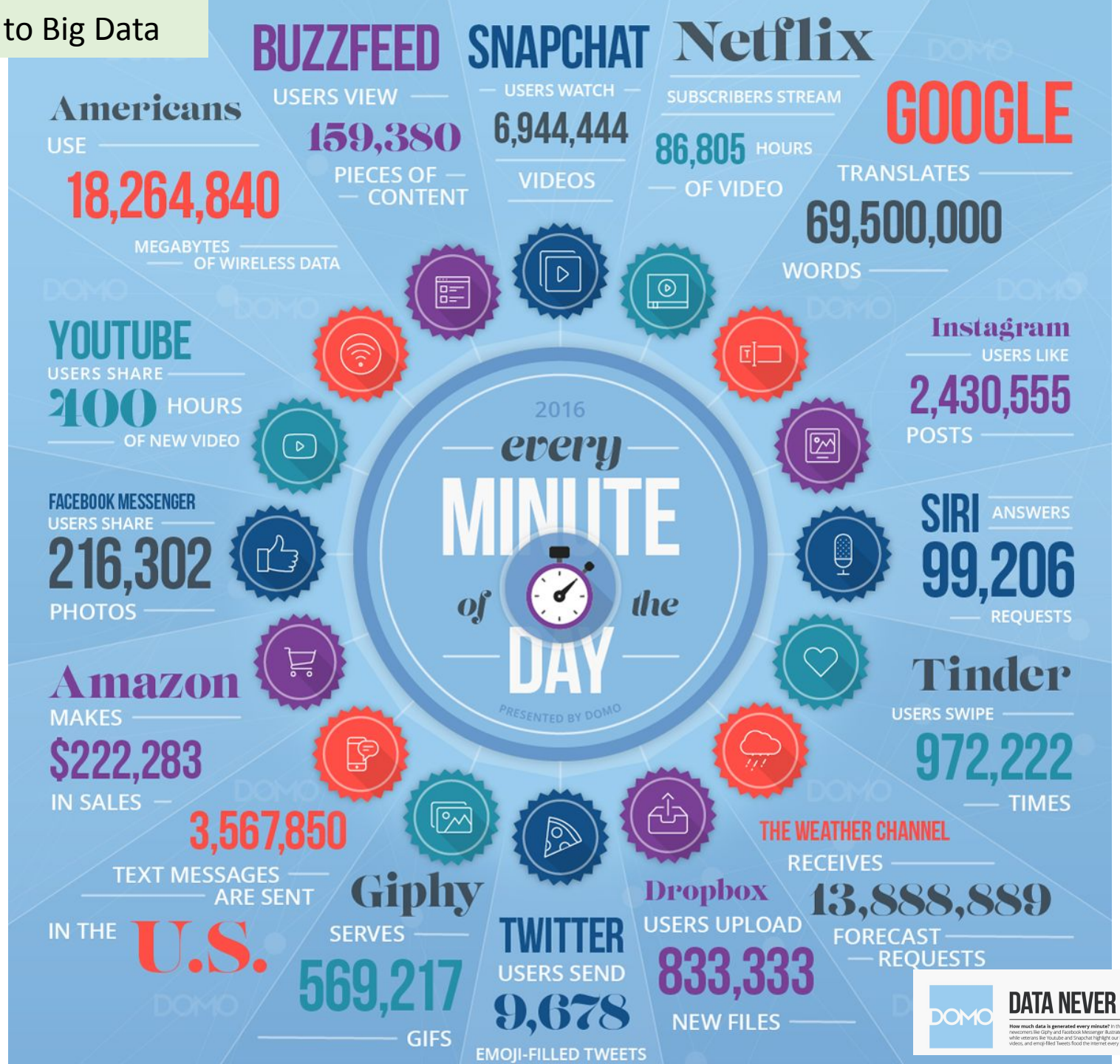
Introduction to Big Data Technologies

Credit to Peerapon Vateekul

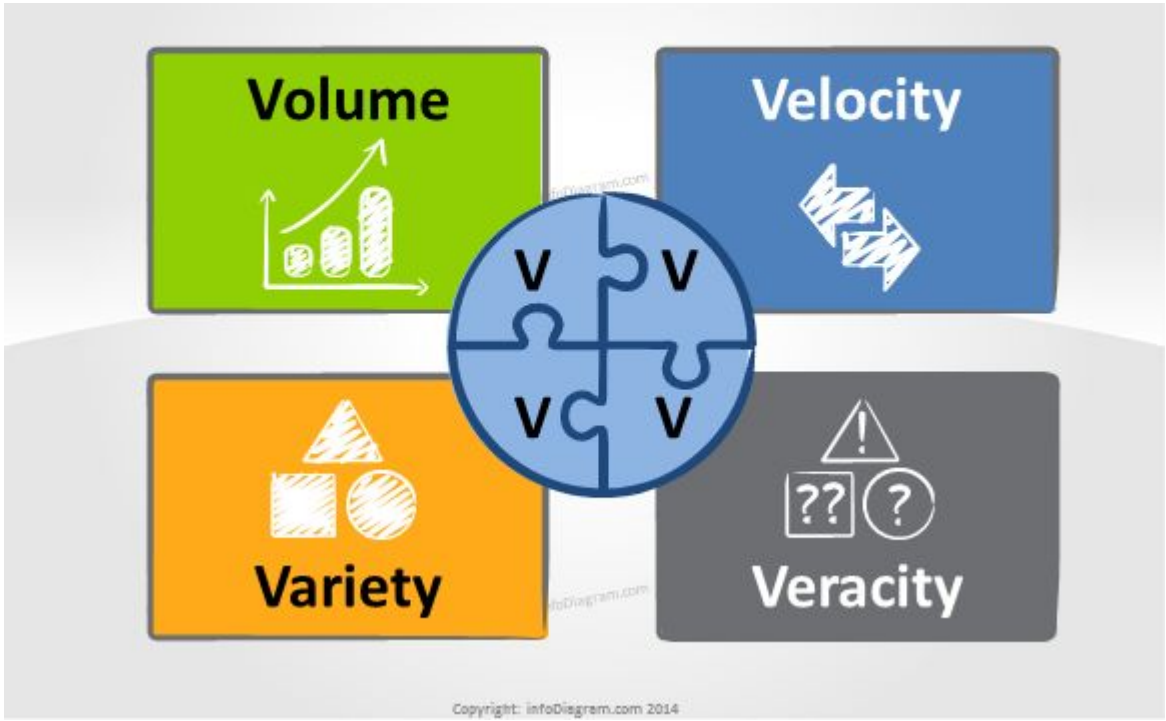
https://github.com/kaopanboonyuen/GISTDA2023

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


BUZZFEED
USERS VIEW
159,380
PIECES OF CONTENT
SNAPCHAT
USERS WATCH
6,944,444
VIDEOS
Netflix
SUBSCRIBERS STREAM
86,805 HOURS
OF VIDEO
GOOGLE
TRANSLATES
69,500,000
WORDS
Americans
USE
18,264,840
MEGABYTES
OF WIRELESS DATA
YOUTUBE
USERS SHARE
100 HOURS
OF NEW VIDEO
Instagram
USERS LIKE
2,430,555
POSTS
2016
every
MINUTE
of the
DAY
PRESENTED BY DOMO
FACEBOOK MESSENGER
USERS SHARE
216,302
PHOTOS
SIRI ANSWERS
99,206
REQUESTS
Amazon
MAKES
$222,283
IN SALES
Tinder
USERS SWIPE
972,222
TIMES
3,567,850
TEXT MESSAGES
ARE SENT
IN THE U.S.
Giphy
SERVES
569,217
GIFS
TWITTER
USERS SEND
9,678
EMOJI-FILLED TWEETS
Dropbox
USERS UPLOAD
833,333
NEW FILES
THE WEATHER CHANNEL
RECEIVES
13,888,889
FORECAST
REQUESTS
DOMO
DATA NEVER SLEEPS 4.0
How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like Youtube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the internet every minute? See for yourself below.

Big Data Definition



Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

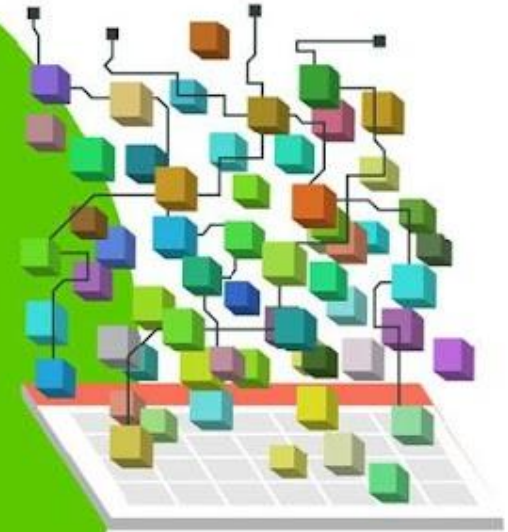
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

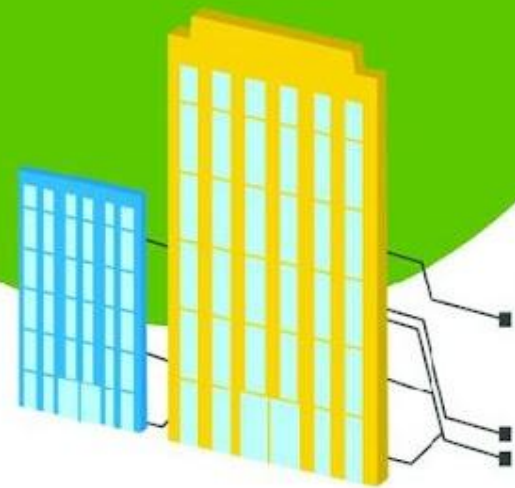
of data are created each day



6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION



Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

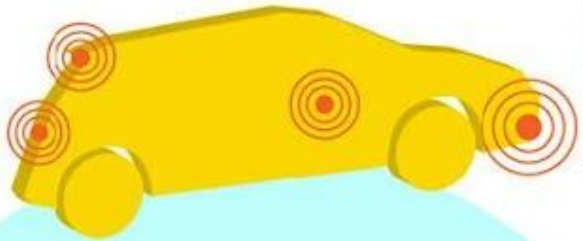
of data stored

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



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

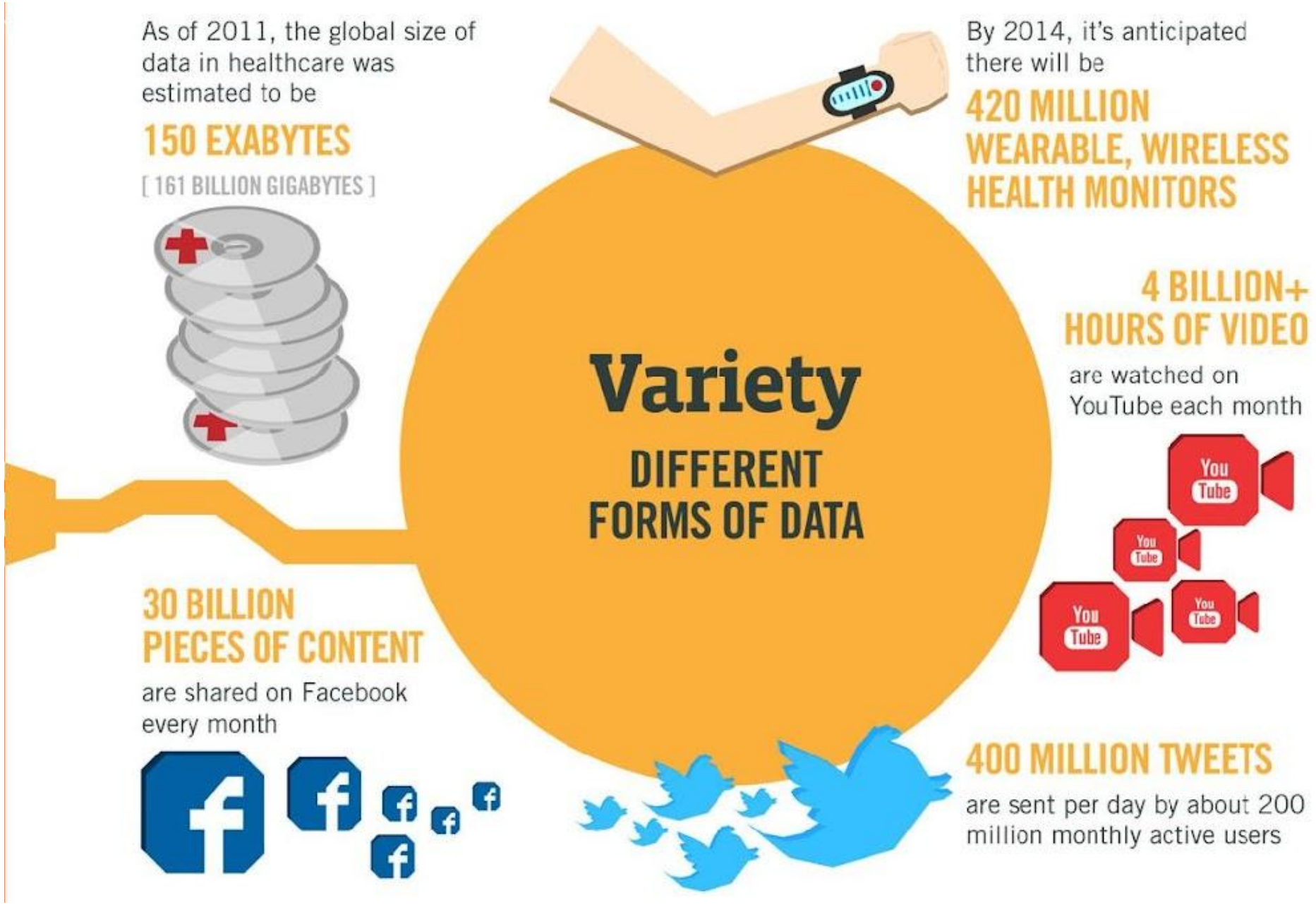
Velocity

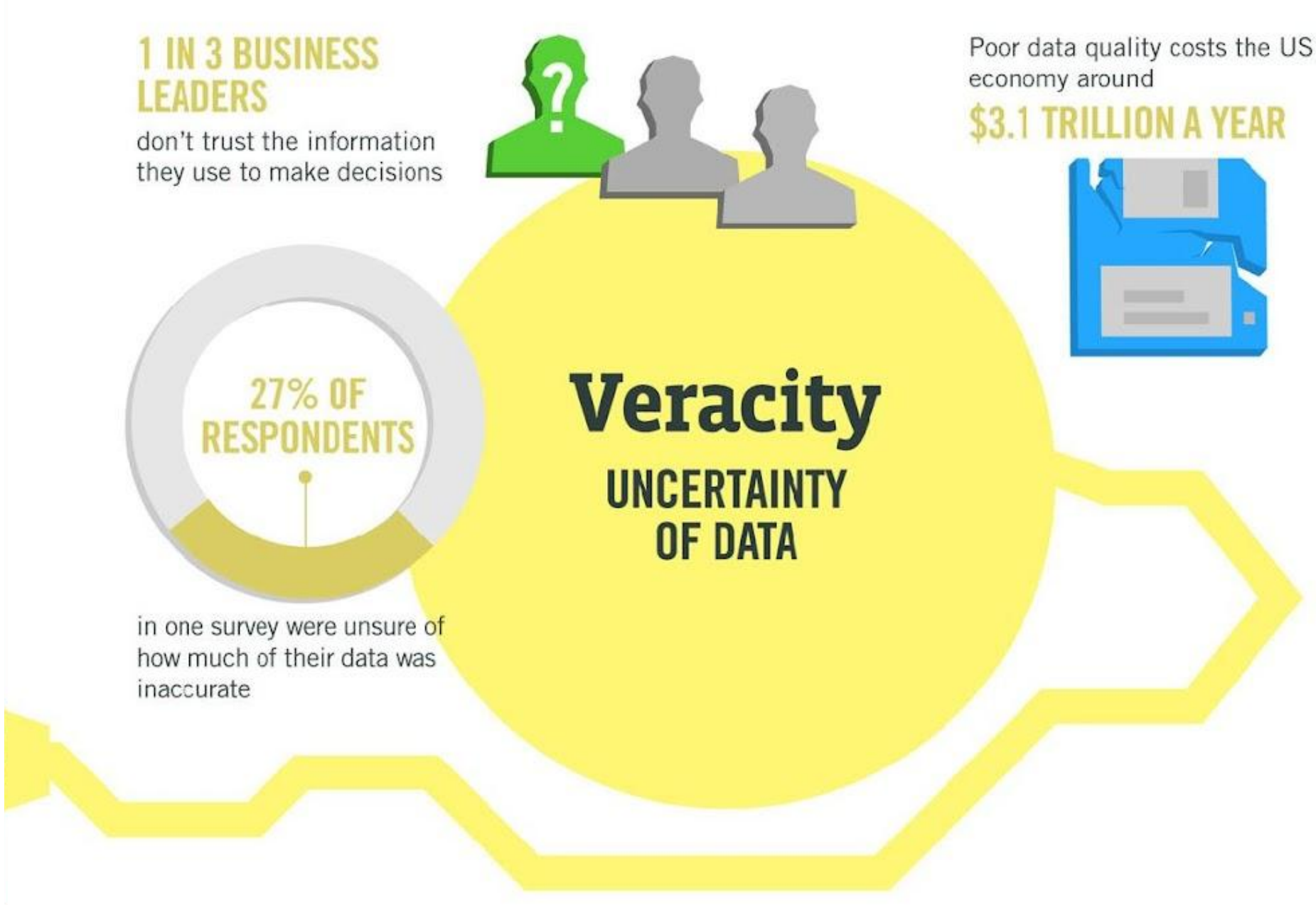
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

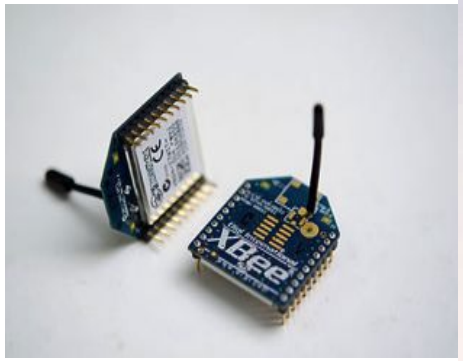
– almost 2.5 connections per person on earth

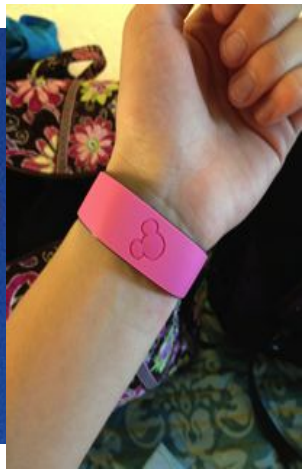

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

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)

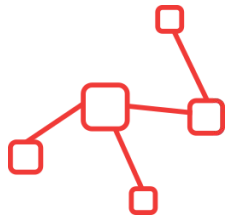
Experiment

Explore
Story-telling

Big Data

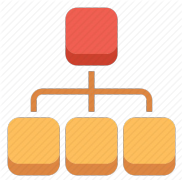
Why Big Data is hard?

- 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



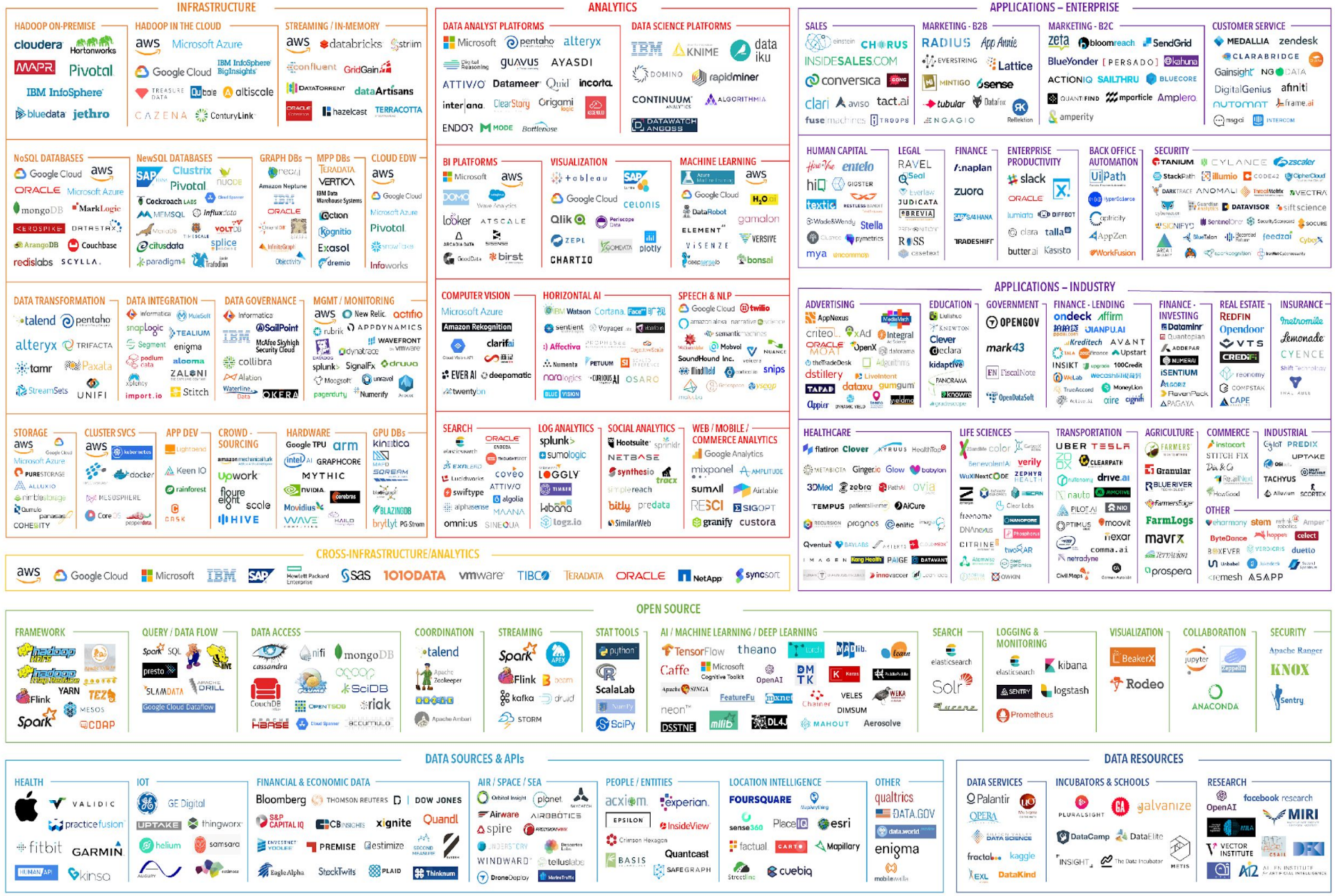
New Infrastructure

New Tools
NoSQL, Analytics, BI

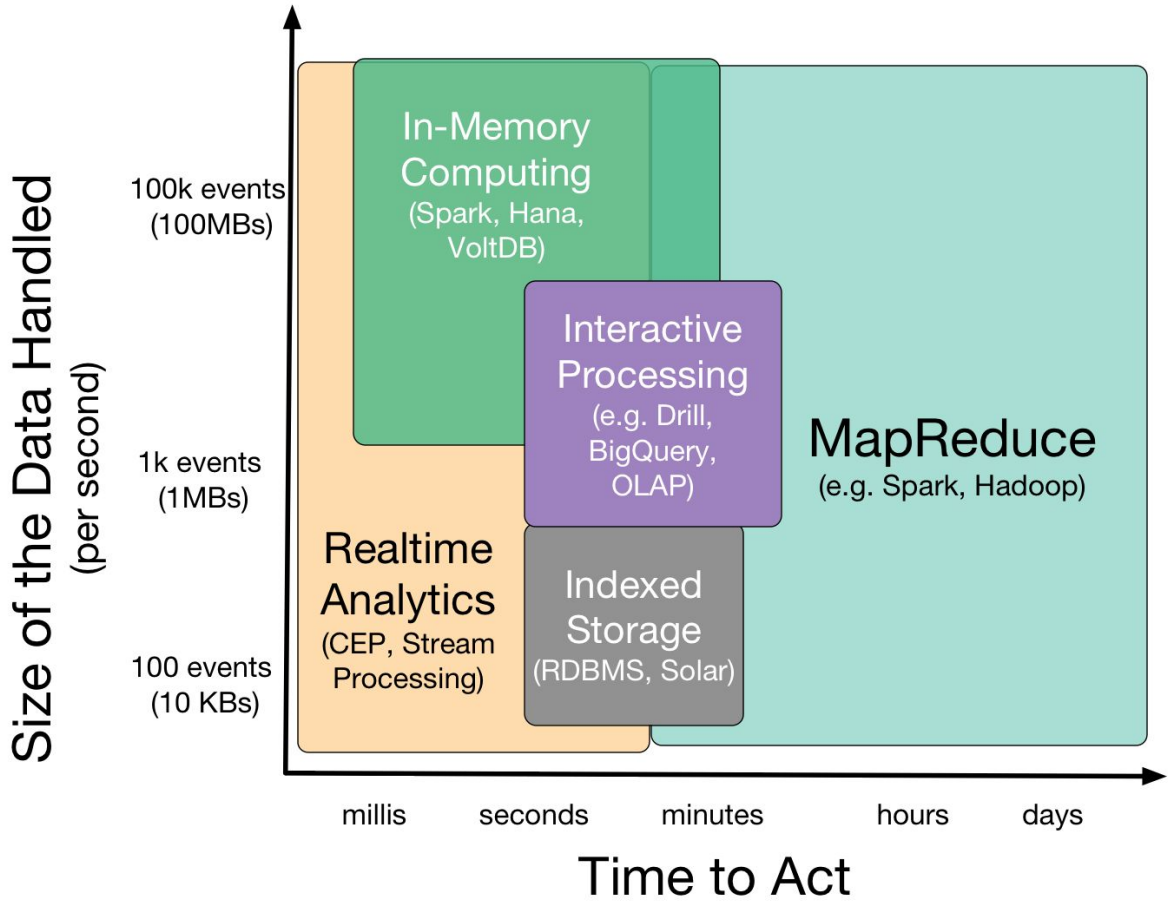


New Algorithms

BIG DATA & AI LANDSCAPE 2018



Big Data Processing Technologies Landscape



Big Data Analytics Process

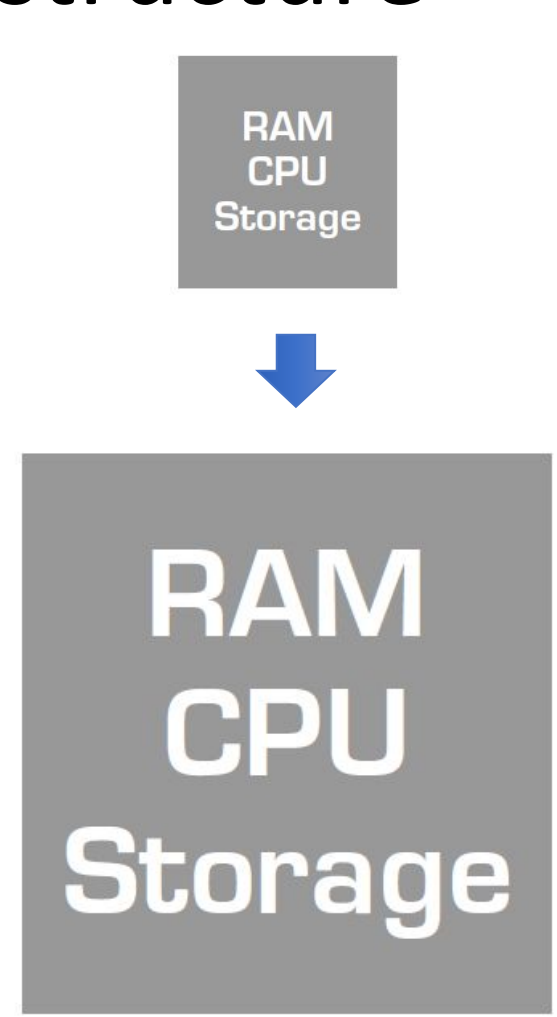
4 • Data Visualization

3 • Data Analytics

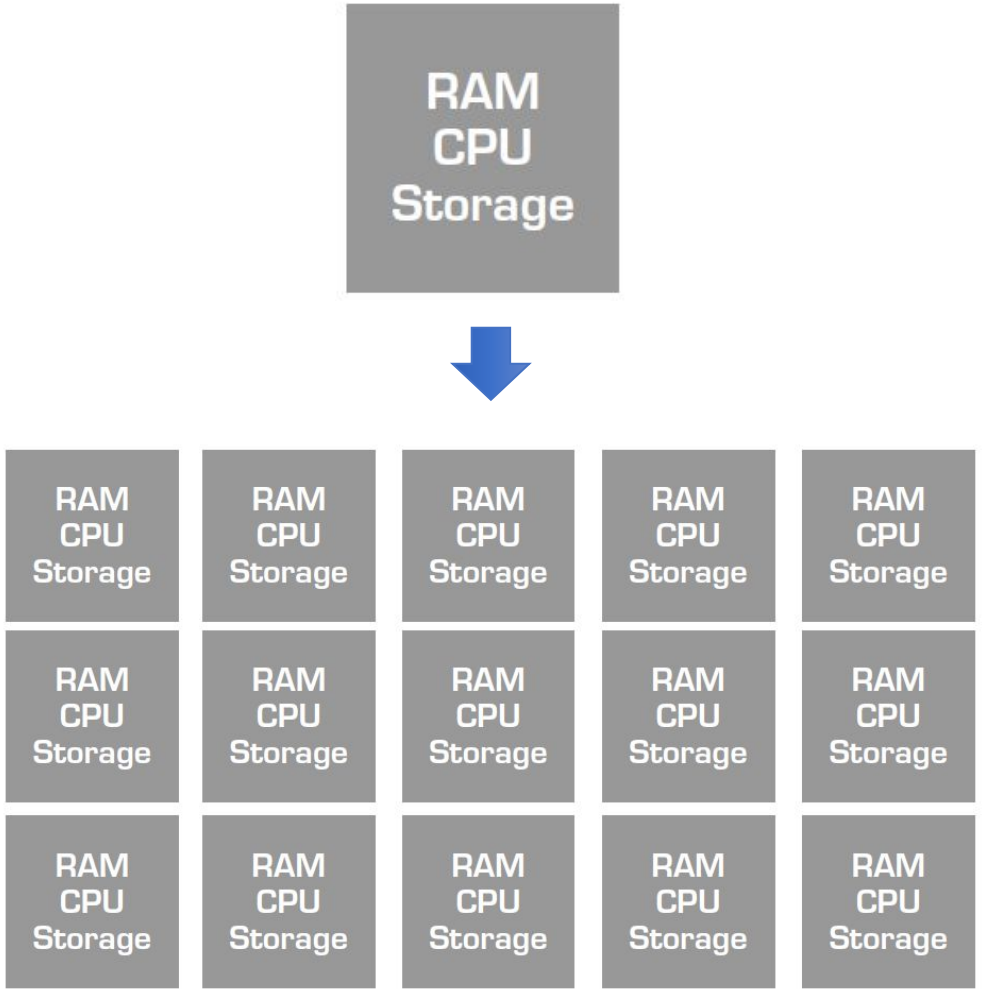
2 • Data Storage

1 • Infrastructure

Infrastructure



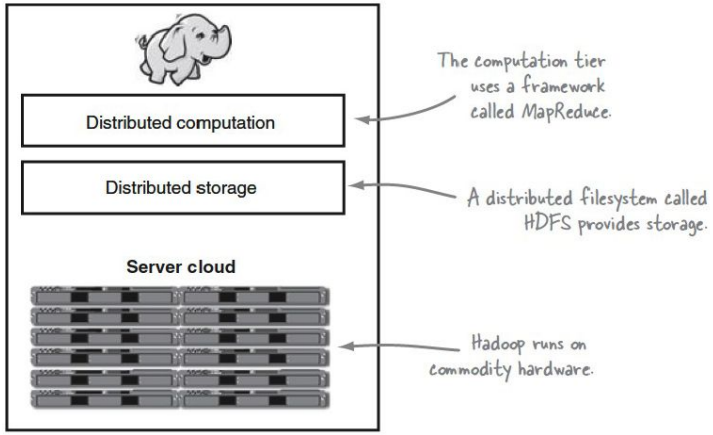
Vertical Scaling
(Scale-up)



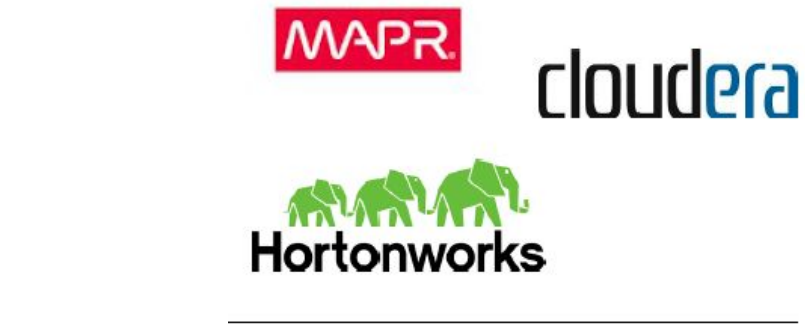
Horizontal Scaling
(Scale-out)

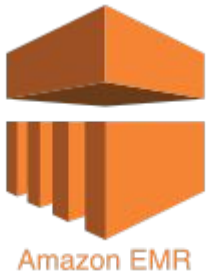
What is Hadoop?

- 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



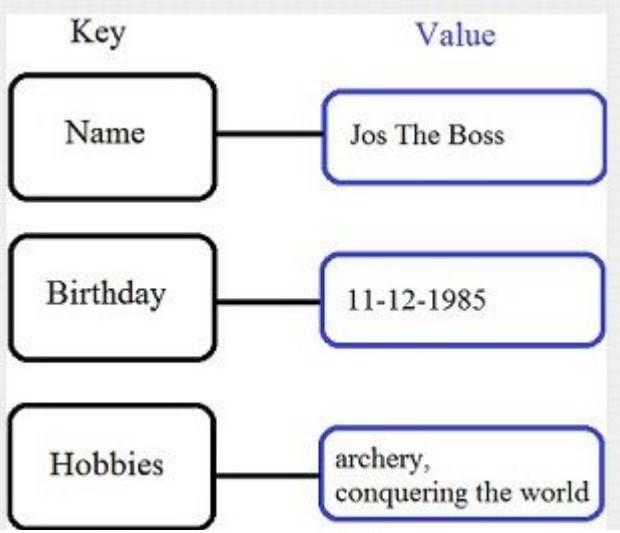


Google Cloud Platform

Cloud Dataproc
Google

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.

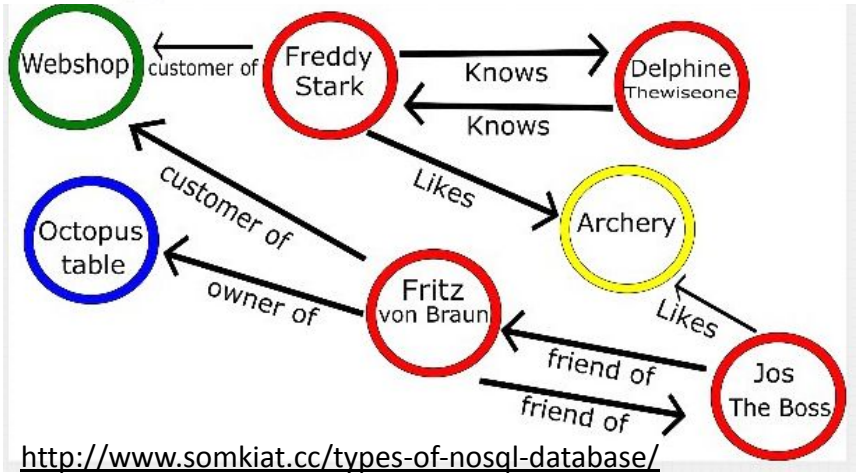
Name	ROWID
Jos The Boss	1
Fritz Schneider	2
Freddy Stark	3
Delphine Thewiseone	4

Birthday	ROWID
11-12-1985	1
27-1-1978	2
16-9-1986	4

Hobbies	ROWID
archery	1, 3
conquering the world	1
building things	2
surfing	2
swordplay	3
lollygagging	3

A column-oriented database stores each column separately

Graph Databases



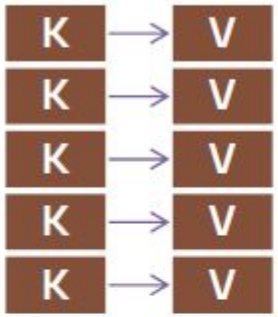
http://www.somkiat.cc/types-of-nosql-database/

Document Stores

Whereas relational databases chop up data, Document stores save documents as a single entity

NoSQL (cont.)

Key-Value Stores

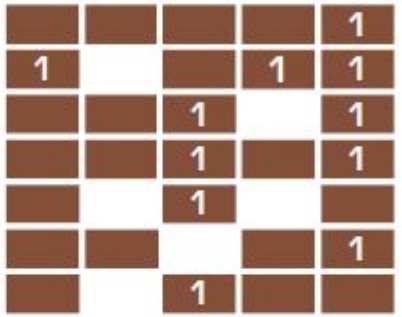


amazon web services
Amazon DynamoDB

redis

riak

Column Stores



Google Bigtable

APACHE HBASE

Cassandra

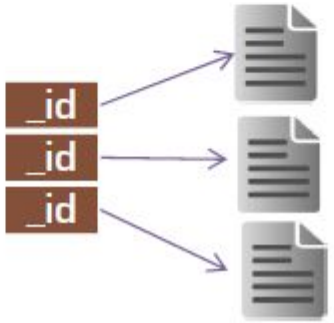
Graph Databases

Neo4j
The World's Leading Graph Database

HYPERGRAPHDB

*dex

Document Stores

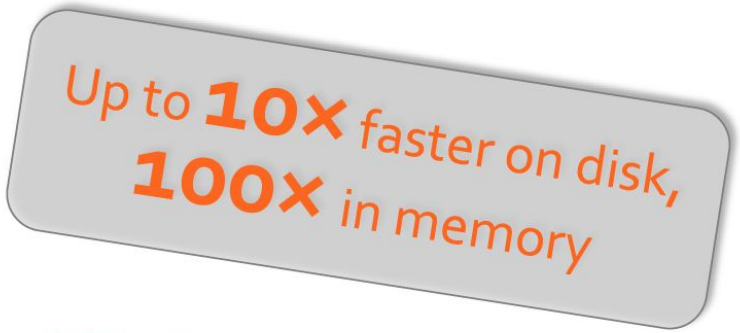


mongoDB

CouchDB
relax

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



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.

http://www.kanouivirach.com/2014/04/elasticsearch-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3/

Big Data Ecosystem