

Terraform

Git



Docker

Kubernetes

Technology Stack overview

Data Lake

Amazon S3 | Google Cloud Storage | Azure Blob Storage

Data Warehouse

Google BigQuery

Azure Synapse

Amazon Redshift

Snowflake

Data Processing

Apache Hive

Apache Spark

Apache Beam

Amazon Glue

Cloud Dataflow

Cloud Dataproc

Data Pipeline

Azure Data Factory

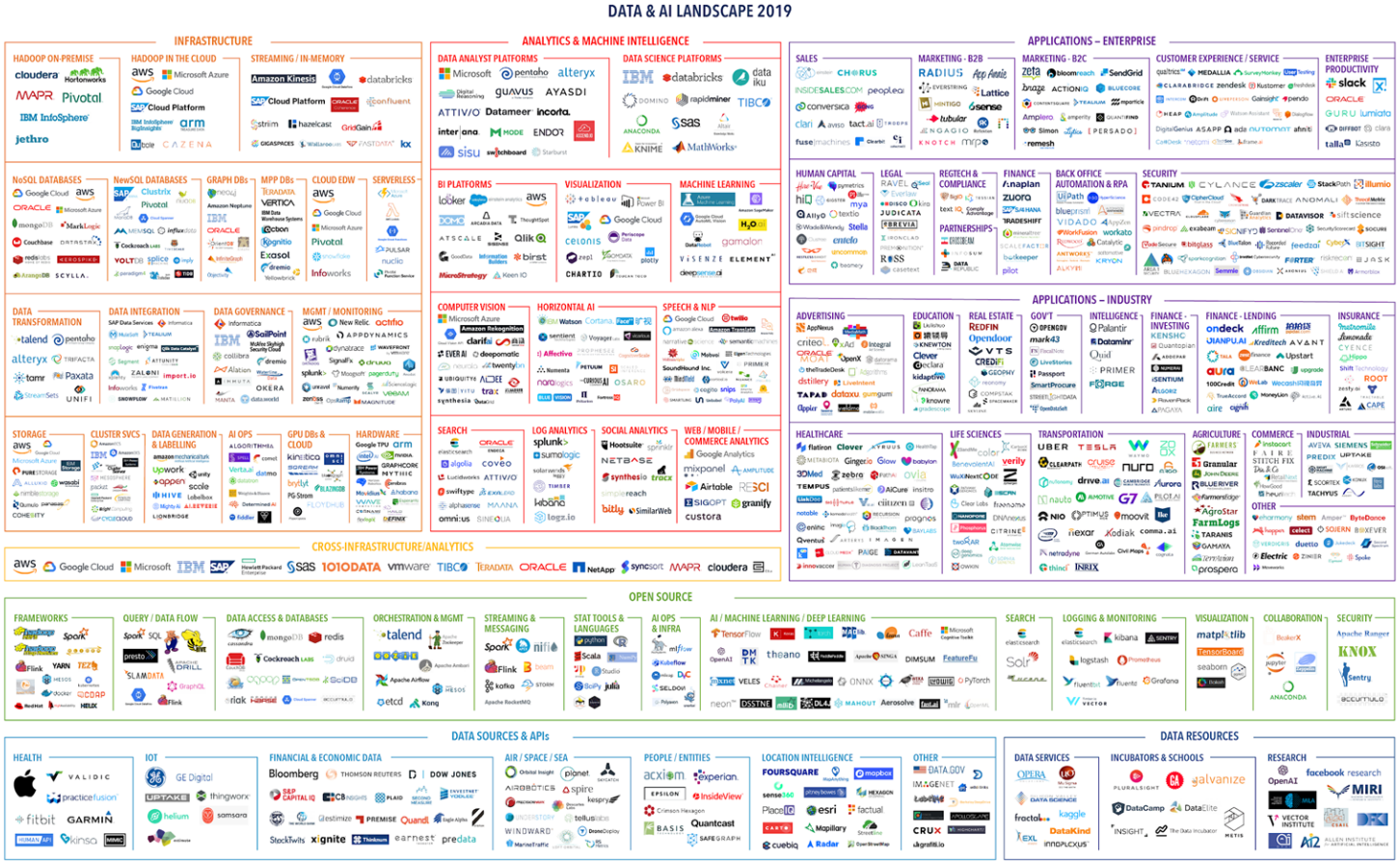
Apache Airflow

Apache Oozie

Luigi

Data & AI Landscape 2019

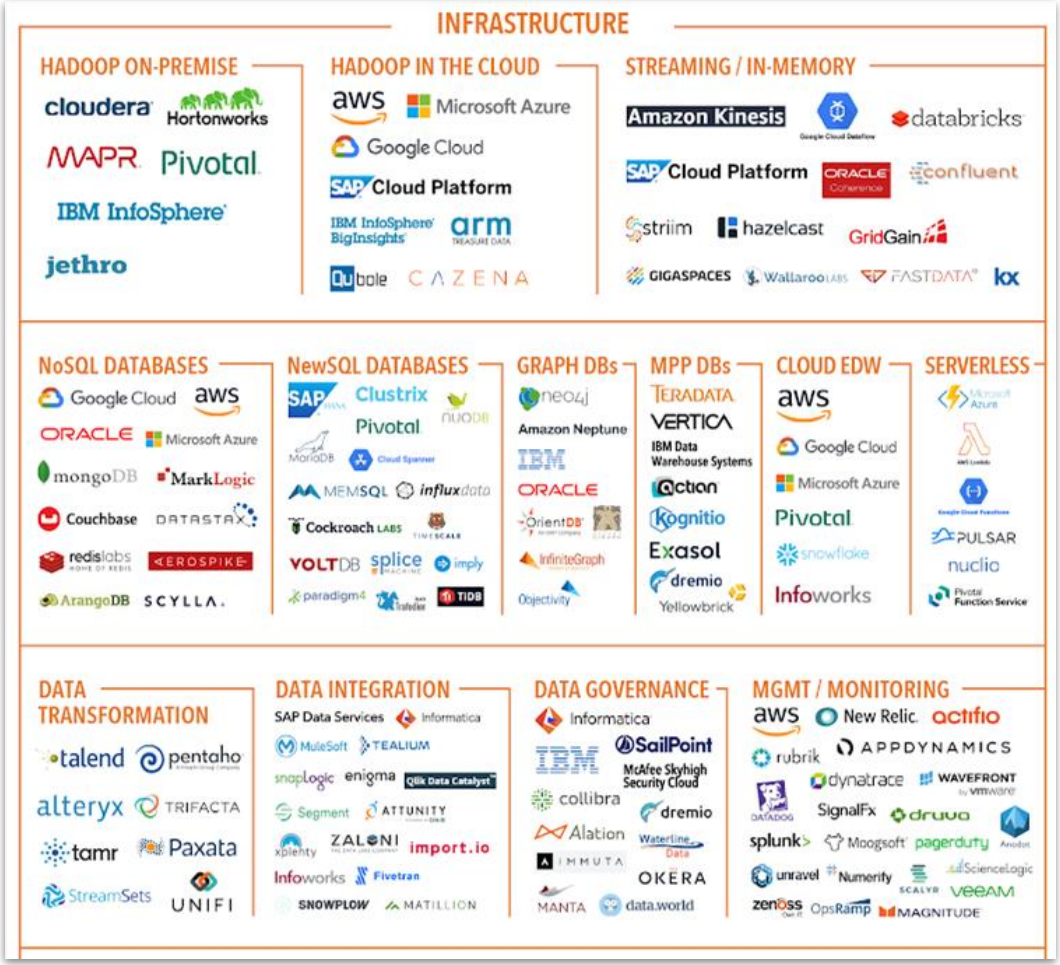
Source: https://mattturck.com/data2019/



FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Data & AI Landscape 2019 (zoom in)

Source: https://mattturck.com/data2019/



สรุป อยากเป็น Data Engineer ต้องรู้อะไรบ้าง

1. Big Data Platform (Hadoop & Cloud)
2. Data Pipeline
3. Data Storage - DB, DW, DL
4. Software Engineering
5. Automation