

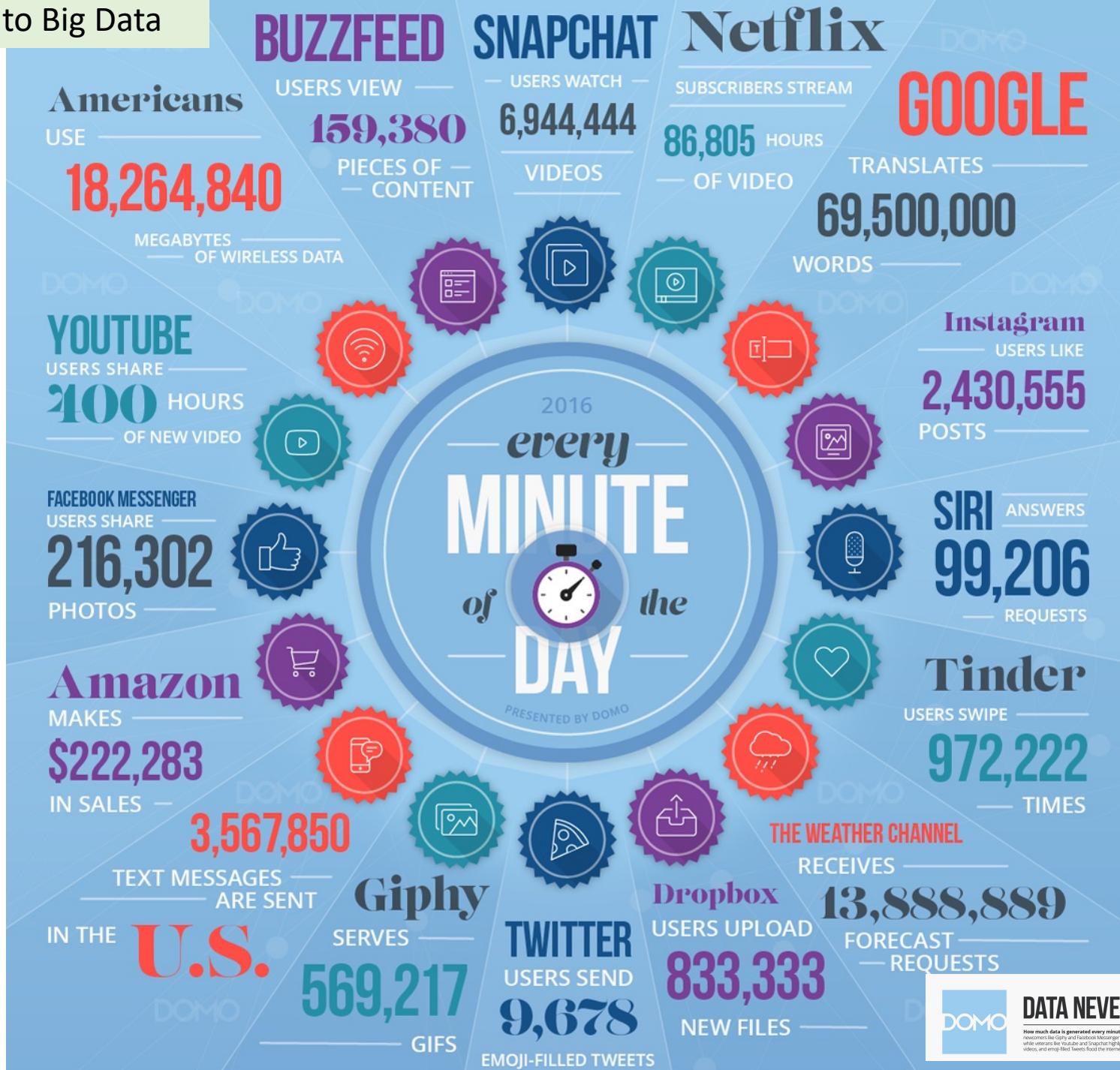
Introduction to Big Data Technologies

Credit to Peerapon Vateekul

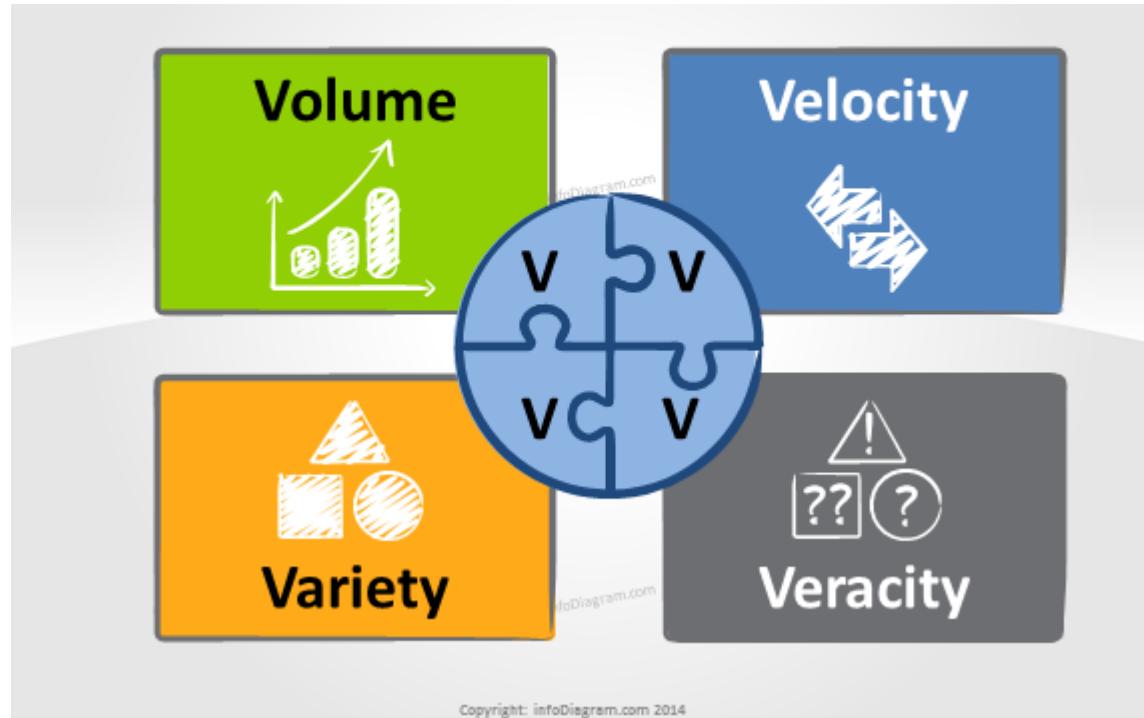
<https://github.com/kaopanboonyuen/GISTDA2022>

Outline

- Introduction to Big Data
 - Big Data Definition
 - Big Data Landscape
 - Internet of Things (IoT)
- Big Data Analytics Process
 - Big Data Infrastructure
 - Big Data Analytics
- Big Data Ecosystem



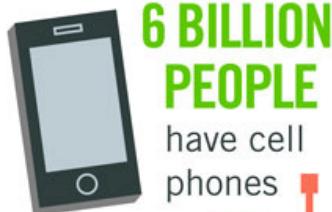
Big Data Definition



40 ZETTABYTES

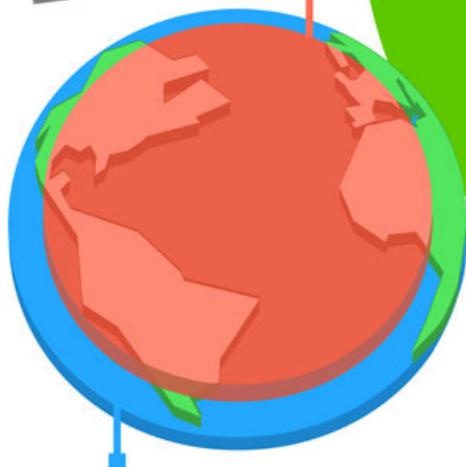
[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005



**6 BILLION
PEOPLE**

have cell phones



WORLD POPULATION: 7 BILLION



Volume SCALE OF DATA

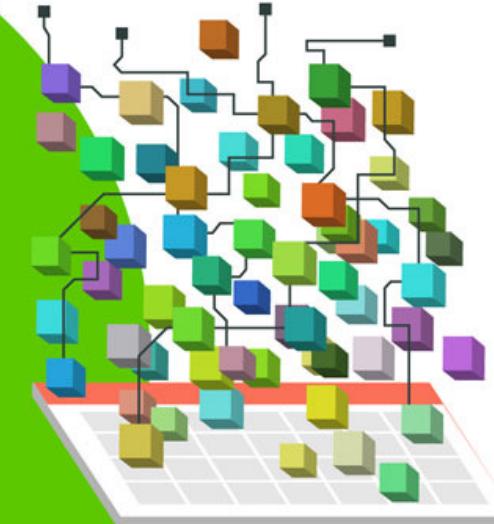


It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



Most companies in the
U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Megabyte	1,000,000 bytes
Gigabyte	1,000,000,000 bytes
Terabyte	1,000,000,000,000 bytes
Petabyte	1,000,000,000,000,000 bytes
Exabyte	1,000,000,000,000,000,000 bytes
Zettabyte	1,000,000,000,000,000,000,000 bytes
Yottabyte	1,000,000,000,000,000,000,000,000 bytes

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

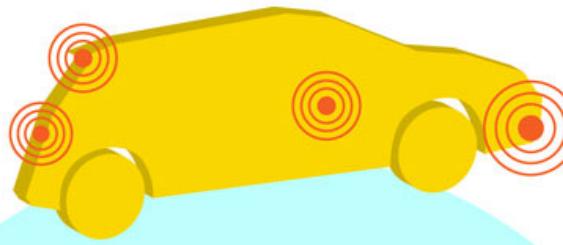
during each trading session



By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

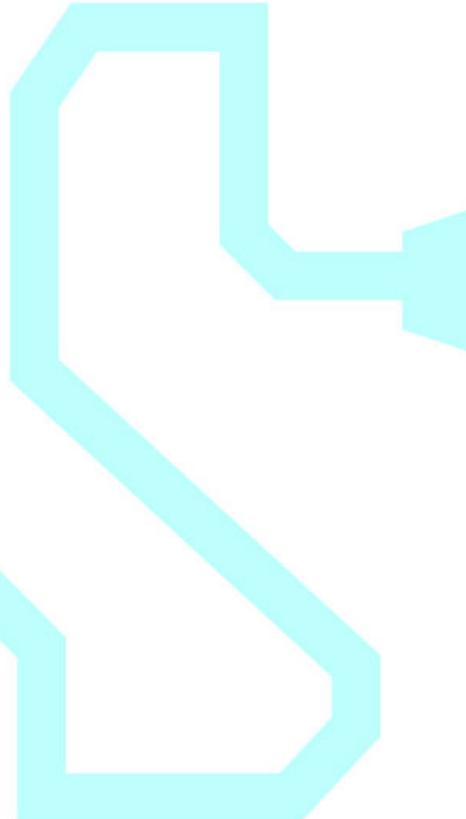
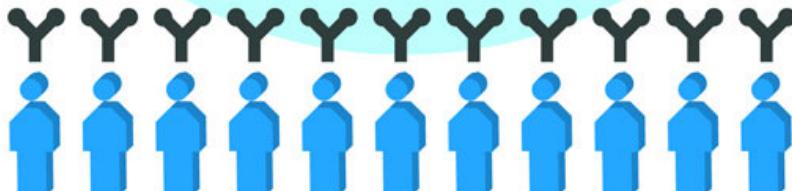
– almost 2.5 connections per person on earth



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

Velocity

ANALYSIS OF STREAMING DATA



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**

are shared on Facebook every month



Variety DIFFERENT FORMS OF DATA

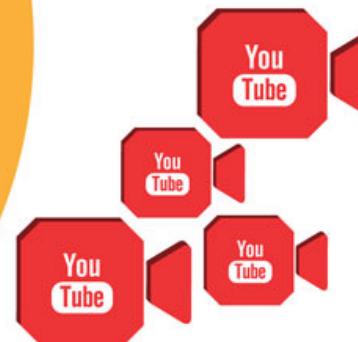


By 2014, it's anticipated there will be

**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**

**4 BILLION+
HOURS OF VIDEO**

are watched on YouTube each month



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

1 IN 3 BUSINESS LEADERS

don't trust the information
they use to make decisions



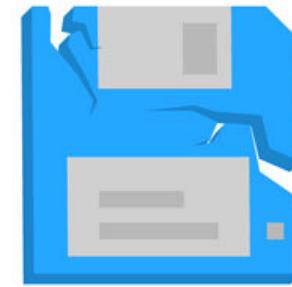
27% OF
RESPONDENTS

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

Veracity

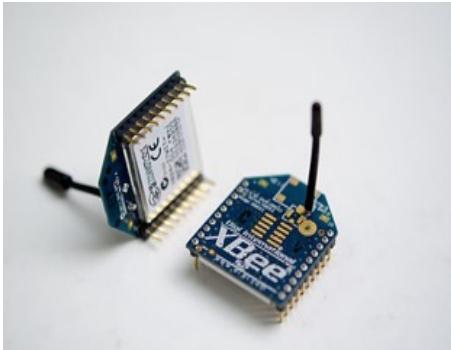
UNCERTAINTY OF DATA

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



Internet of Things (IoT)

- Currently physical world and software worlds are detached
- Internet of things promises to bridge this
 - It is about sensors and actuators everywhere
 - In your fridge, in your blanket, in your chair, in your carpet.. Yes even in your socks
 - Umbrella that light up when there is rain and medicine cups



Big Data Goal



Predict
(Supervised Learning)



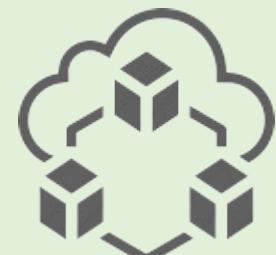
Discover
(Unsupervised Learning)



Experiment



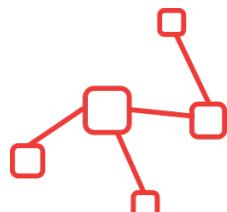
Explore
Story-telling



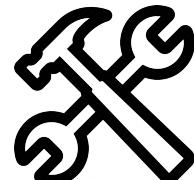
Big Data

Why Big Data is hard?

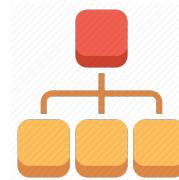
- How store? Assuming 1TB bytes it takes **1,000 computers** to store a 1PB
- How to move? Assuming 10Gb network, it takes **2 hours** to copy 1TB, or **83 days** to copy a 1PB
- How to search? Assuming each record is 1KB and one machine can process 1,000 records per sec, it needs **277CPU days** to process a 1TB and **785 CPU years** to process a 1 PB



New Infrastructure



New Tools
NoSQL, Analytics, BI

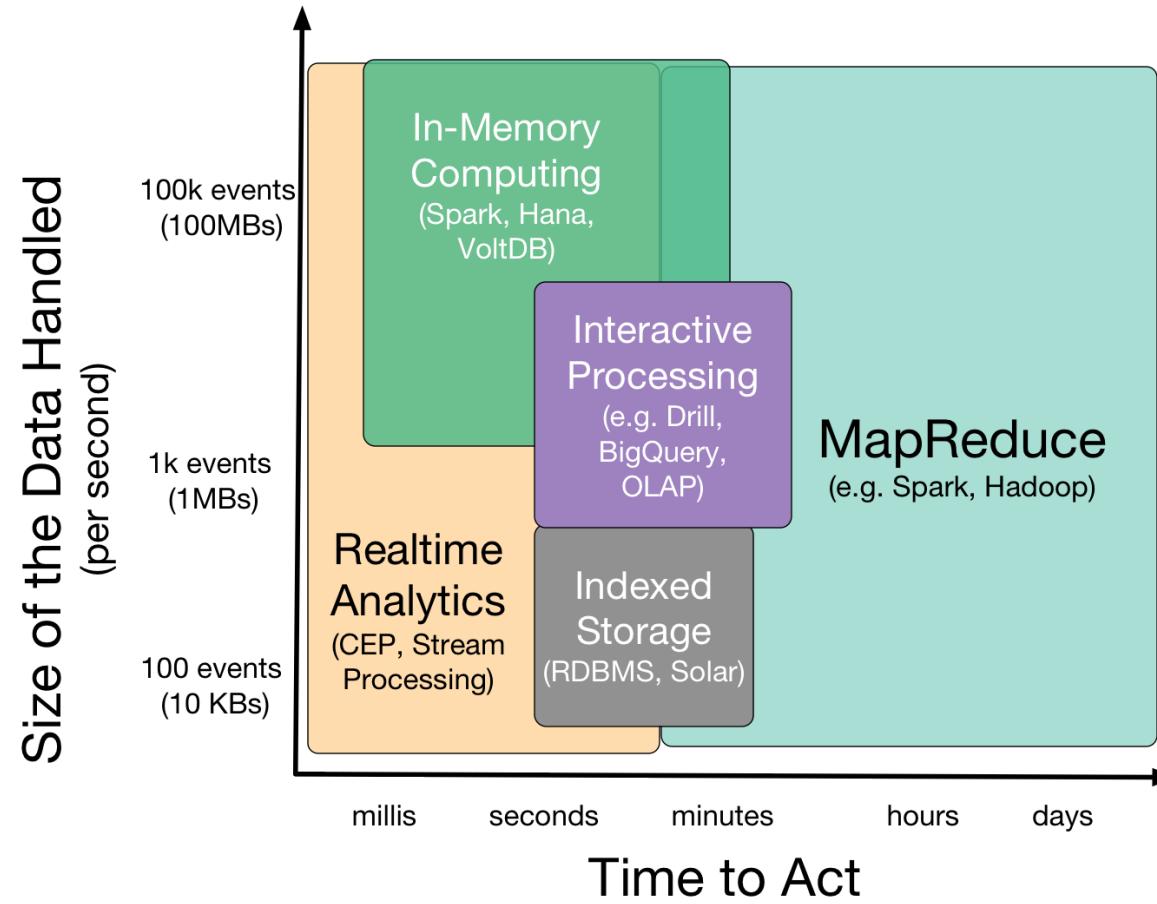


New Algorithms

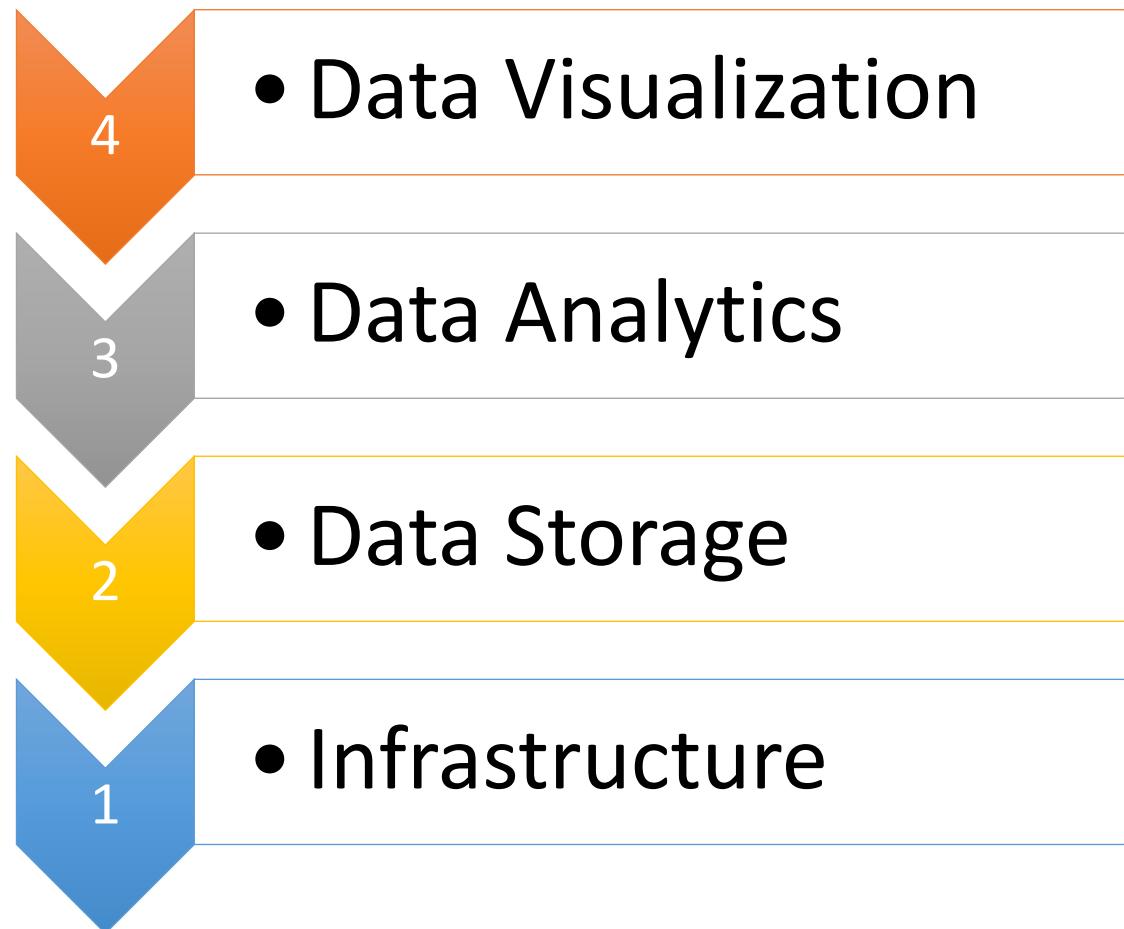
BIG DATA & AI LANDSCAPE 2018



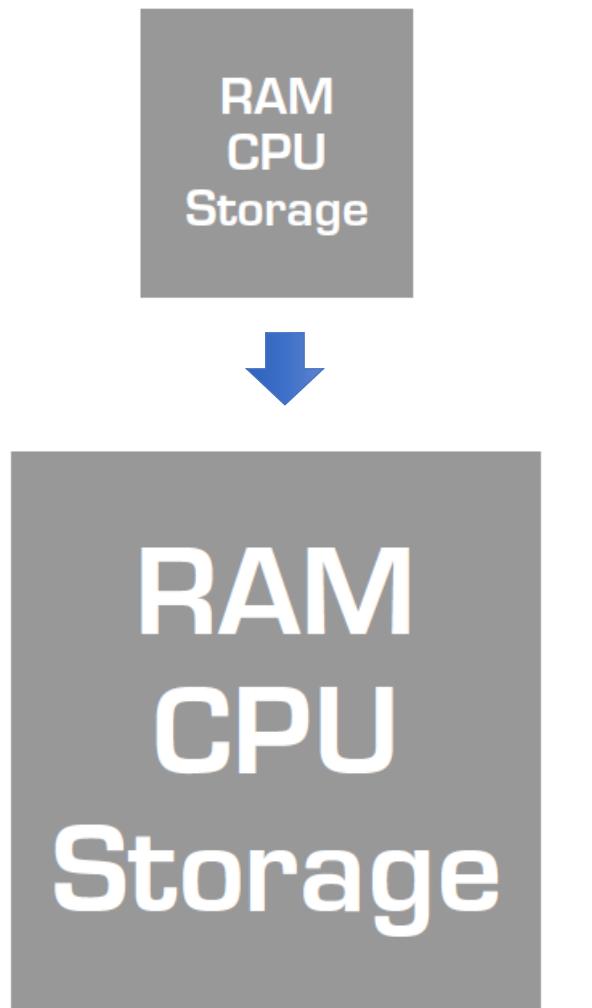
Big Data Processing Technologies Landscape



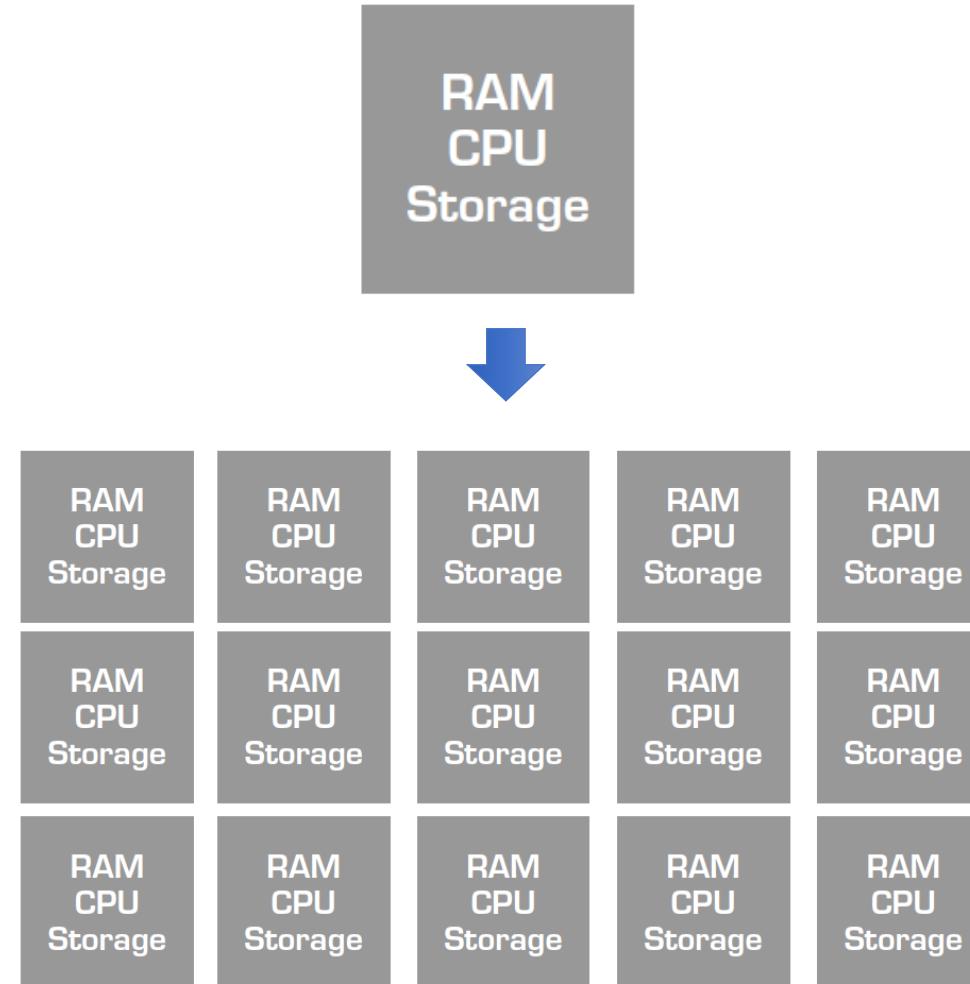
Big Data Analytics Process



Infrastructure



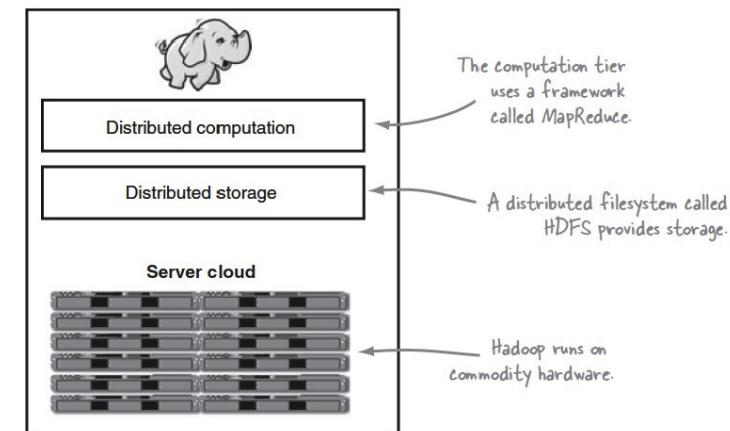
Vertical Scaling
(Scale-up)



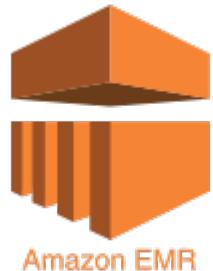
Horizontal Scaling
(Scale-out)

What is Hadoop?

- A scalable fault-tolerant **distributed system** for (1) data storage and (2) processing
- Completely written in java
- Open source & distributed under Apache license
- Two main components
 - Map/Reduce System
 - Hadoop Distributed File System (HDFS)



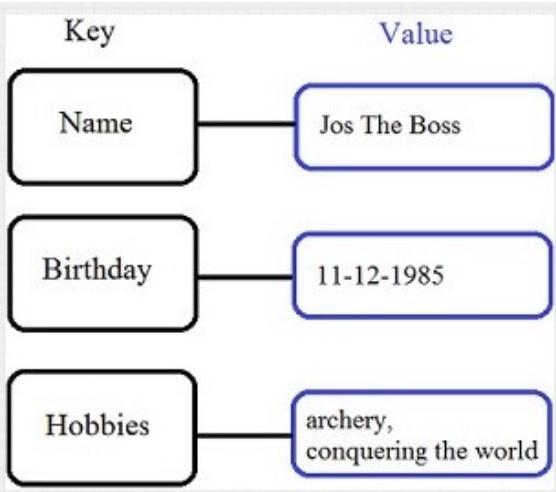
Hadoop Distribution



Google Cloud Platform

NoSQL

Key-Value Stores



Column Stores

ROWID	Name	Birthday	Hobbies
1	Jos The Boss	11-12-1985	archery, conquering the world
2	Fritz von Braun	27-1-1978	building things, surfing
3	Freddy Stark		swordplay, lollygagging, archery
4	Delphine Thewiseone	16-9-1986	

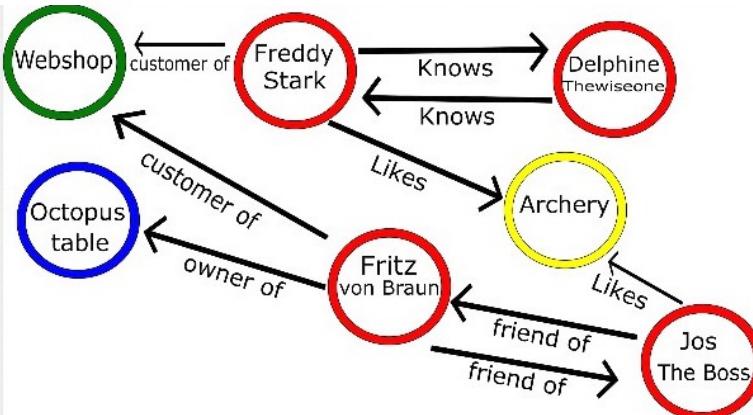
Row-oriented lookup: from top to bottom and for every entry all columns are taken in memory.



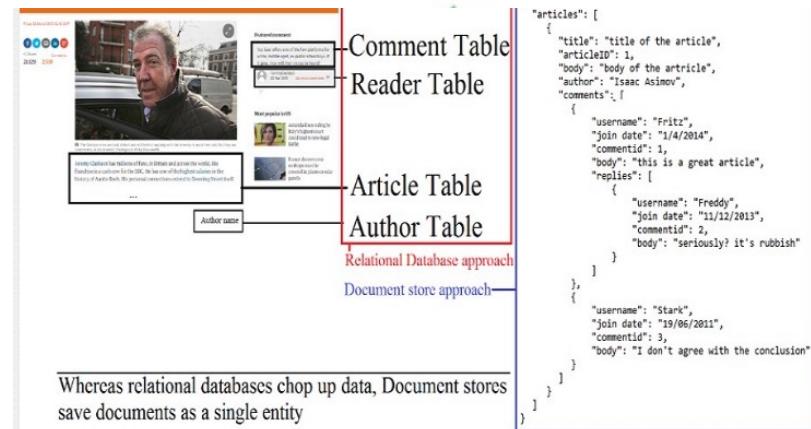
Name	ROWID	Birthday	ROWID	Hobbies	ROWID
Jos The Boss	1	11-12-1985	1	archery	1, 3
Fritz Schneider	2	27-1-1978	2	conquering the world	1
Freddy Stark	3			building things	2
Delphine Thewiseone	4	16-9-1986	4	surfing	2
				swordplay	3
				lollygagging	3

A column-oriented database stores each column separately

Graph Databases

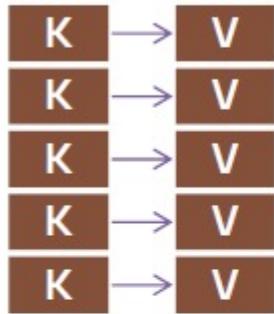


Document Stores

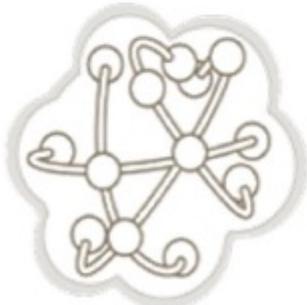


NoSQL (cont.)

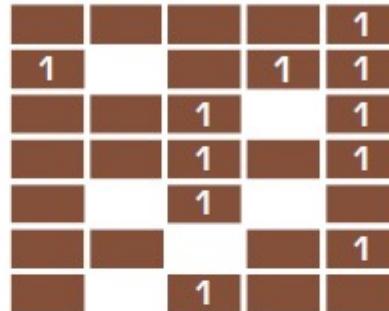
Key-Value Stores



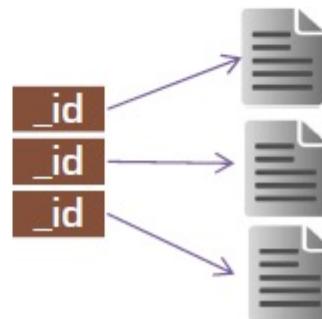
Graph Databases



Column Stores



Document Stores



Spark



- Apache Spark is a general-purpose cluster **in-memory computing** system (**no data storage**)
- Support Hadoop environment
- Provide high-level APIs in Java, Scala, and Python
- Provide optimized engine that supports general execution graphs
- Provide various level tools, e.g., Spark SQL, MLlib

Up to **10x** faster on disk,
100x in memory

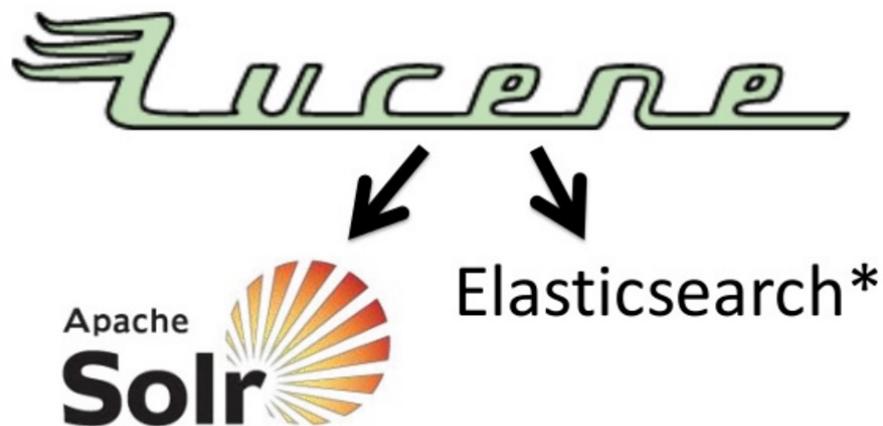


Data Analytics & Data Visualization

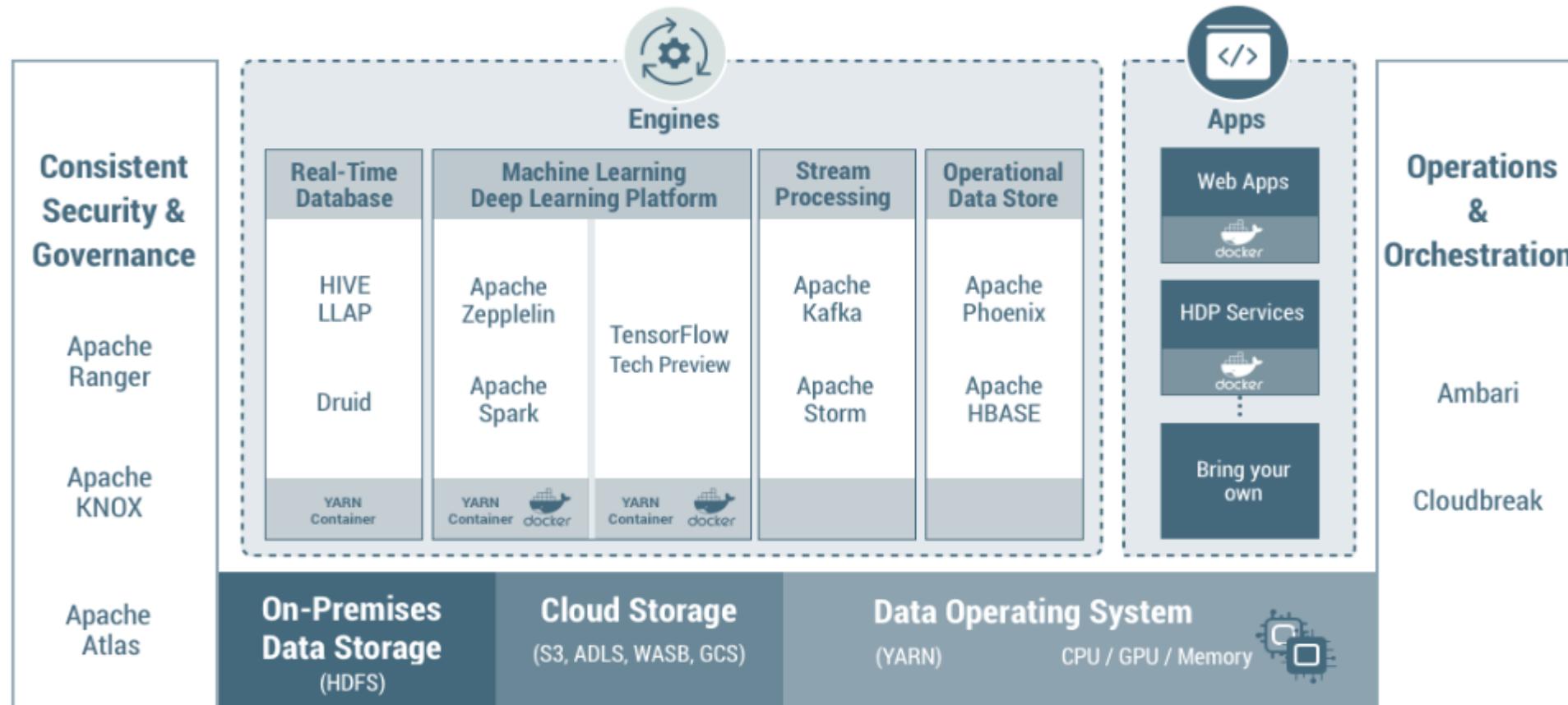


Search

- Open-source, broadly-distributed, readily scalable search engine
- Fast direct access to the data
- To achieve fast search responses because, instead of searching the text directly, it searches an index instead.



Big Data Ecosystem



Large, Shared Workloads, Multi-Tenant Clusters

