Cloud-based Data Analytics

Lecture 1

Cloud-based Data Analytics – Fall 2022

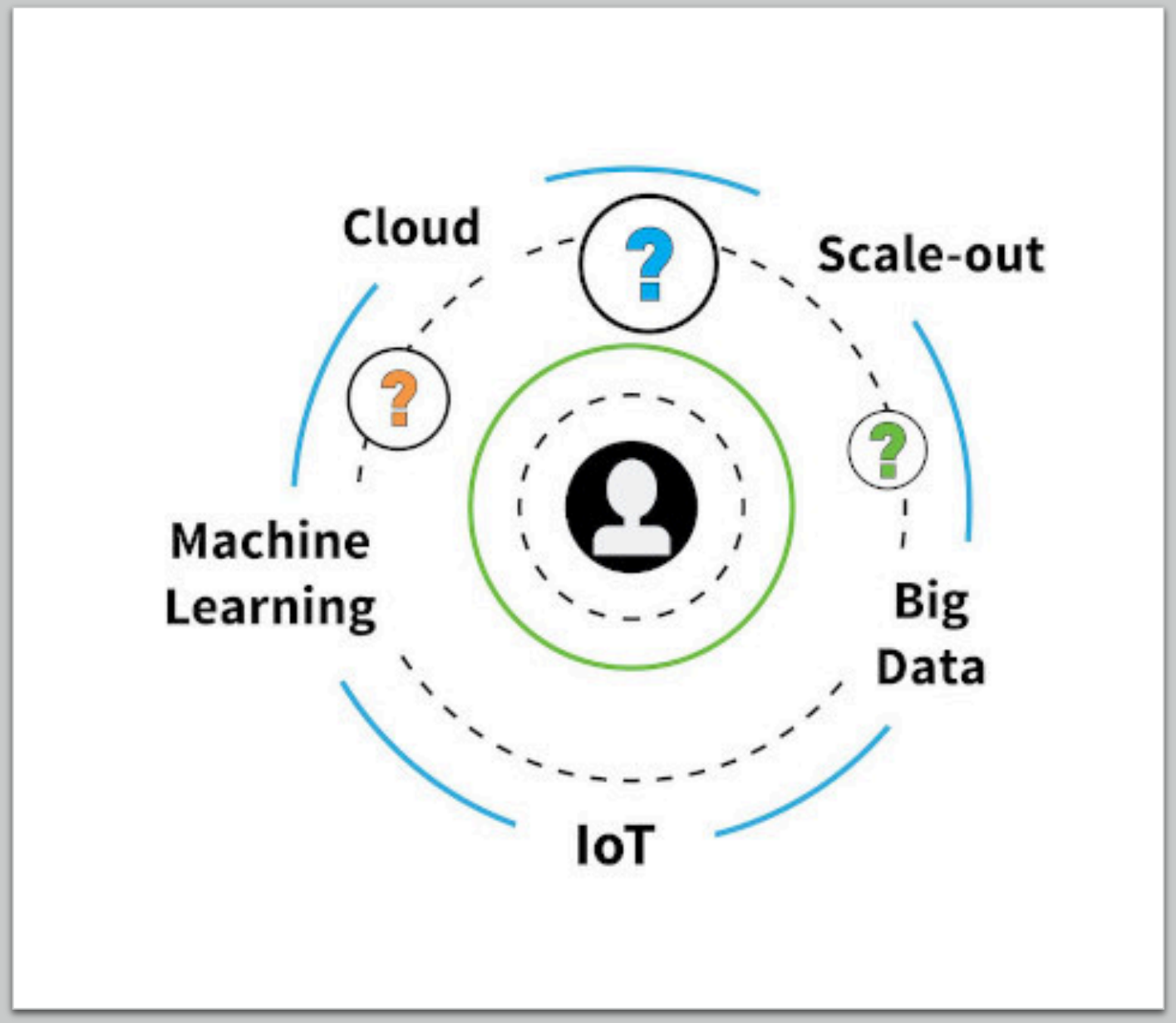
- Instructor:
  - Sneha : s.sneha@utoronto.ca

- TAs:
  - Mustafa Ammous: mustafa.ammous@mail.utoronto.ca
  - Ziyue Dong: zy.dong@mail.utoronto.ca
  - Faraz Khoshbakhtian: faraz.khoshbakhtian@mail.utoronto.ca

- Office Hours: Online – on zoom

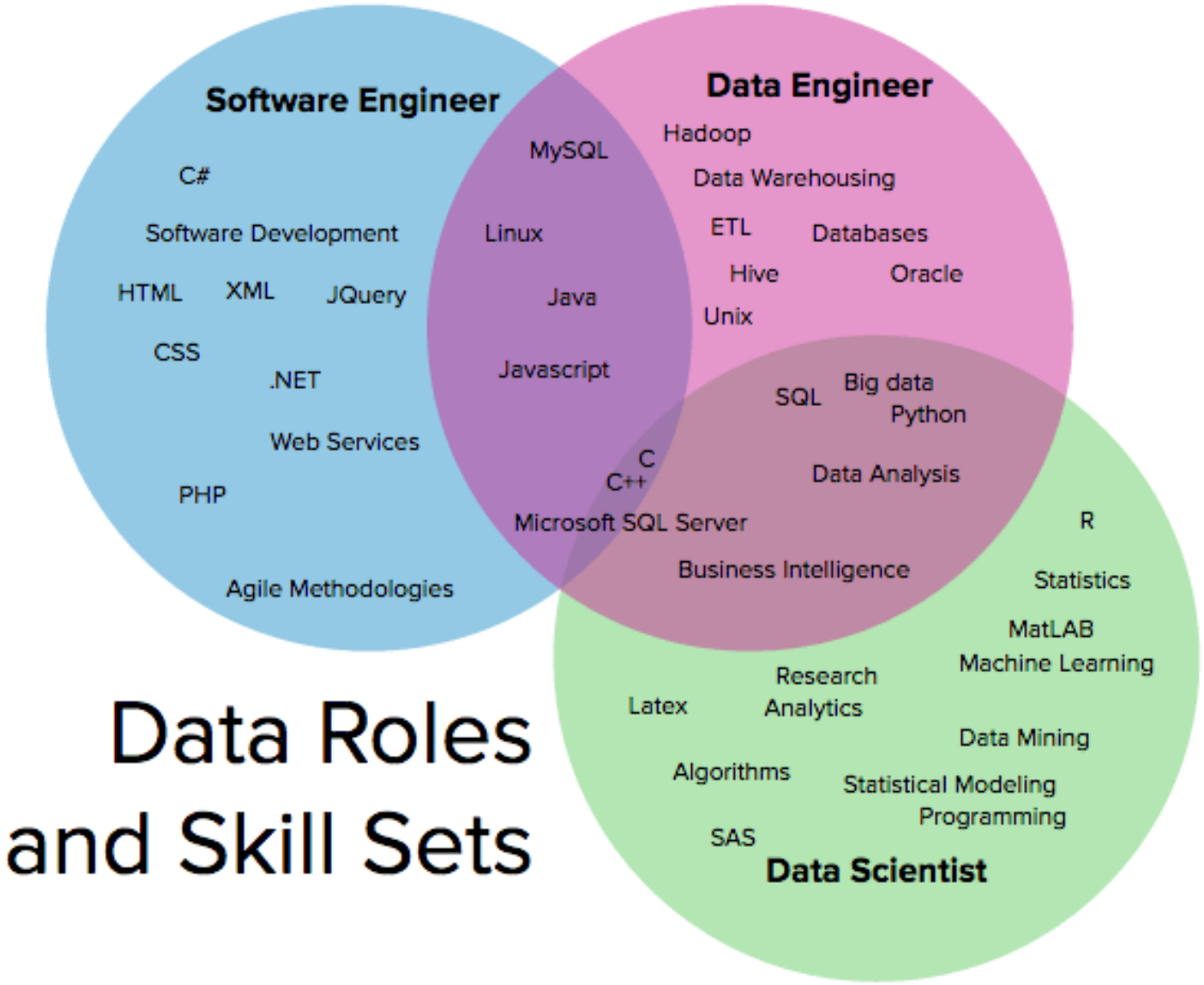
Cloud-based Data Analytics 1628 - Course Outline

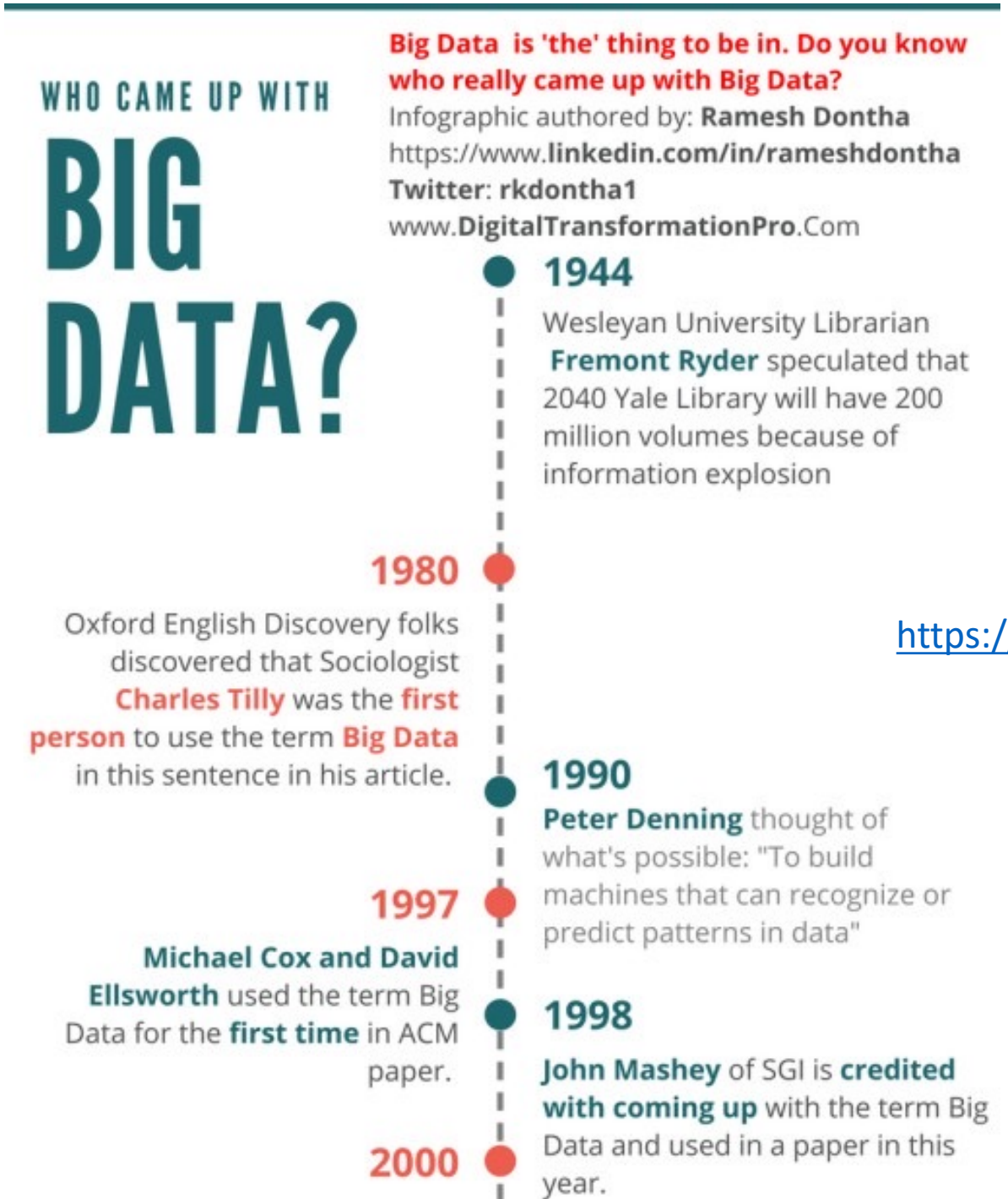
Grading: Assignment/Exam	Weight (%)	Due Date / Time
Assignment 1	10	Jan 30 @ 24:00
Assignment 2	10	Feb 13 @ 24:00
Midterm	25	Feb 25
Assignment 3	10	Mar 6 @ 24:00
Assignment 4	10	Mar 20 @ 24:00
Assignment 5	10	Apr 03 @ 24:00
Final Exam	25	Apr 08



Cloud-based Data Analytics?

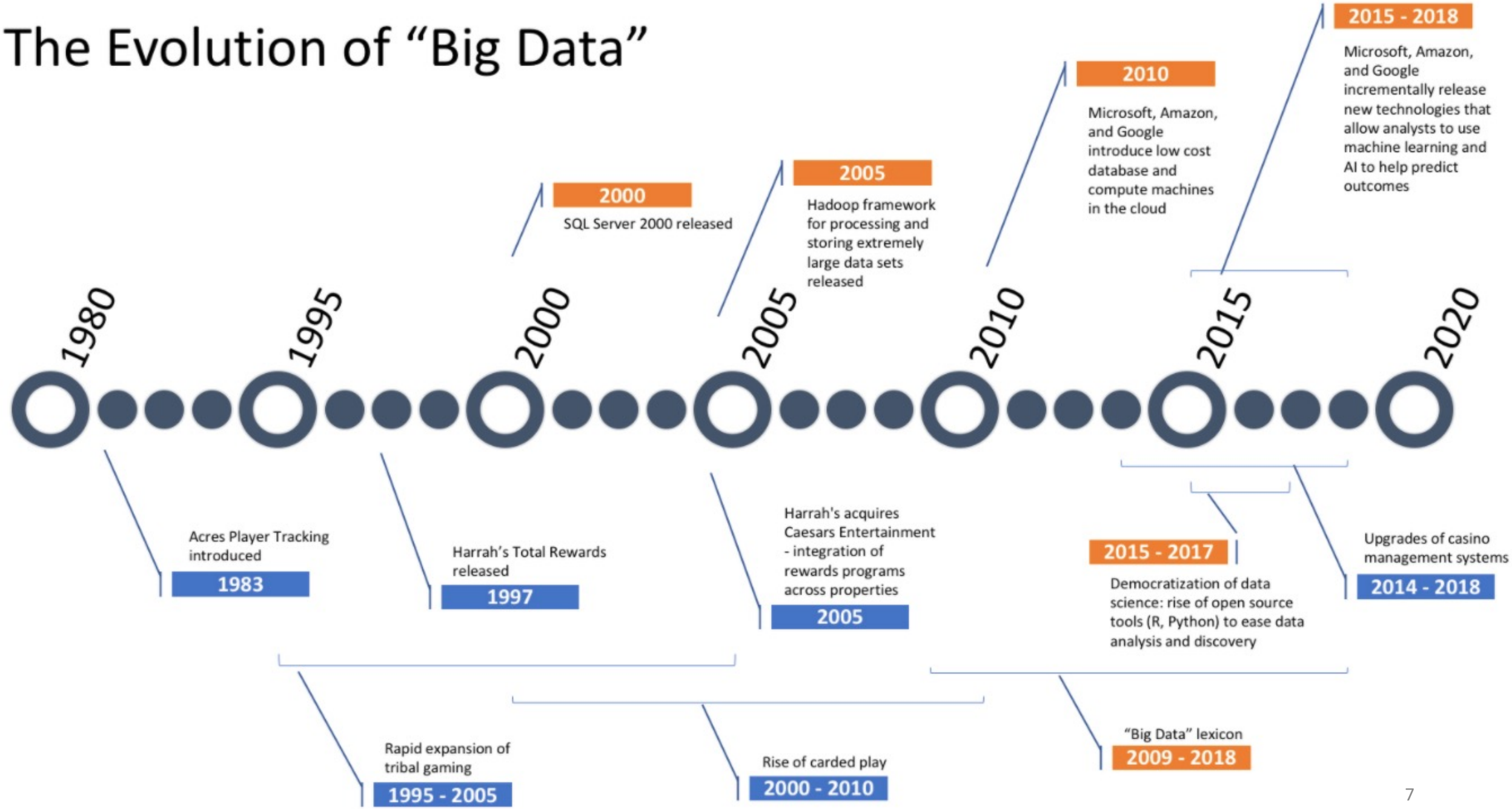
This course will focus on some Data Scientist skills and some Data Engineer skills

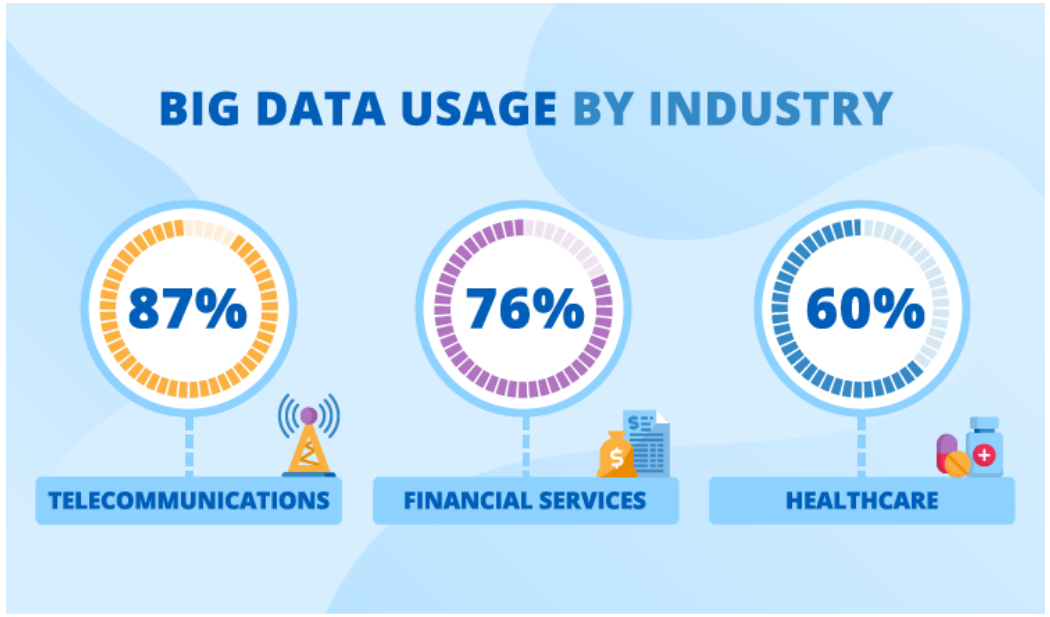




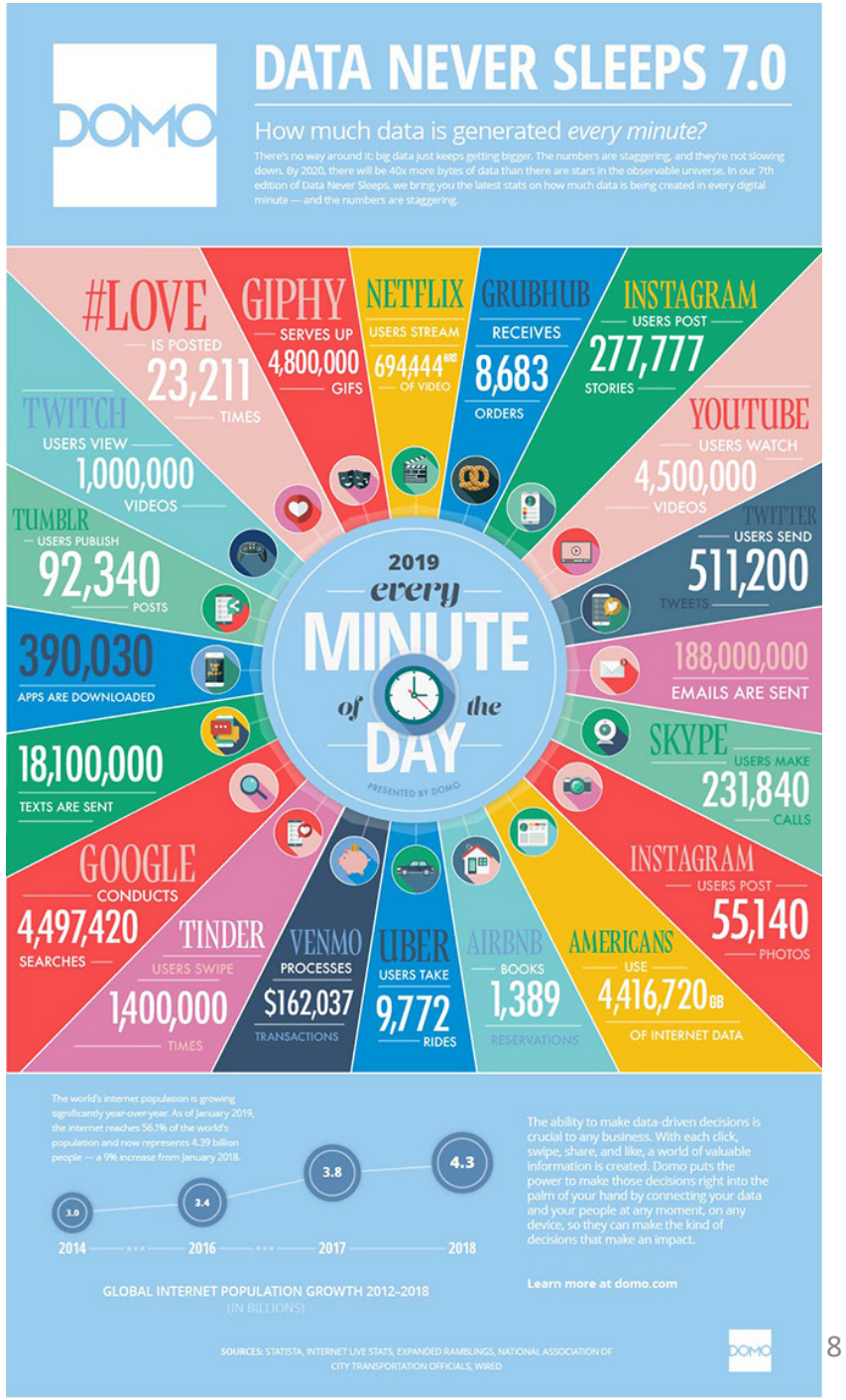
https://digitaltransformationpro.com/

The Evolution of "Big Data"



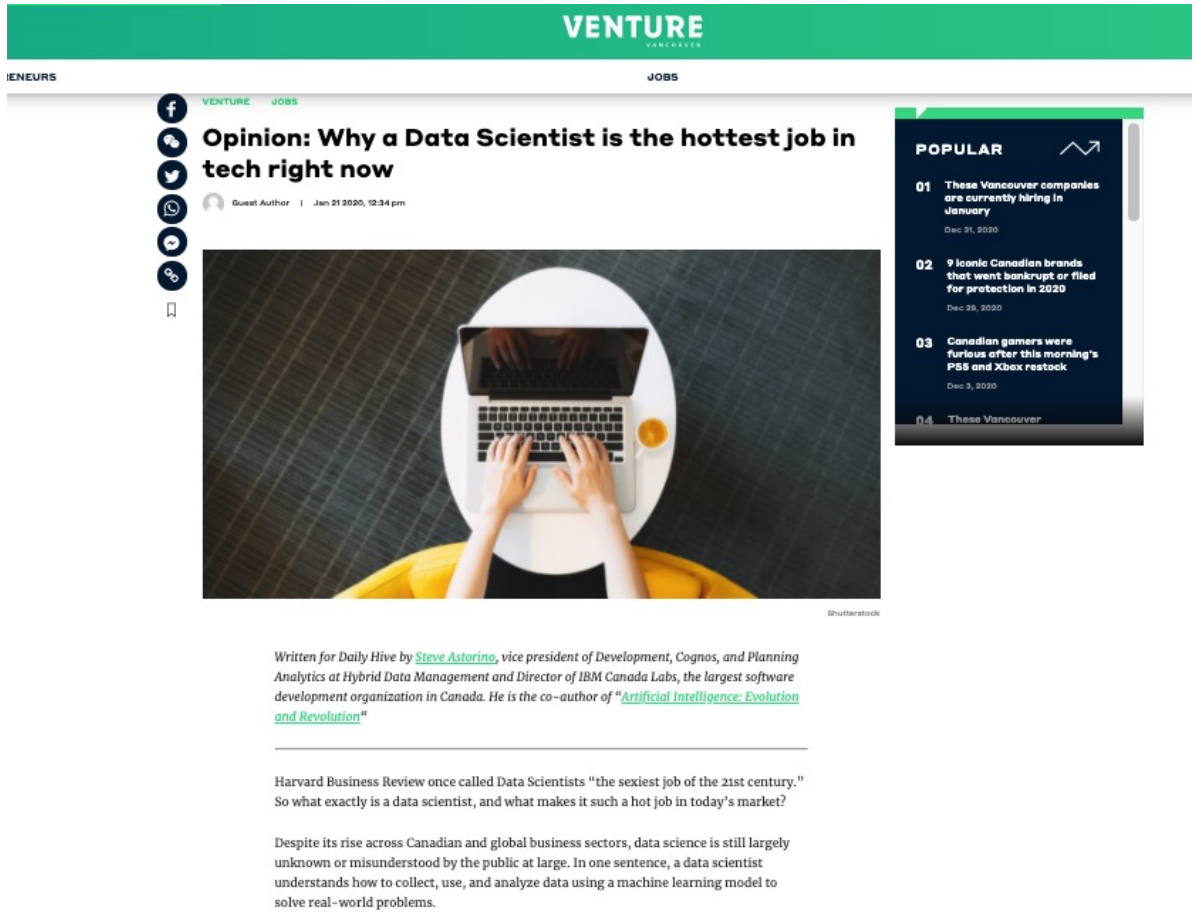

BIG DATA USAGE BY INDUSTRY
87%
TELECOMMUNICATIONS
76%
FINANCIAL SERVICES
60%
HEALTHCARE


Growth in healthcare data
1 exabyte = 1 billion gigabytes
2013
153
EXABYTES
2020
2,314
EXABYTES
Source: Stanford Medicine 2017, IDC 2014
To put that into perspective, data centers globally will only have enough room for an estimated
985 exabytes by 2020—
meaning that almost two and a half times this capacity would be required to house all the healthcare data.
Source: Cisco Global Cloud Index 2016
DATA STORAGE VS MEDICAL DATA (2020)
STORAGE CAPACITY
985
EXABYTES
MEDICAL DATA GENERATED
2,314
EXABYTES


DOMO
DATA NEVER SLEEPS 7.0
How much data is generated every minute?
There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute — and the numbers are staggering.
2019
every MINUTE of the DAY
PRESENTED BY DOMO
#LOVE IS POSTED 23,211 TIMES
GIPHY SERVES UP 4,800,000 GIFS
NETFLIX USERS STREAM 694444 HRS OF VIDEO
GRUBHUB RECEIVES 8,683 ORDERS
INSTAGRAM USERS POST 277,777 STORIES
TWITCH USERS VIEW 1,000,000 VIDEOS
YOUTUBE USERS WATCH 4,500,000 VIDEOS
TUMBLR USERS PUBLISH 92,340 POSTS
TWITTER USERS SEND 511,200 TWEETS
390,030 APPS ARE DOWNLOADED
188,000,000 EMAILS ARE SENT
18,100,000 TEXTS ARE SENT
SKYPE USERS MAKE 231,840 CALLS
GOOGLE CONDUCTS 4,497,420 SEARCHES
INSTAGRAM USERS POST 55,140 PHOTOS
TINDER USERS SWIPE 1,400,000 TIMES
VENMO PROCESSES $162,037 TRANSACTIONS
UBER USERS TAKE 9,772 RIDES
AIRBNB BOOKS 1,389 RESERVATIONS
AMERICANS USE 4,416,720 GB OF INTERNET DATA
The world's internet population is growing significantly year-over-year. As of January 2019, the internet reaches 56.1% of the world's population and now represents 4.39 billion people — a 9% increase from January 2018.
3.0
3.4
3.8
4.3
2014
2016
2017
2018
GLOBAL INTERNET POPULATION GROWTH 2012-2018
(IN BILLIONS)
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.
Learn more at domo.com
SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED
DOMO

Cloud and Big Data Science Skills are in Demand

- Best Big Data Careers to Explore in 2020
- What's next for the data science and analytics job market?
- 10 in-demand roles for 2020 driven by data science
- Top Emerging Job In 2020 According To LinkedIn
- Top 5 Most In-Demand Big Data Jobs To Earn Big Bucks
- This Is America's Hottest Job



VENTURE

JOBS

Opinion: Why a Data Scientist is the hottest job in tech right now

Guest Author | Jan 21 2020, 12:34 pm

POPULAR

01 These Vancouver companies are currently hiring in January

Dec 31, 2020

02 9 iconic Canadian brands that went bankrupt or filed for protection in 2020

Dec 29, 2020

03 Canadian gamers were furious after this morning's PS5 and Xbox restock

Dec 3, 2020

04 These Vancouver

Shutterstock

Written for Daily Hive by Steve Astorino, vice president of Development, Cognos, and Planning Analytics at Hybrid Data Management and Director of IBM Canada Labs, the largest software development organization in Canada. He is the co-author of "Artificial Intelligence: Evolution and Revolution"

Harvard Business Review once called Data Scientists "the sexiest job of the 21st century." So what exactly is a data scientist, and what makes it such a hot job in today's market?

Despite its rise across Canadian and global business sectors, data science is still largely unknown or misunderstood by the public at large. In one sentence, a data scientist understands how to collect, use, and analyze data using a machine learning model to solve real-world problems.

Why a Data Scientist is the hottest job in tech right now

Introduction of this Lecture

What is Big Data?

What are the main characteristics of the Big Data?

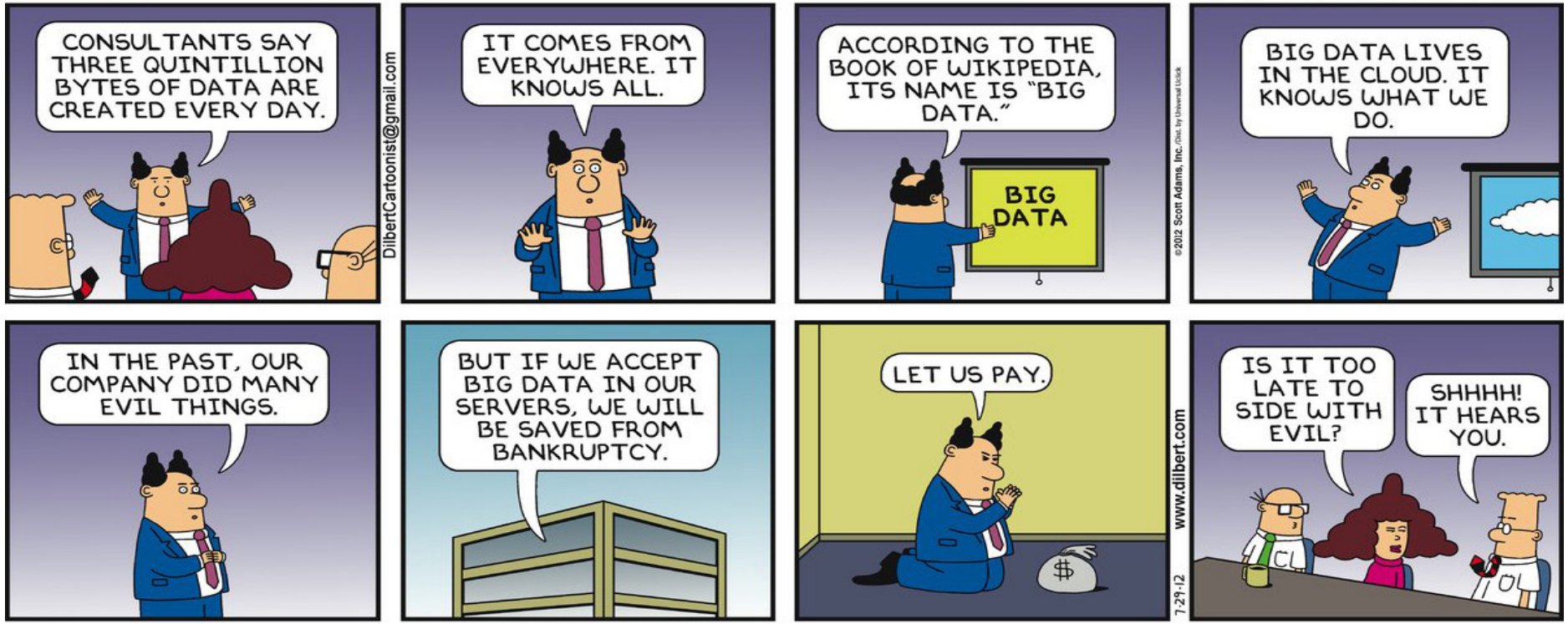
How to deal with Big Data ?
Big Data Enabling Technologies

Hadoop Stack for Big Data

Big Data Landscape

Hadoop Installation

What's the Big Data?



- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Big Data Definition

- ***"Big data"*** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
- Systems/Enterprises generate huge amount of data from Terabytes to and even Petabytes of information.

Some Facts and Figures

- Walmart handles 1 million customer transactions/hour
- Facebook handles 40 billion photos from its user base
- Facebook inserts 500 terabytes of data every day.
- Facebook stores, access and analyses 30+ petabytes of user generated data.
- YouTube users upload 48 hours of new video every minute of the day.
- A flight generates 240 terabytes of flight data in 6-8 hours of flight.
- More than 5 million people are calling, texting, tweeting and browsing on cell phones worldwide.
- Decoding the human genome originally took 10 years to process, now it can be achieved in one week.
- The largest AT&T database boasts titles including the largest volume of data in one of the unique database (312 terabytes) and the second largest number of rows in a unique database (1.9 trillion), which comprises AT&T's extensive calling records.
- 80-90% of the data we generate today is unstructured.
- By the end of 2020, 44 zettabytes will make up the entire digital universe.
- 463 exabytes of data will be generated each day by humans as of 2025.

Bytes – With an Example

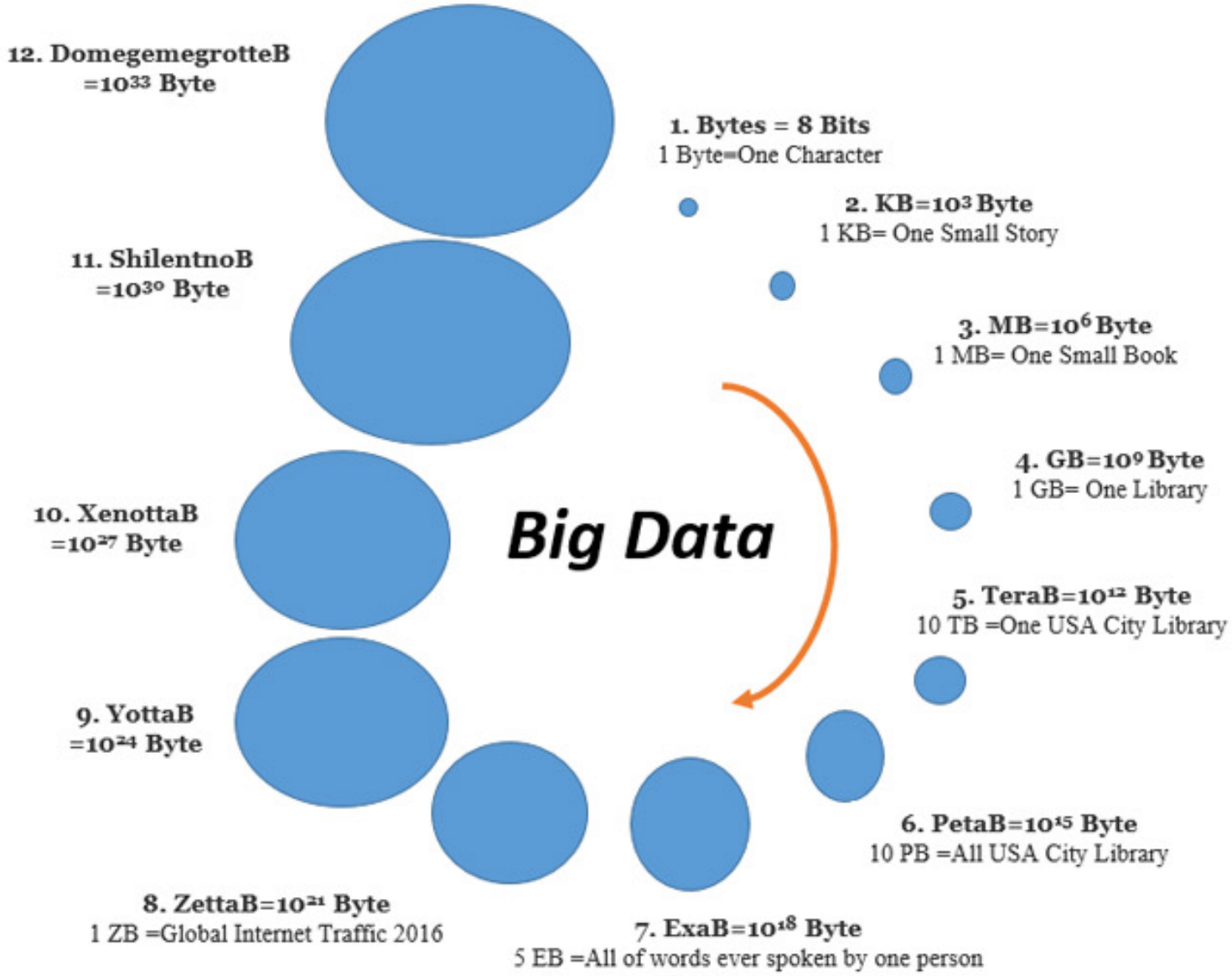
- Byte - One grain of rice
- Kilobyte (10^3) - One cup of rice
- Megabyte (10^6) - 8 bags of rice
- Gigabyte (10^9) - 3 Semi trucks of rice
- Terabyte (10^{12}) - 2 container ships of rice
- Petabyte (10^{15}) - Blankets ½ of Toronto
- Exabyte (10^{18}) - Blankets over west coast or ¼ of Canada
- Zettabyte (10^{21}) - Fills Pacific ocean
- Yottabyte (10^{24}) - An earth sized rice bowl

Desktop

Internet

Big Data

Future

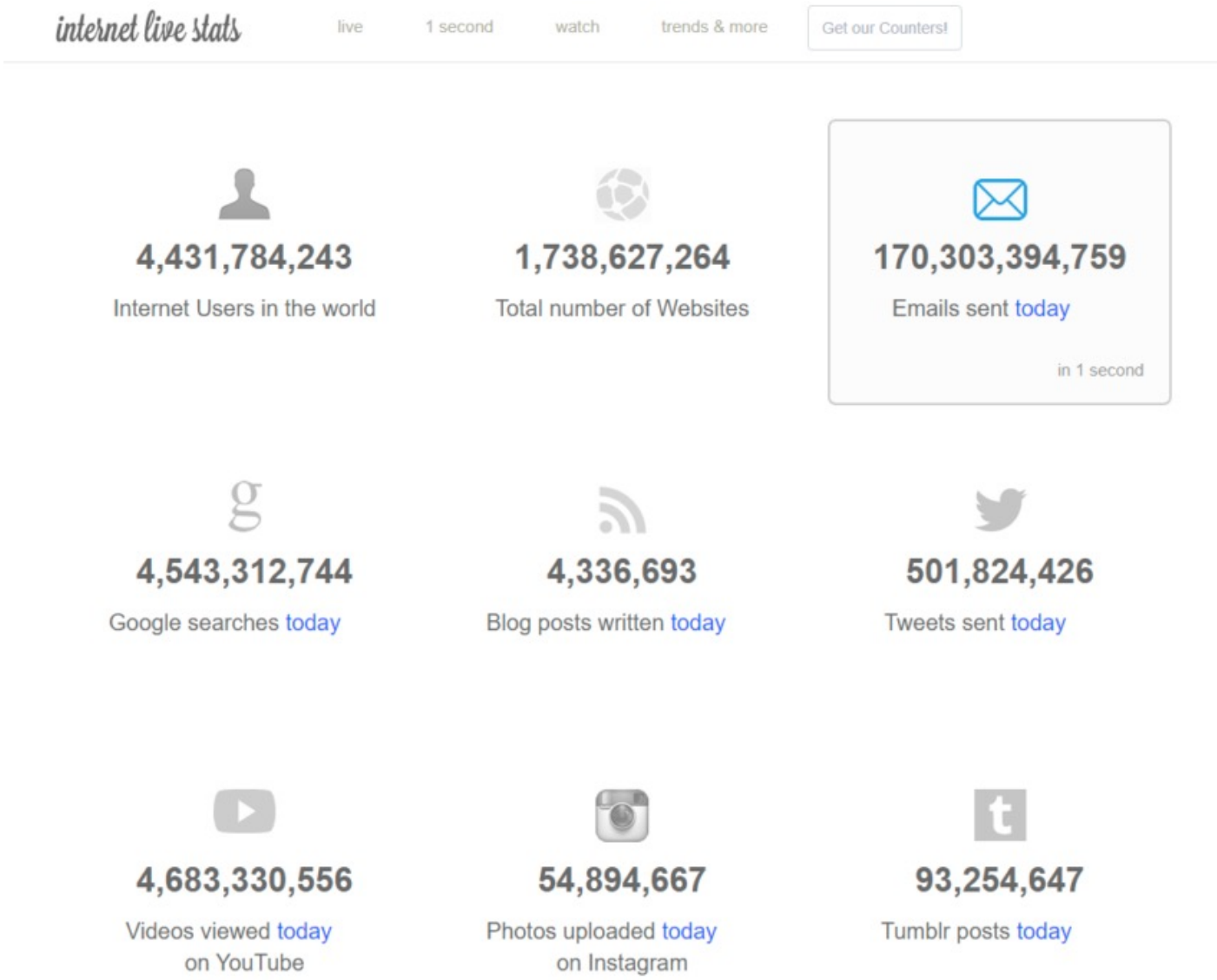

12. DomegemegrotteB
=10³³ Byte
11. ShilentnoB
=10³⁰ Byte
10. XenottaB
=10²⁷ Byte
9. YottaB
=10²⁴ Byte
8. ZettaB=10²¹ Byte
1 ZB =Global Internet Traffic 2016
7. ExaB=10¹⁸ Byte
5 EB =All of words ever spoken by one person
6. PetaB=10¹⁵ Byte
10 PB =All USA City Library
5. TeraB=10¹² Byte
10 TB =One USA City Library
4. GB=10⁹ Byte
1 GB= One Library
3. MB=10⁶ Byte
1 MB= One Small Book
2. KB=10³ Byte
1 KB= One Small Story
1. Bytes = 8 Bits
1 Byte=One Character
Big Data

Question

What’s making so much of data?

What's making so much of data?

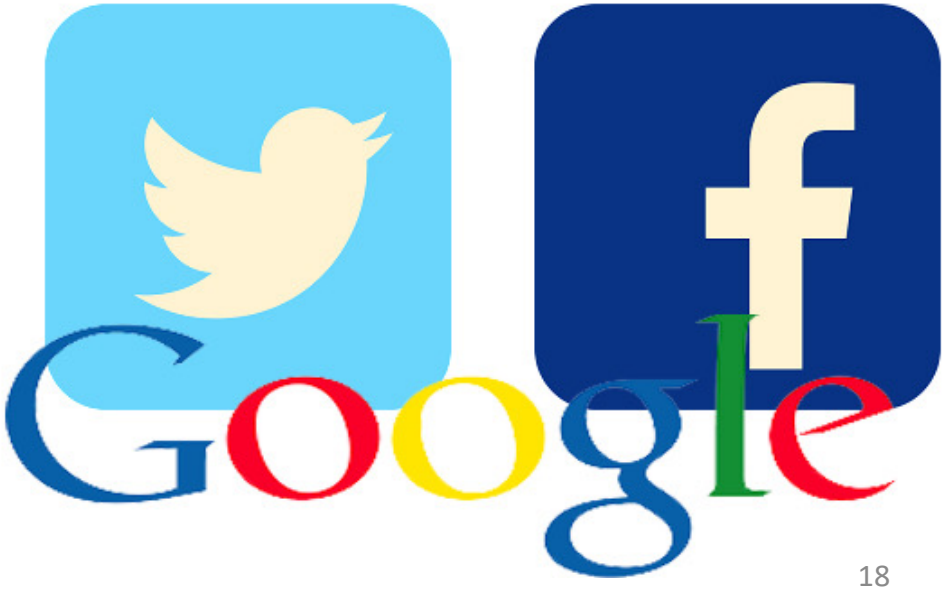
- Sources: People, machine, organization: ubiquitous computing
- More people carrying data-generating devices (cell phones with Facebook, Camera, GPS etc.)
- Data on the Internet:



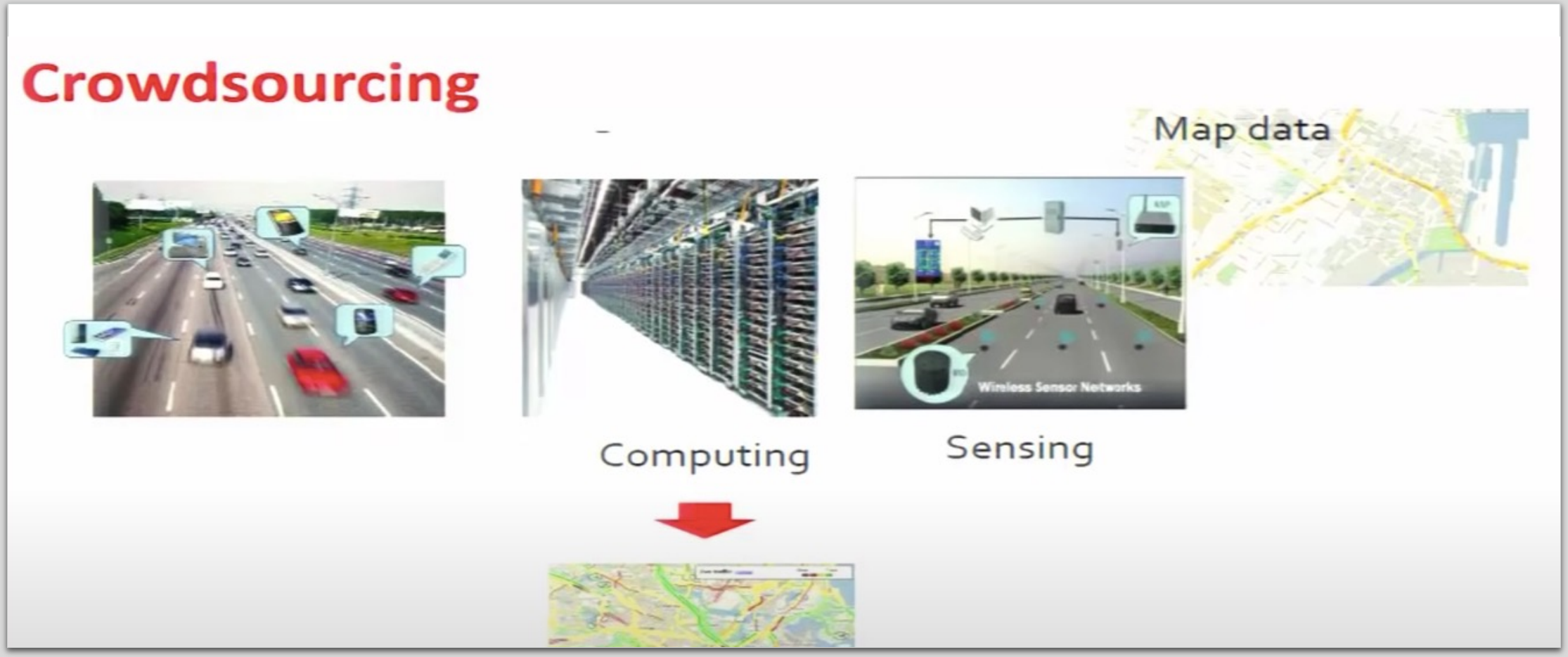
Sources of Data Generation

12+ TBs of tweet data every day

25+ TBs of log data every day

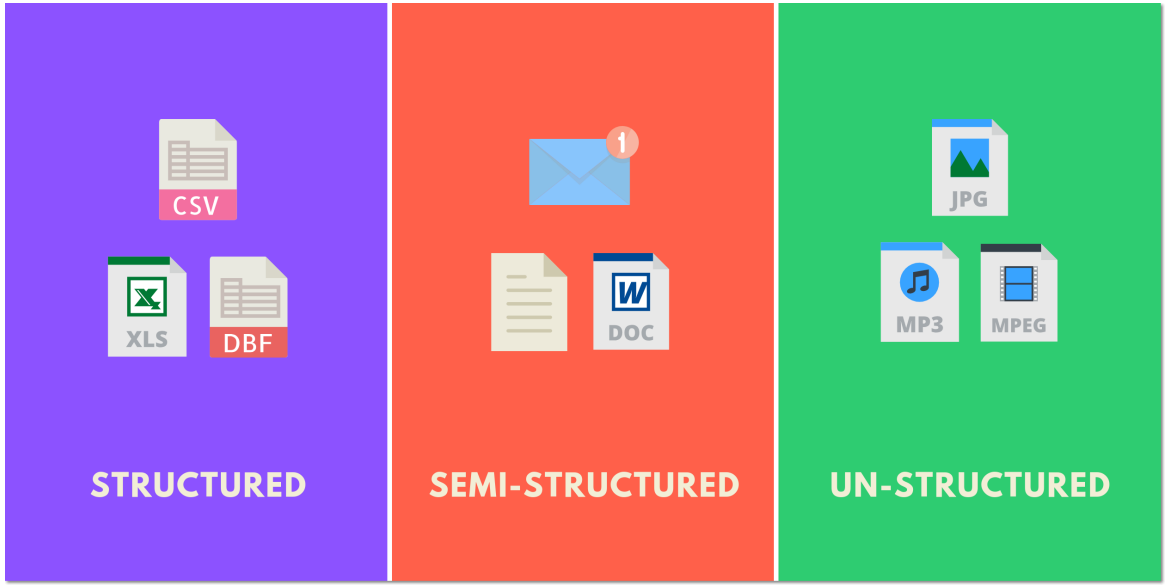


An Example of Big Data at Work



Types of Big Data

- Big Data could be of three types:
  - ***Structured:*** Organised data format with a fixed schema. Ex: RDBMS
  - ***Semi-Structured/Multi-Structured:*** Partially organized data which does not have a fixed format. Ex: XML, JSON
  - ***Unstructured:*** Unorganized data with an unknown schema. Ex: Audio, video files, etc.



Where is the problem?

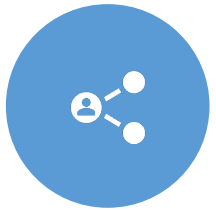
- Traditional RDBMS queries isn’t sufficient to get useful information out of huge volume of data
- To search it with traditional tools to find out if a particular topic was trending would take so long that result will be meaningless by the time it was completed.
- Big data provides a solution to store this data in novel ways in order to make it more accessible and provide methods of performing analysis on it.

Challenges

Capturing

Storing

Searching



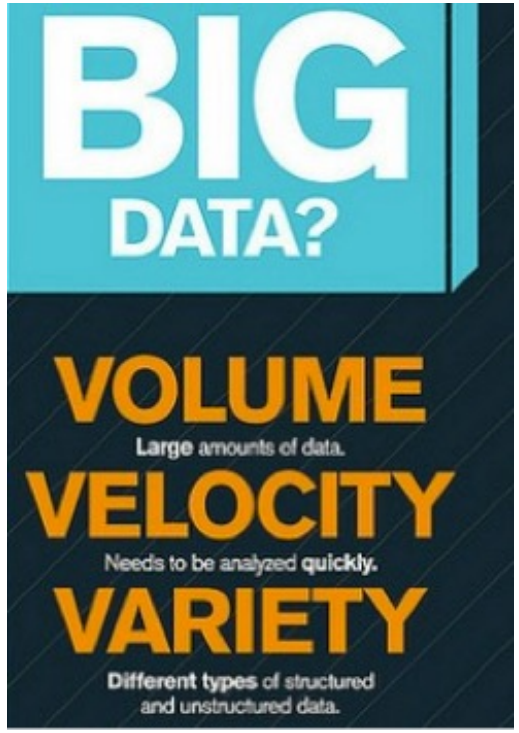
Sharing

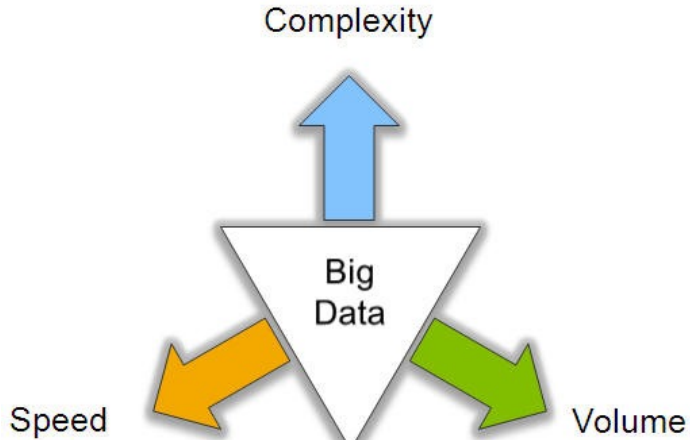
Analyzing

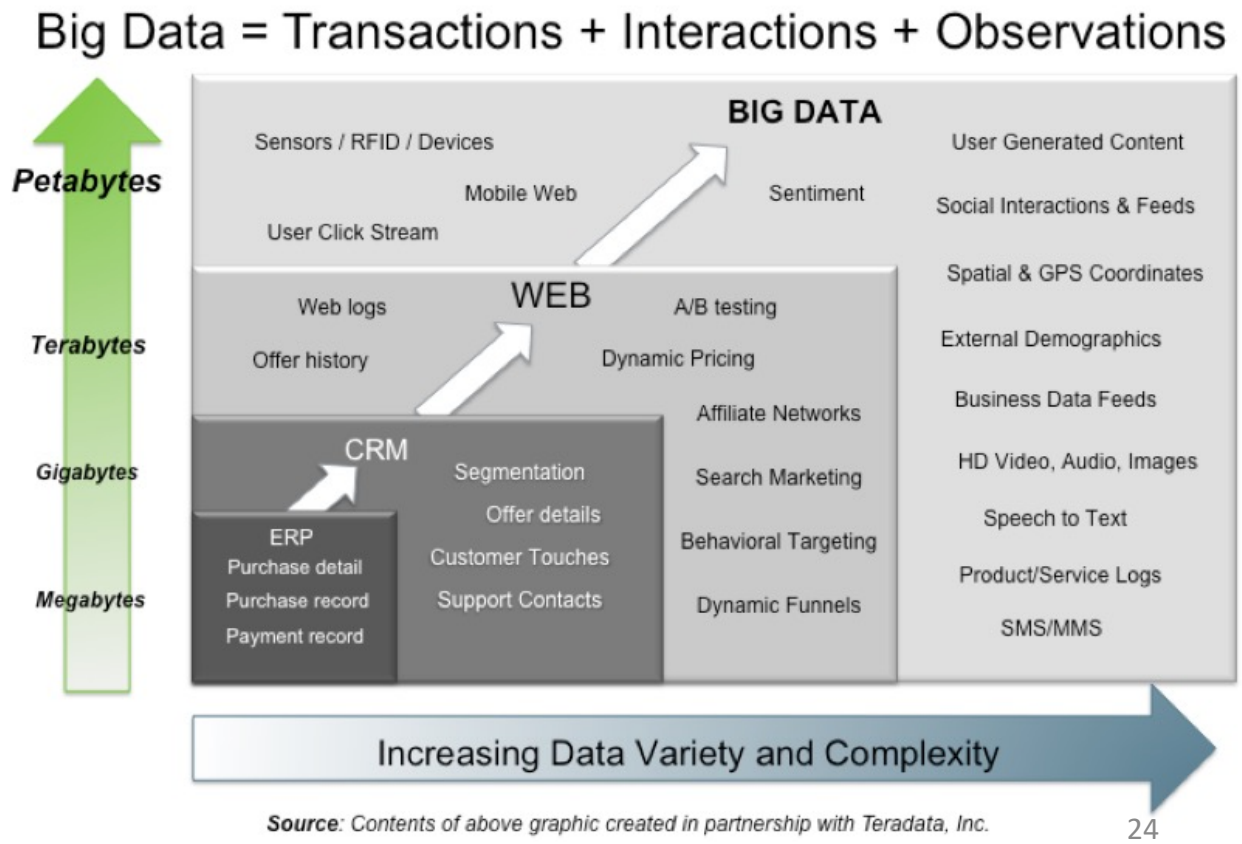
Visualization

What are the main characteristics of Big Data ??

Big Data: 3V's







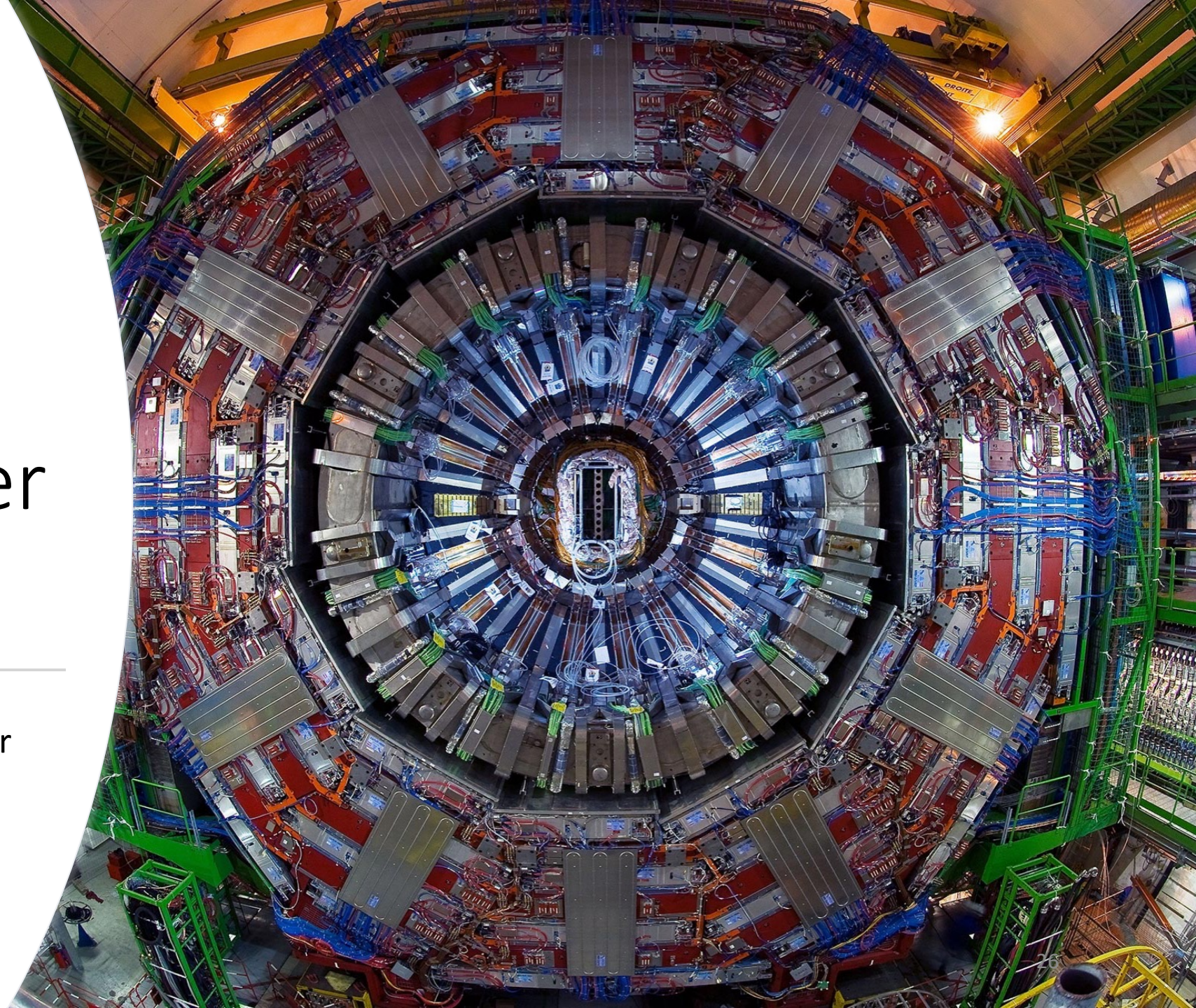
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Volume (Scale)

- Volume: Enterprises are overflowing with ever-growing data of all types, easily amassing terabytes even petabytes of information.
  - Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
  - Convert 350 billion annual meter readings to better predict power consumption.

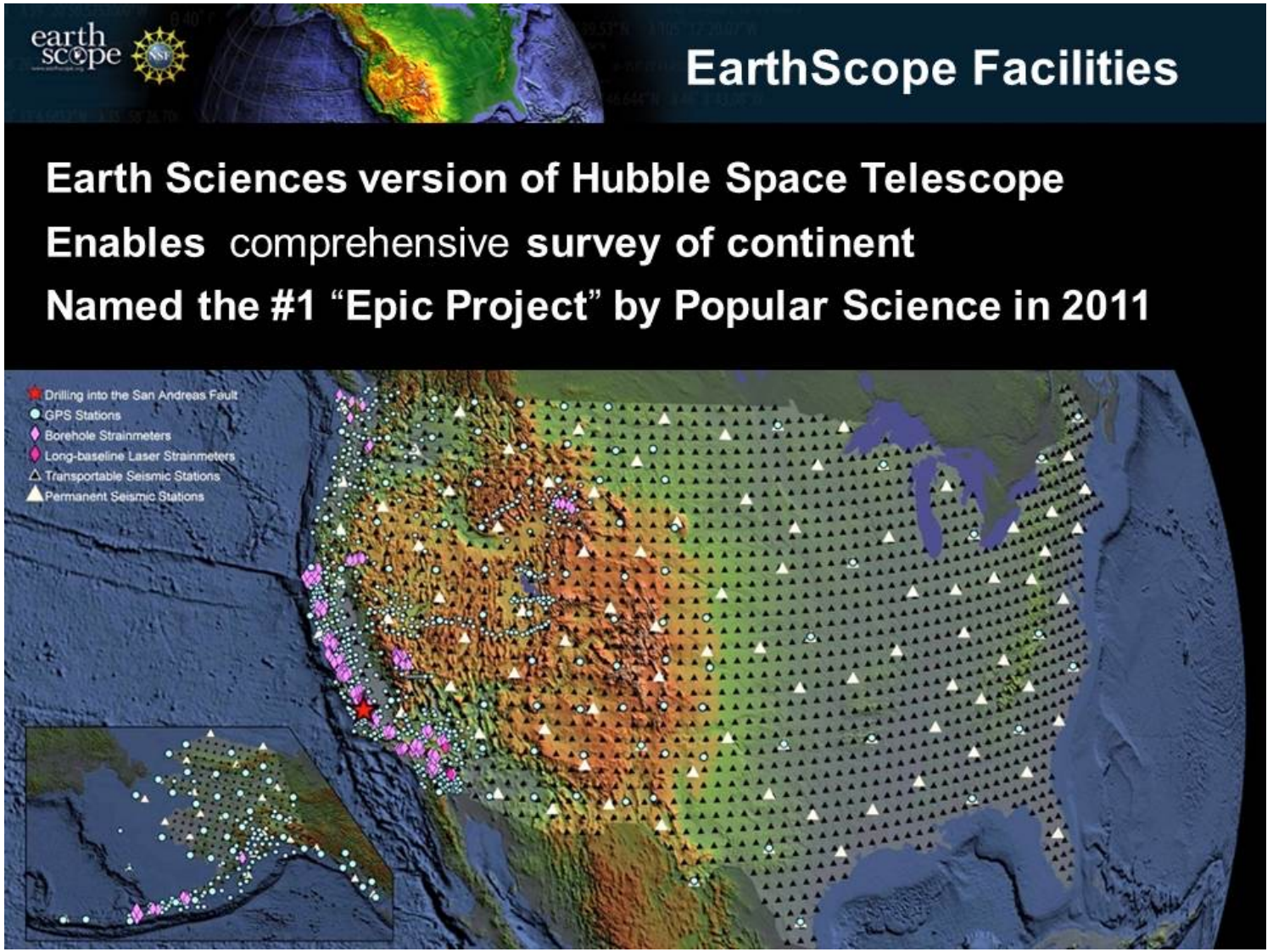
Example 1 : CERN’s Large Hydron Collider (LHC)

CERN’s LHC generates 15 PB a year

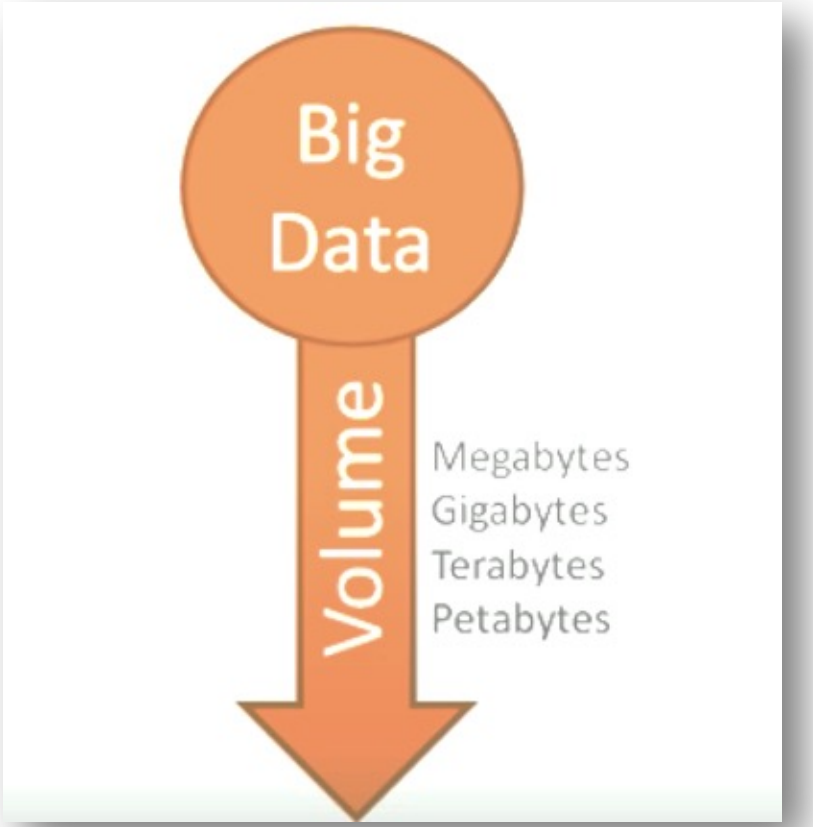


Example 2: The Earthscope

The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault as well as the plume of magna underneath Yellowstone and much, much more.



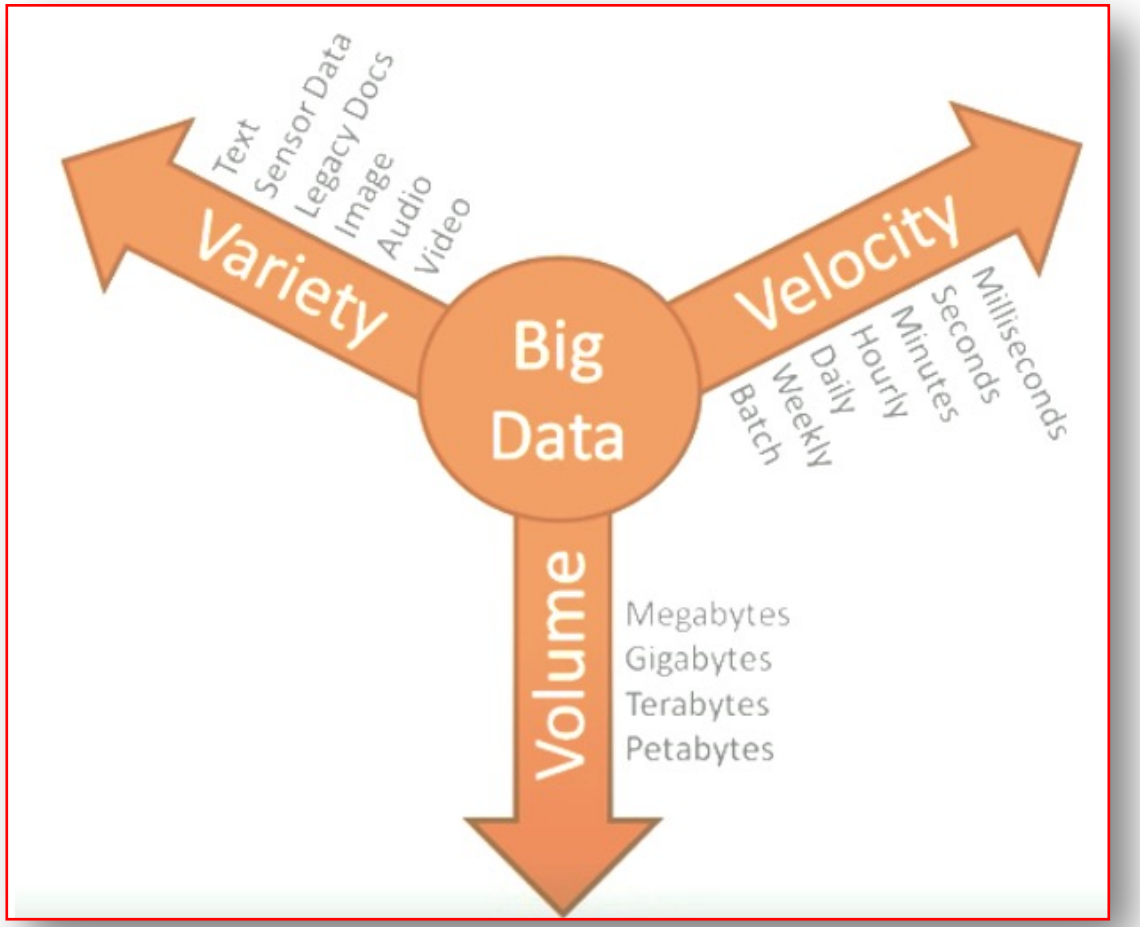
Question: Do you call a 150 GB dataset a big data?



150 GB

Question: How big is Big Data?

Text



Minute

150 GB

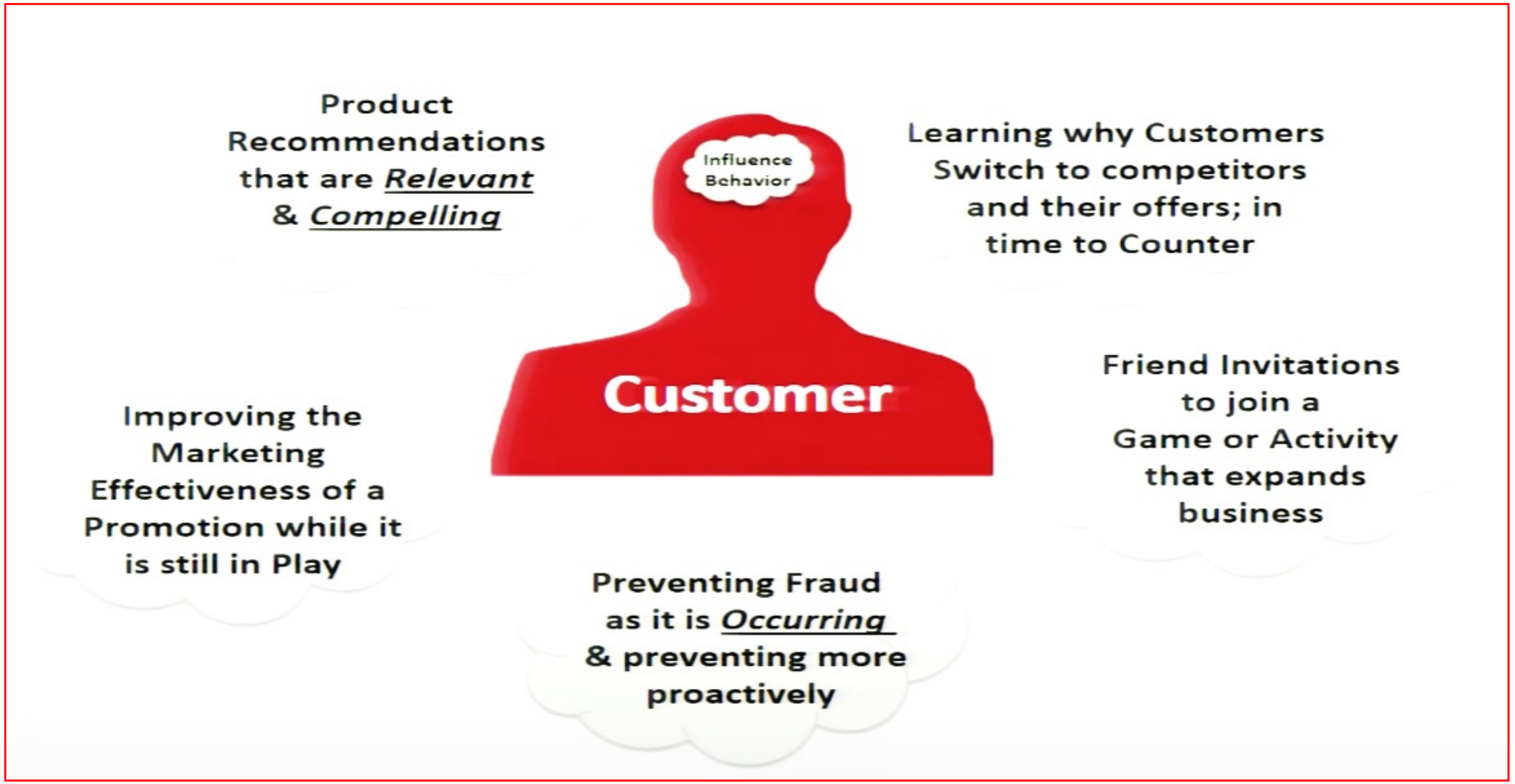
Velocity (Speed)

- Velocity: Sometimes 2 minutes is too late. For time sensitive processes such as catching fraud, big data must be used as it streams into the enterprise in order to maximize its value.
  - Scrutinize 5 million trade events created each day to identify potential fraud.
  - Analyze 500 million daily call details records in real-time to predict customer churn faster.

Examples: Velocity (Speed)

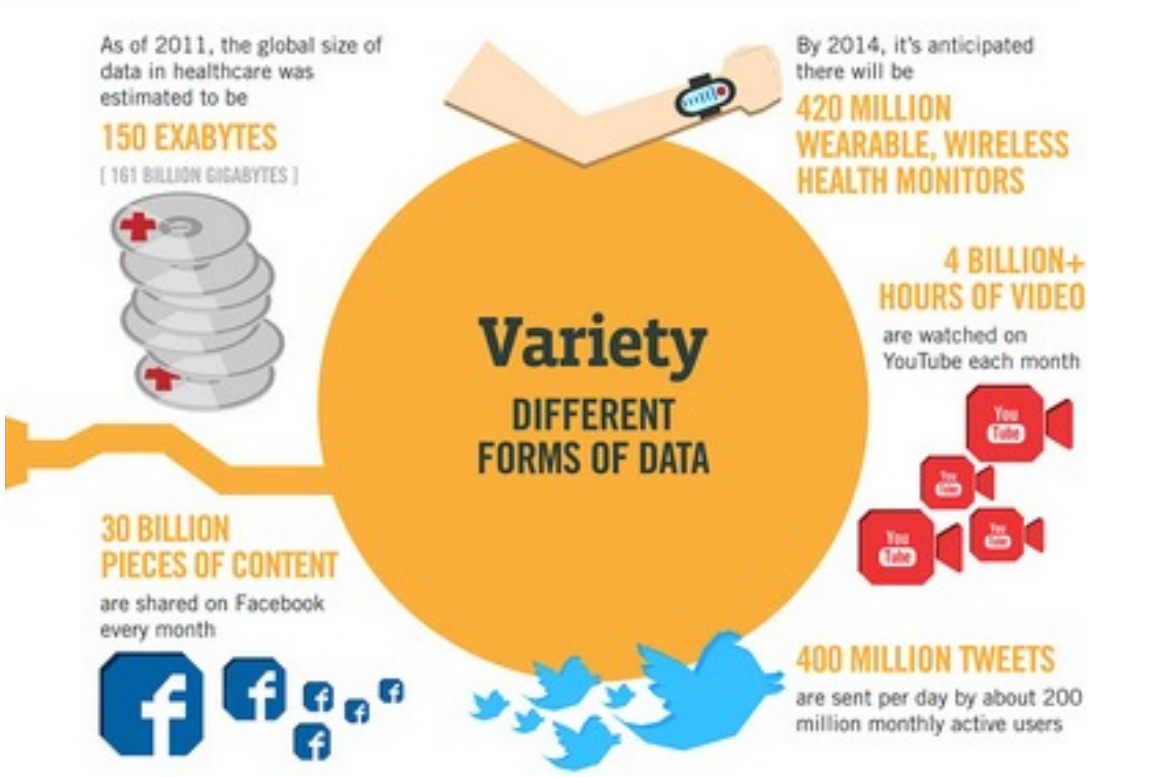
- Data is generated fast and need to be processed fast
- Online Data Analytics
- Late Decisions -> missing opportunities
- Examples:
  - E-Promotions: Based on your current location, your purchase history, what you like -> send you promotions right now for store next to you
  - Healthcare Monitoring: Sensors monitoring your activities and body for any abnormal measurements that require immediate action
  - Scientific instruments, Sensor Technology and Networks etc.

Real-time Analytics/Decision Requirements



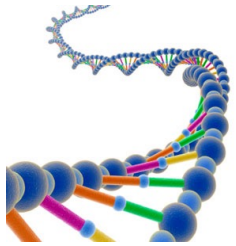
Variety (Complexity)

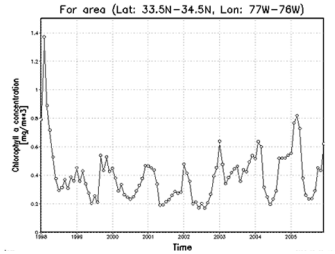
- Variety: Big data is any type of data-
  - Structured Data - example: tabular data
  - Unstructured Data - text, pdf, audio, video
  - Semi-structured Data - web data, log files, json documents

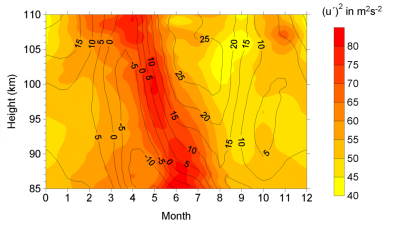


Examples: Variety (Complexity)

- Relational data: Tables/ Transactions/ Legacy data
- Text data: Web
- Semi-Structured data: XML, JSON
- Graph data: Social Network, Semantic Web (RDF),
- Streaming data: You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public data: online, weather, finance etc.



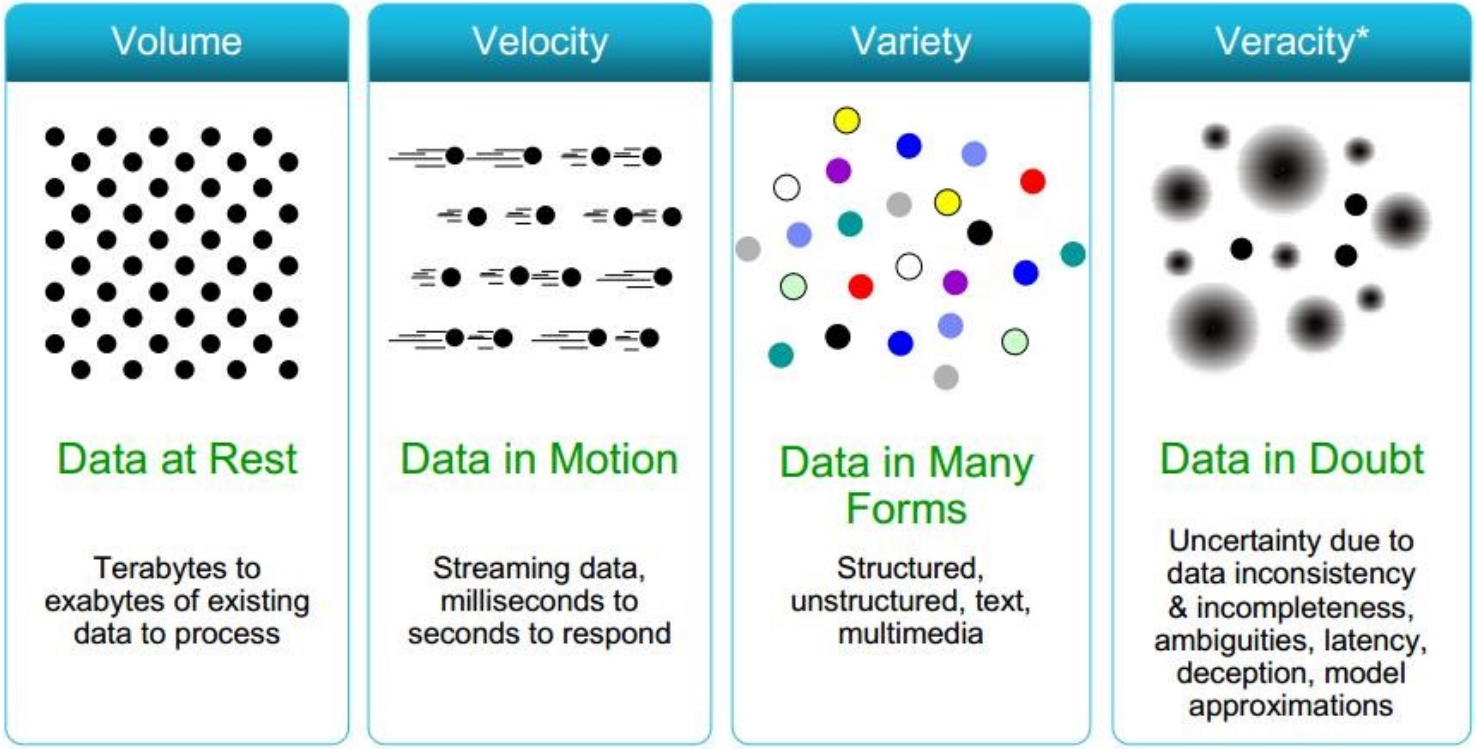




To extract knowledge➔ all these types of data need to linked together

The 3 Big V's (+1)

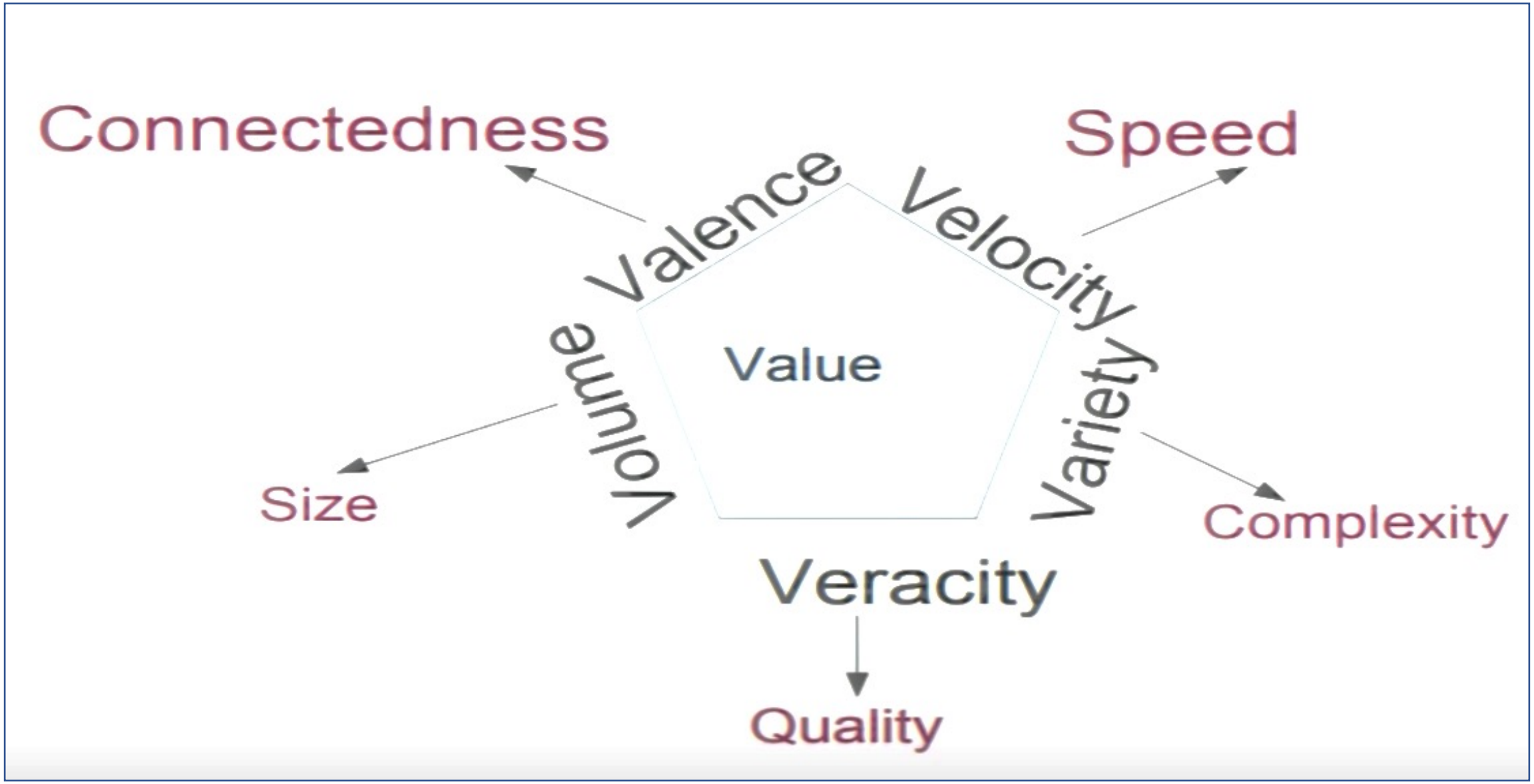
- Big 3 V's
  - Volume
  - Velocity
  - Variety
- Plus 1
  - Veracity



The 3 Big V's (+1) and (+N more)

- Plus many more
  - Veracity
  - Validity
  - Variability
  - Viscosity and Volatility
  - Viability
  - Venue
  - Vocabulary and Vagueness

  . . . .


Connectedness
Speed
Valence
Velocity
Volume
Value
Variety
Size
Complexity
Veracity
Quality

Value

- Integrating Data
  - Reducing data complexity
  - Increase data availability
  - Unify data Systems
  - All these points together will lead to increased data collaboration

    -> add value to your big data

Veracity

- Veracity refers to the biases, noise and abnormality in the data, trustworthiness of the data
- 1 in 3 businesses leaders don’t trust the information they use to make the decisions
  - How can you act upon an information if you don’t trust it?
  - Establishing the trust in big data presents a huge challenge as the variety and number of sources grow.

Valence

- Valence refers to the connectedness of big data
- Such as in the form of graph networks

Validity

- Accuracy and correctness of the data relative to a particular use

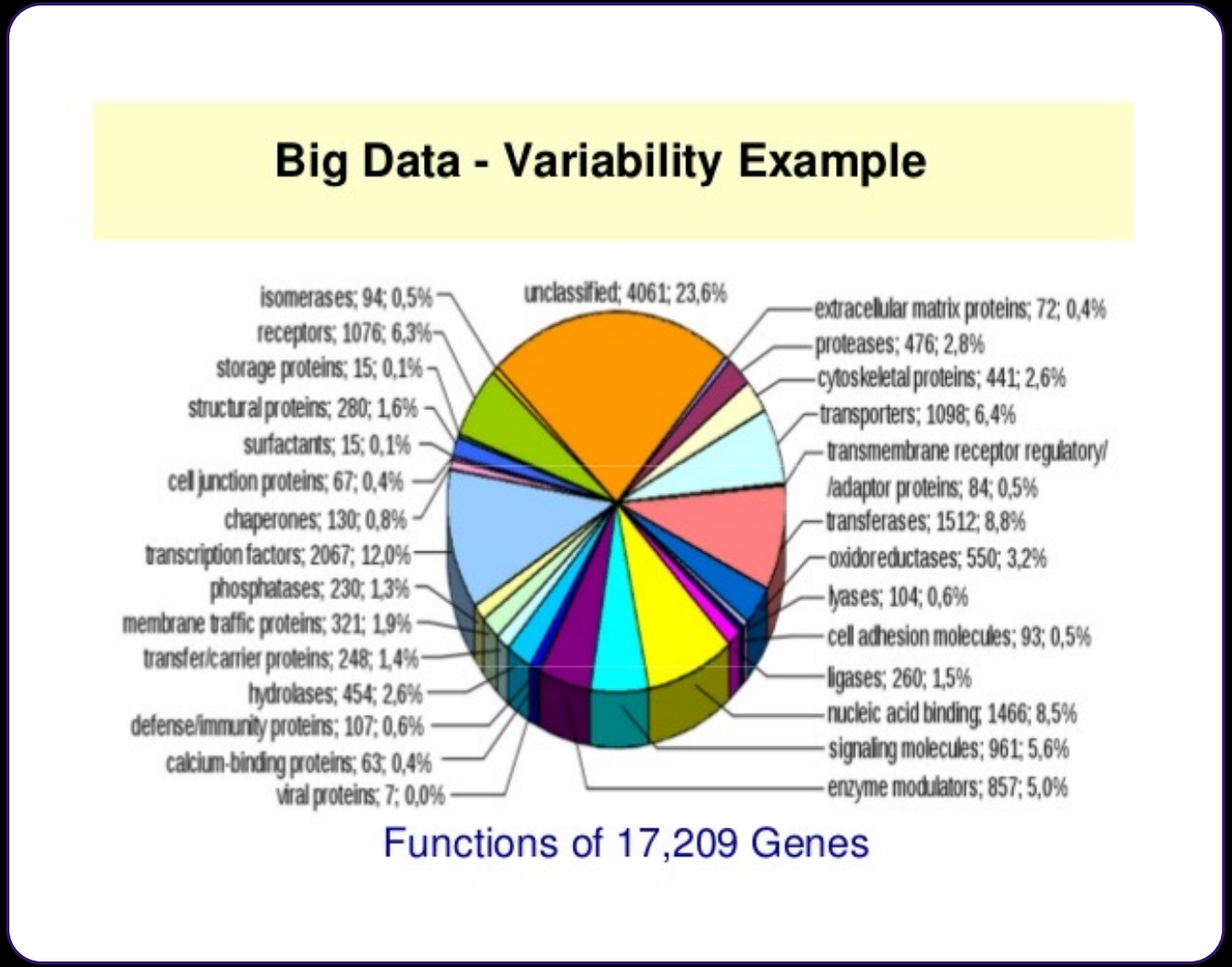
Example: **Gauging storm intensity**

satellite imagery vs social media posts

prediction quality vs human impact

Variability

- How the meaning of data changes over time
- Language evolution
- Data availability
- Sampling processes
- Changes in characteristics of the data source



Viscosity and Volatility

- Both are related to velocity
- Viscosity: Data velocity relative to timescale of event being studied
- Volatility: Rate of data loss and stable lifetime of data
  - Scientific data often has practically unlimited lifespan, but social/business data may evaporate in finite time

More V’s

Viability

Which data has meaningful relations to questions of interest?

What affect does time of a day have on buying behavior?

Venue

Where does the data live and how do you get it?

Vocabulary

Metadata describing structure, content and provenance

Schemas, semantics, ontologies, taxonomies and vocabularies

Vagueness

Confusion or an interpretation issue with results being returned

How to deal with Big Data ??

Dealing with Big Data

Distill big data down to small information

Parallel and automated analysis

Automation requires standardization

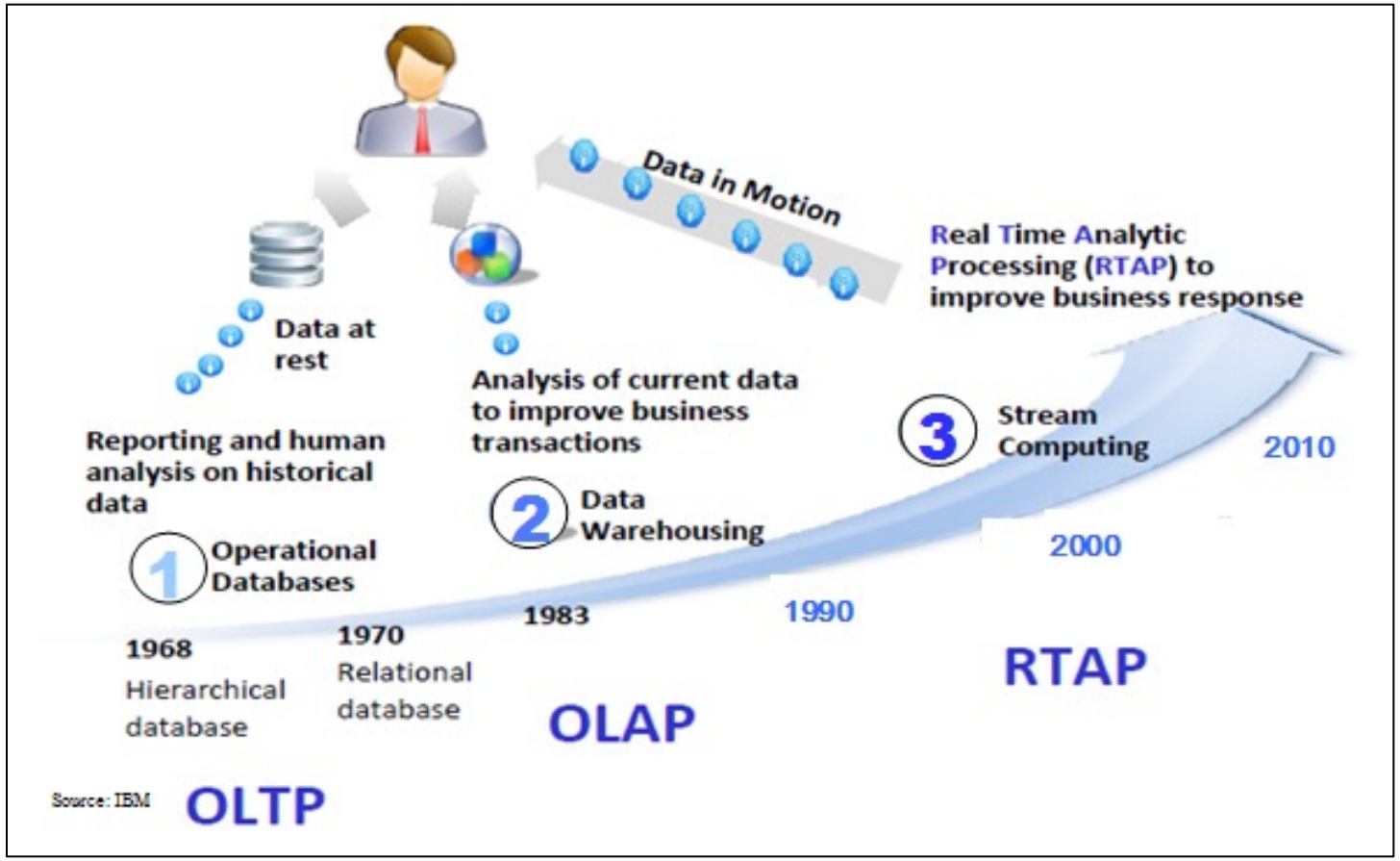
Standardize by reducing Variety:

Format

Standards

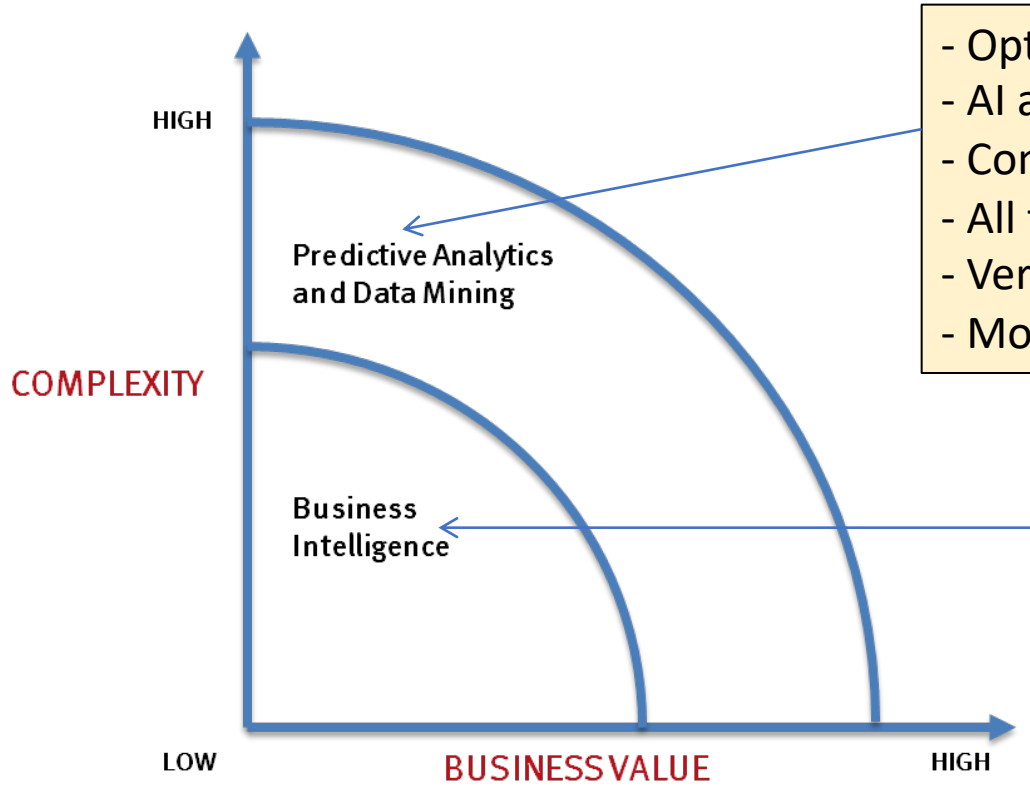
Structure

Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

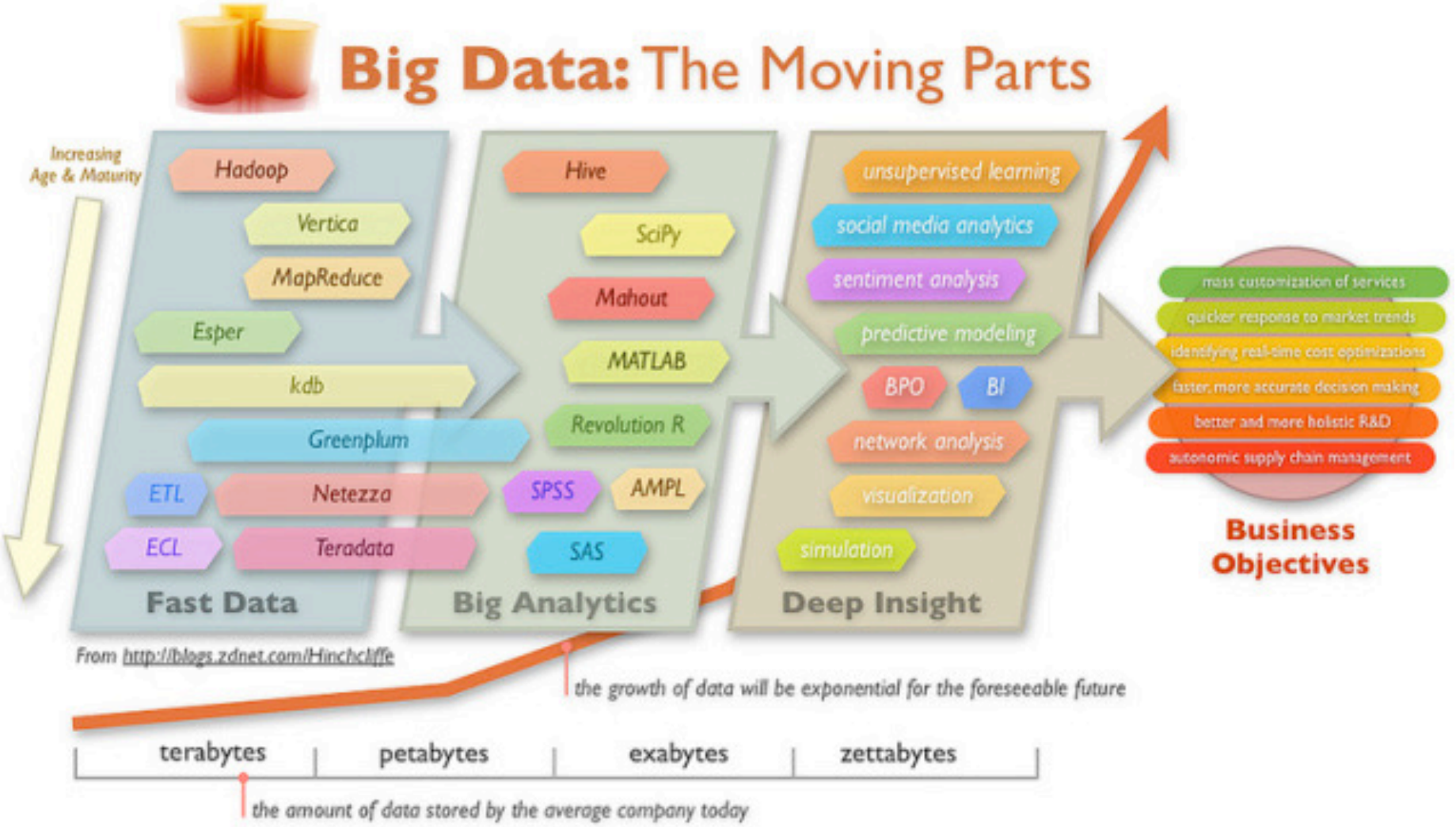
What is driving Big Data?



- Optimizations and predictive analytics
- AI and Machine Learning
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Big Data Technology



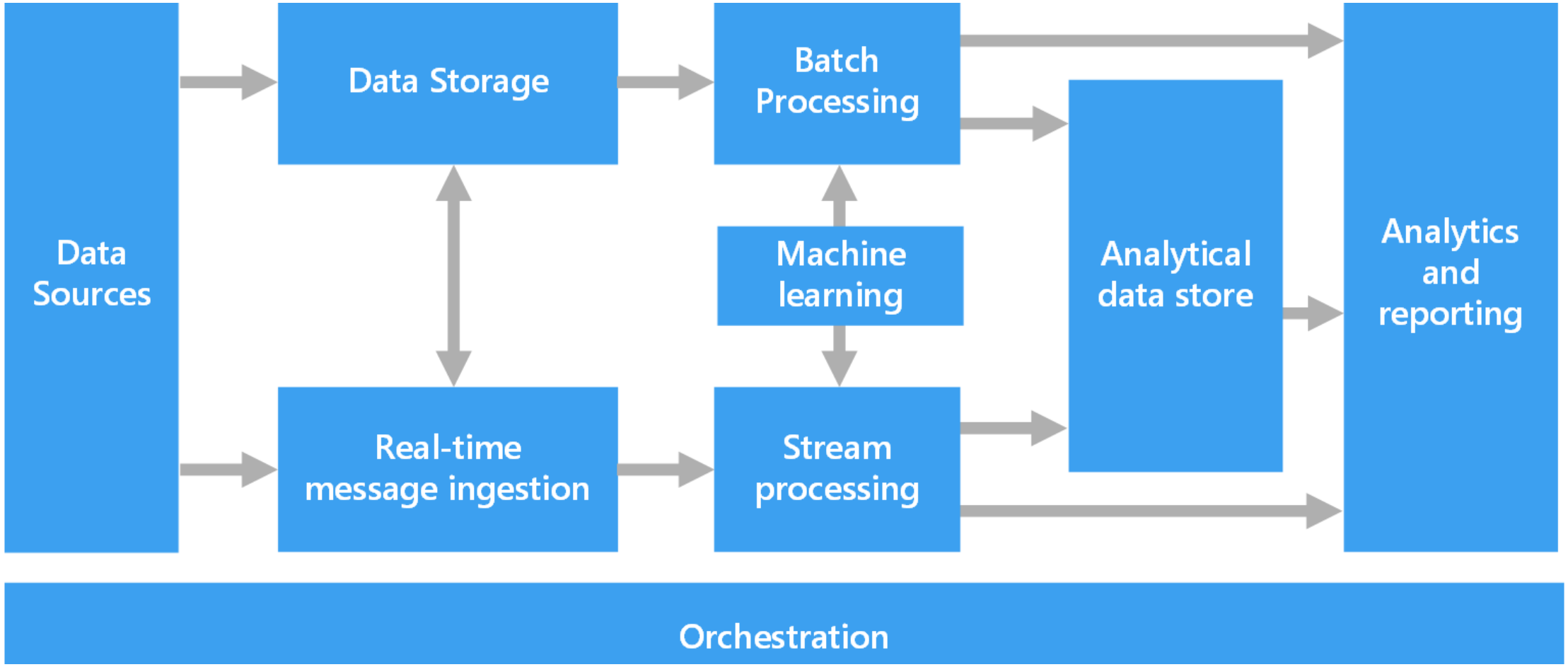
Benefits of Big Data Processing

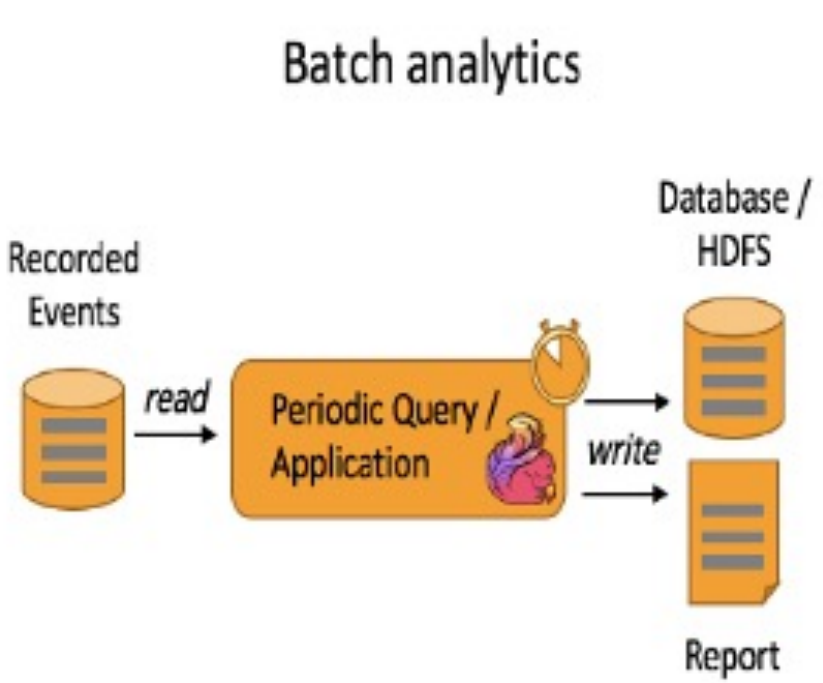
- Ability to process Big Data brings in multiple benefits, such as-
  - Businesses can utilize outside intelligence while taking decisions
- Access to social data from search engines and sites like Facebook, Twitter are enabling organizations to fine tune their business strategies.
  - Improved customer service
- Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.
  - Early identification of risk to the product/services, if any
  - Better operational efficiency
- Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

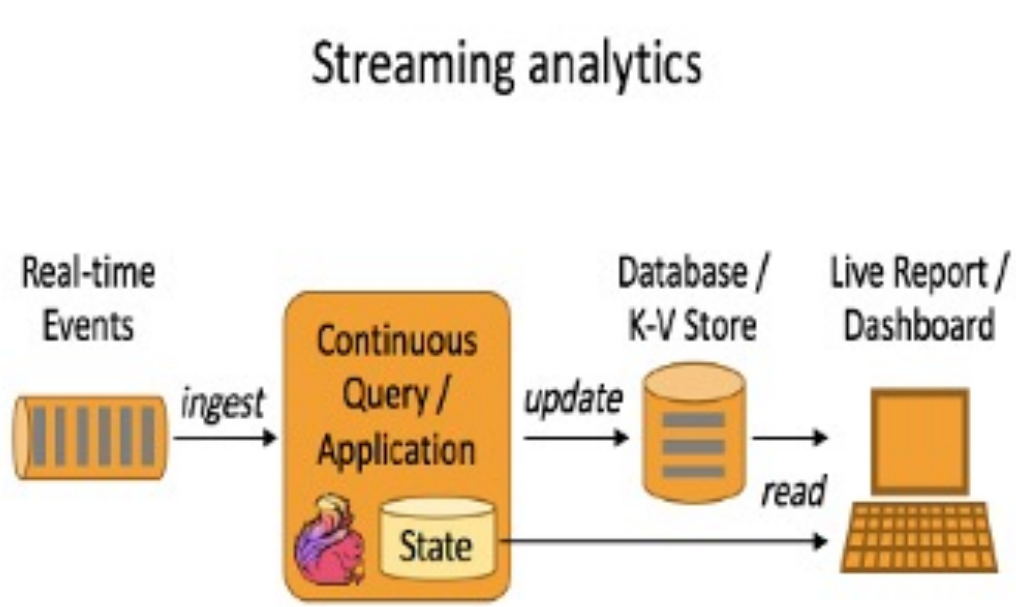
How does a Big Data Life Cycle Looks Like?

- The general categories of activities involved with big data processing are:
  - Ingesting data into the system
  - Persisting the data in storage
  - Computing and Analyzing data
  - Visualizing the results

Big Data Life Cycle/Architecture







Batch processing is defined as a series of jobs that are connected to each other or executed one after another in sequence or in parallel.

After executing all the jobs, an output is generated, and the information is consolidated to generate a final result.

The real-time data processing continuously receives data that is under constant change, such as information relating to air traffic control, customer services and so on.

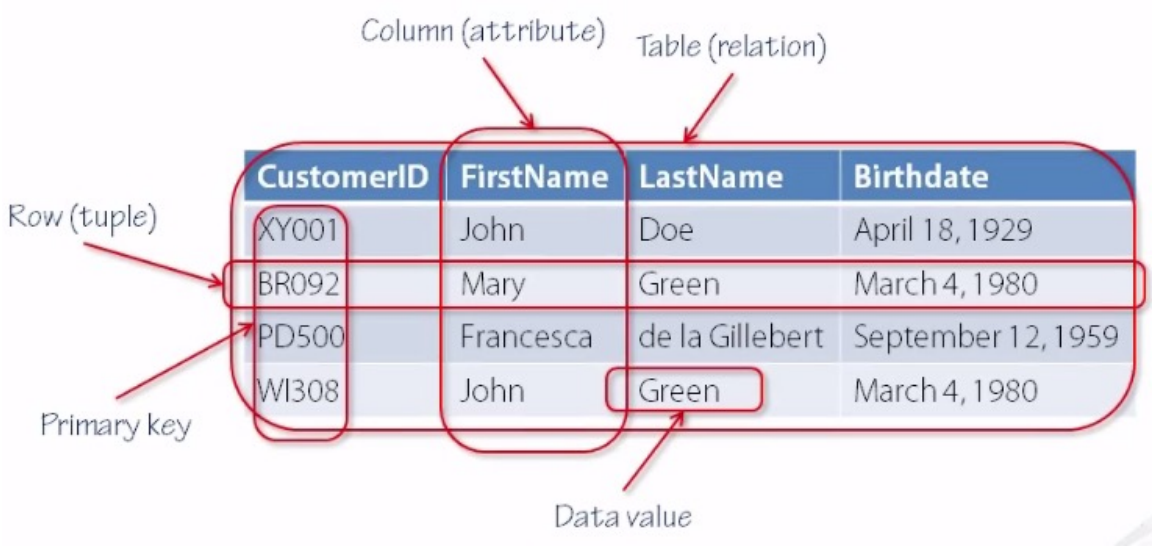
Relational Databases

SQL databases:

- SQL databases are primarily called as Relational Databases (RDBMS)
- Tabular Format
- Predefined schema
- SQL is used to work with RDBMS

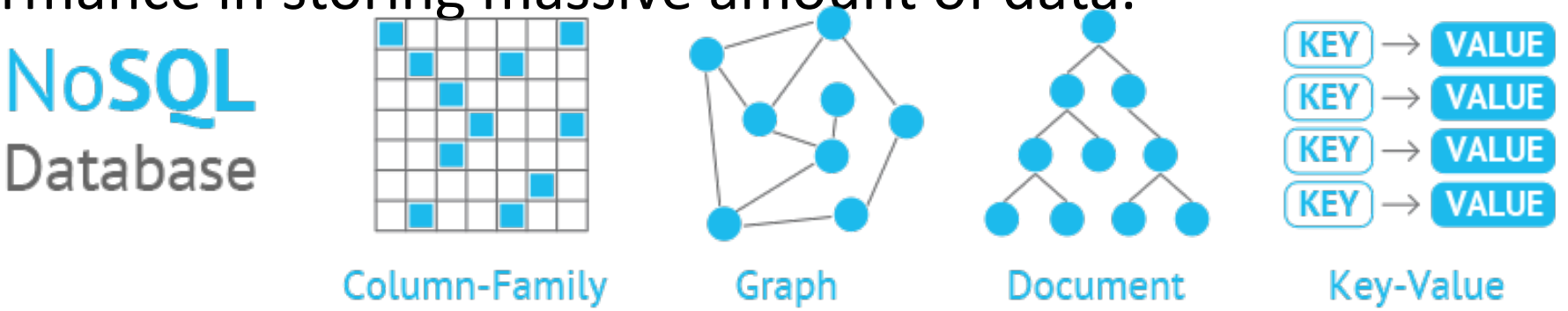
This model organizes data into one or more tables (or relations) of columns and rows, with a unique key identifying each row.

- Rows are also called records or tuples.
- Columns are also called attributes.



NoSQL

- While the traditional SQL can be effectively used to handle large amount of structured data, we need NoSQL (Not Only SQL) to handle semi-structured data.
- NoSQL databases store semi-structured data with no particular schema.
- Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data.



NoSQL Databases

NoSQL databases:

– Not Only SQL

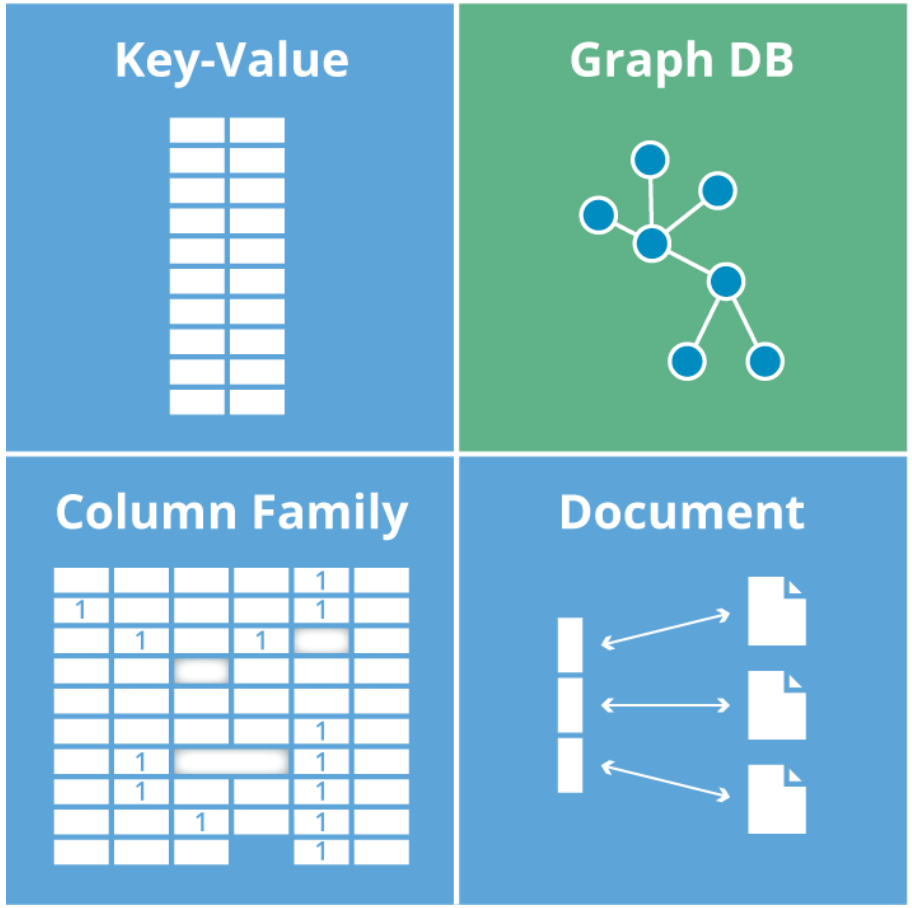
– NoSQL database are primarily called as non-relational or distributed database

– Collection of key-value pairs

– No predefined schema

Examples of NoSQL data:

- Graph data
- Geo-spatial
- Engineering data
- Sensor data
- Network log data



NoSQL data formats

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

Key-Value store

Document oriented database

Document 1

```
{
  "id": "1",
  "name": "John Smith",
  "isActive": true,
  "dob": "1964-30-08"
}
```

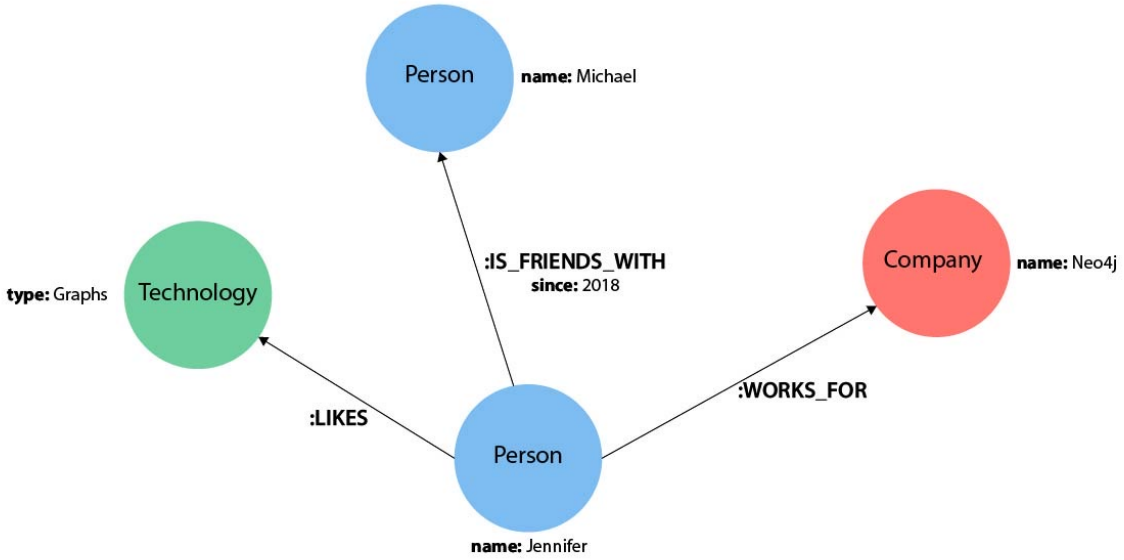
Document 2

```
{
  "id": "2",
  "fullName": "Sarah Jones",
  "isActive": false,
  "dob": "2002-02-18"
}
```

Document 3

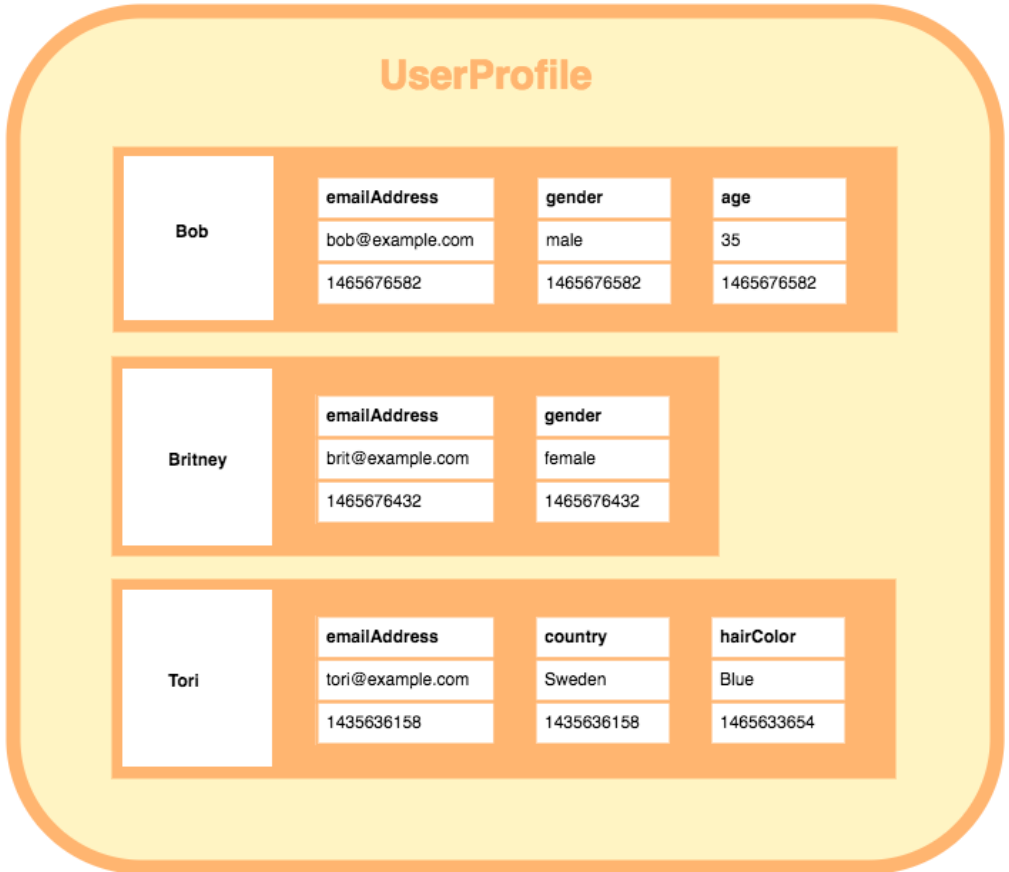
```
{
  "id": "3",
  "fullName":
  {
    "first": "Adam",
    "last": "Stark"
  },
  "isActive": true,
  "dob": "2015-04-19"
}
```

NoSQL data formats

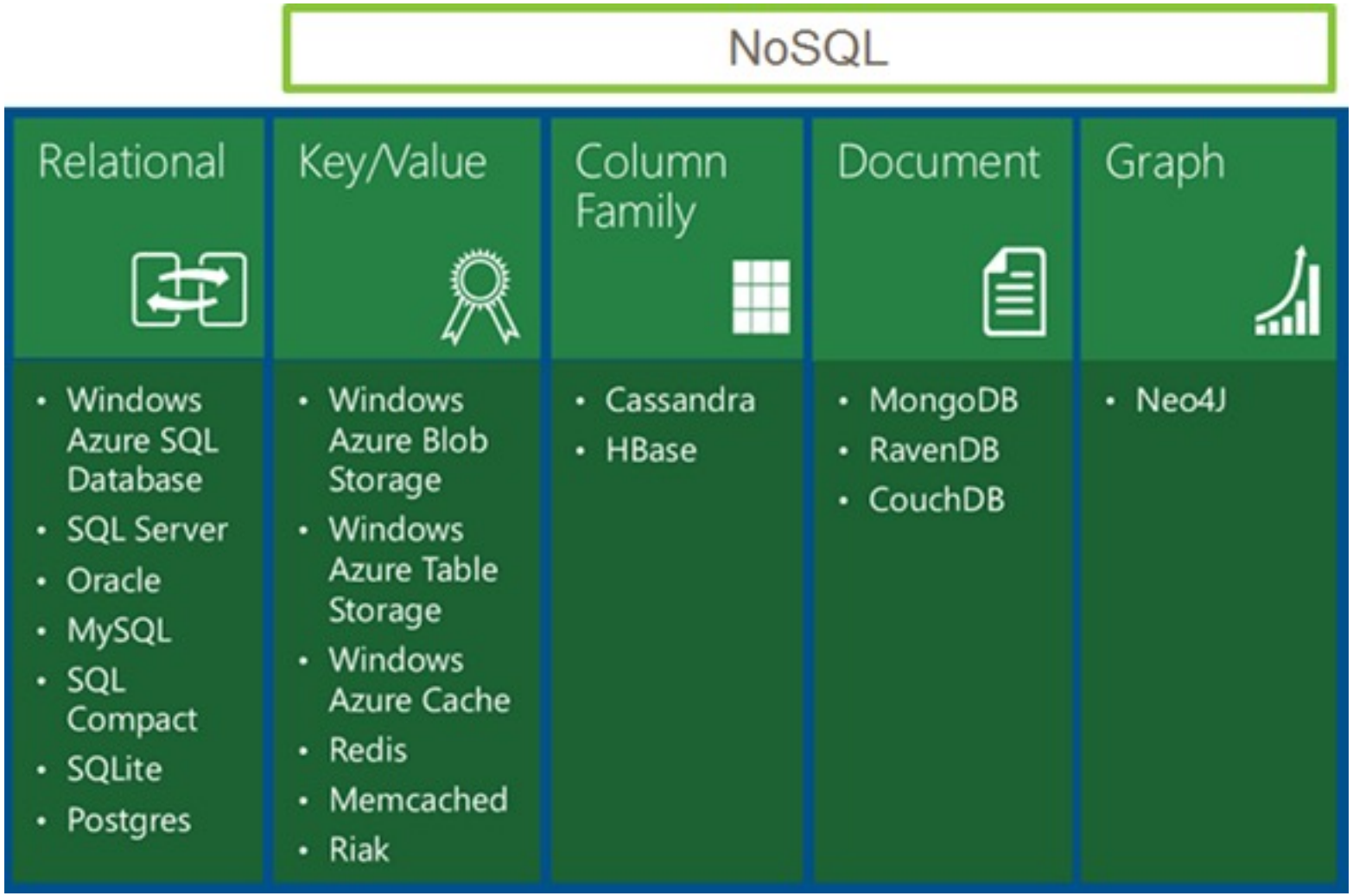


Graph database

Column family database



Database Options



No SQL Database Providers

Top 4 NoSQL Databases

	MongoDB	Cassandra	Elasticsearch	Couchbase
Description	One of the most popular document stores	Wide-column store based on ideas of BigTable and DynamoDB	A modern search and analytics engine based on Apache Lucene	JSON-based document store derived from CouchDB with a Memcached-compatible interface
Database model	Document store	Wide Column store	Search engine	Document store
Developer	MongoDB, Inc.	Apache Software Foundation	Elastic	Couchbase, Inc.
Release	2009	2008	2010	2011
Language	C++	Java	Java	C, C+ + and Erlang
Server-side scripts	JavaScript	No	Yes	View functions in JavaScript
Replication methods	Master-slave replication	Selectable replication factor	Yes	Master-master replication, Master-slave replication
Best use	If you need dynamic queries. If you prefer to define indexes, not map and reduced functions. If you need good performance on a big DB and when your data changes too much	When data you need to store doesn't fit on server, but requires friendly familiar interface to it	When you have objects with flexible fields, and you need "advanced search" functionality	Any application that requires low-latency data access, high concurrency support and high availability

Types of Databases

- Relational Databases: SQL Server, PostgreSQL, SQLite, MySQL, Azure SQL
- Non-relational Databases: Azure Cosmos DB
  - Key-Value Pair (KVP) Databases: Data is stored as Key:Value, e.g., Riak Key-Value Database, Redis, and Oracle NoSQL database
  - Document Databases: Store documents or web pages, e.g., MongoDB, CouchDB
  - Columnar Databases: Store data in columns, e.g., Hbase, Cassandra
  - Graph Databases: Stores nodes and relationship, e.g., Neo4j
  - In-Memory Database (IMDB): All data in memory. For real time applications
- Cloud Databases: Any data that is run in a cloud using IAAS, DAAS, PAAS

Applications of Big Data

- **Smarter Healthcare**: Making use of the petabytes of patient's data, the organization can extract meaningful information and then build applications that can predict the patient's deteriorating condition in advance.
- **Telecom**: Telecom sectors collects information, analyzes it and provide solutions to different problems. By using Big Data applications, telecom companies have been able to significantly reduce data packet loss, which occurs when networks are overloaded, and thus, providing a seamless connection to their customers.
- **Retail**: Retail has some of the tightest margins, and is one of the greatest beneficiaries of big data. The beauty of using big data in retail is to understand consumer behavior. Amazon's recommendation engine provides suggestion based on the browsing history of the consumer.
- **Traffic control**: Traffic congestion is a major challenge for many cities globally. Effective use of data and sensors will be key to managing traffic better as cities become increasingly densely populated.
- **Manufacturing**: Analyzing big data in the manufacturing industry can reduce component defects, improve product quality, increase efficiency, and save time and money.
- **Search Quality**: Every time we are extracting information from google, we are simultaneously generating data for it. Google stores this data and uses it to improve its search quality.

Problems with Big Data

- **Data Quality** – The problem here is the 4th V i.e. Veracity. The data here is very messy, inconsistent and incomplete. Dirty data cost $600 billion to the companies every year in the United States.
- **Discovery** – Finding insights on Big Data is like finding a needle in a haystack. Analyzing petabytes of data using extremely powerful algorithms to find patterns and insights are very difficult.
- **Storage** – The more data an organization has, the more complex the problems of managing it can become. The question that arises here is “Where to store it?”. We need a storage system which can easily scale up or down on-demand.
- **Analytics** – In the case of Big Data, most of the time we are unaware of the kind of data we are dealing with, so analyzing that data is even more difficult.
- **Security** – Since the data is huge in size, keeping it secure is another challenge. It includes user authentication, restricting access based on a user, recording data access histories, proper use of data encryption etc.
- **Lack of Talent –** There are a lot of Big Data projects in major organizations, but a sophisticated team of developers, data scientists and analysts who also have sufficient amount of domain knowledge is still a challenge.

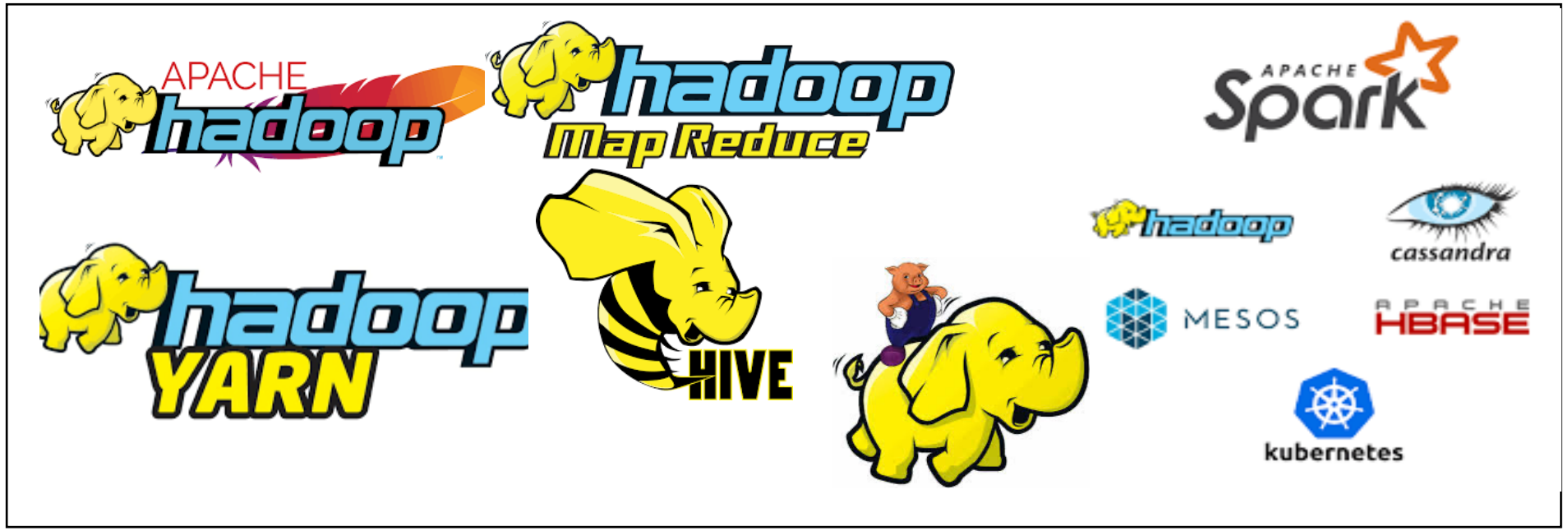
Recap

- In this part of the lecture, we have defined Big Data and discussed the challenges and applications of the Big Data.
- We have also described characteristics of Big Data that are Volume, Velocity, Variety and many more V’s.

Big Data Technologies

Big Data Technologies

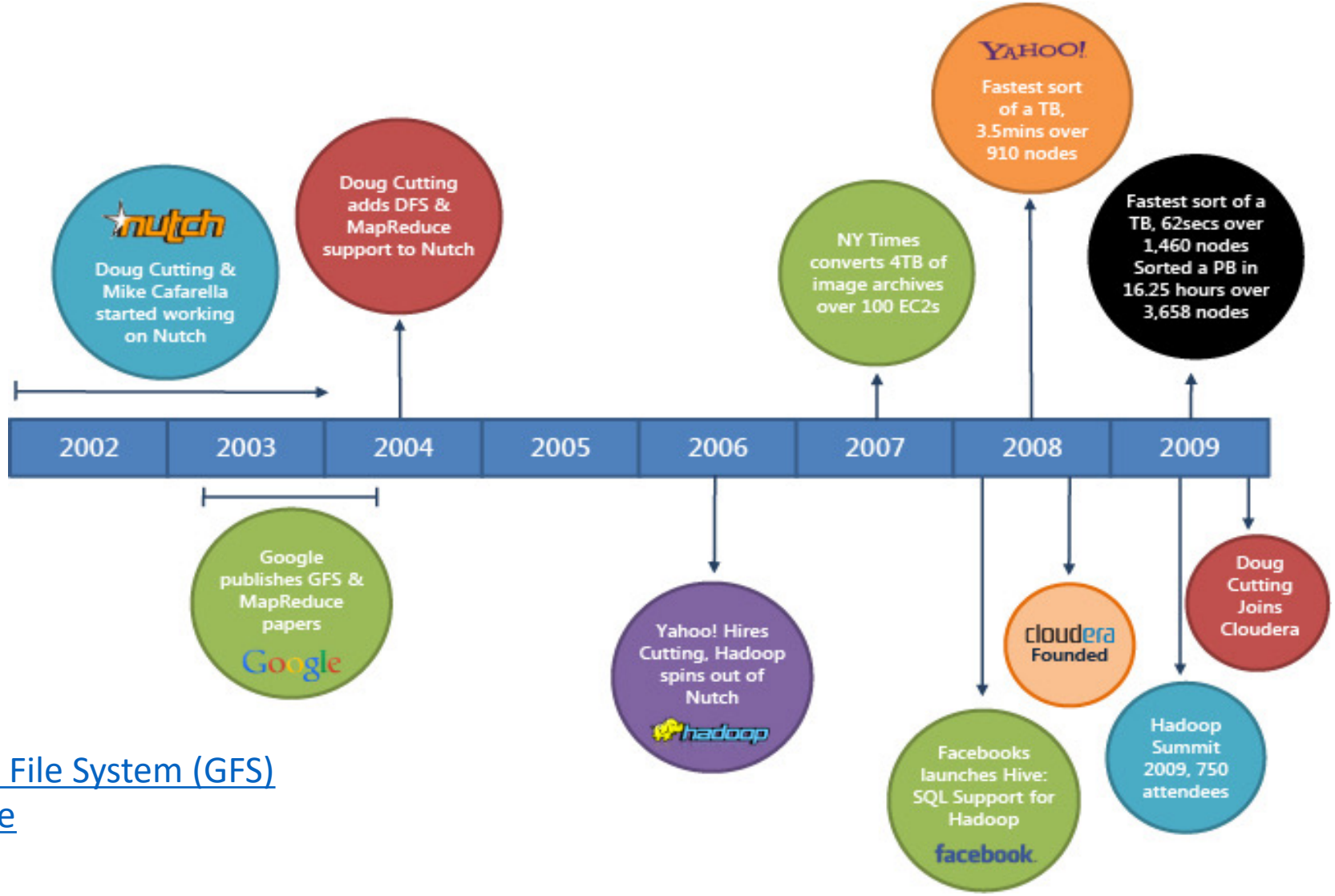
- In this part, we will discuss a brief introduction to Big Data Enabling Technologies.



Big Data Technologies

- Big Data is used for a collection of datasets so large and complex that it is difficult to process using traditional tools.
- A recent survey says that 80% of the data created in the world is unstructured.
- One challenge is how can we store and process this big amount of data. In this part of the lecture, we will discuss the top technologies used to store and analyze Big Data.

Hadoop History

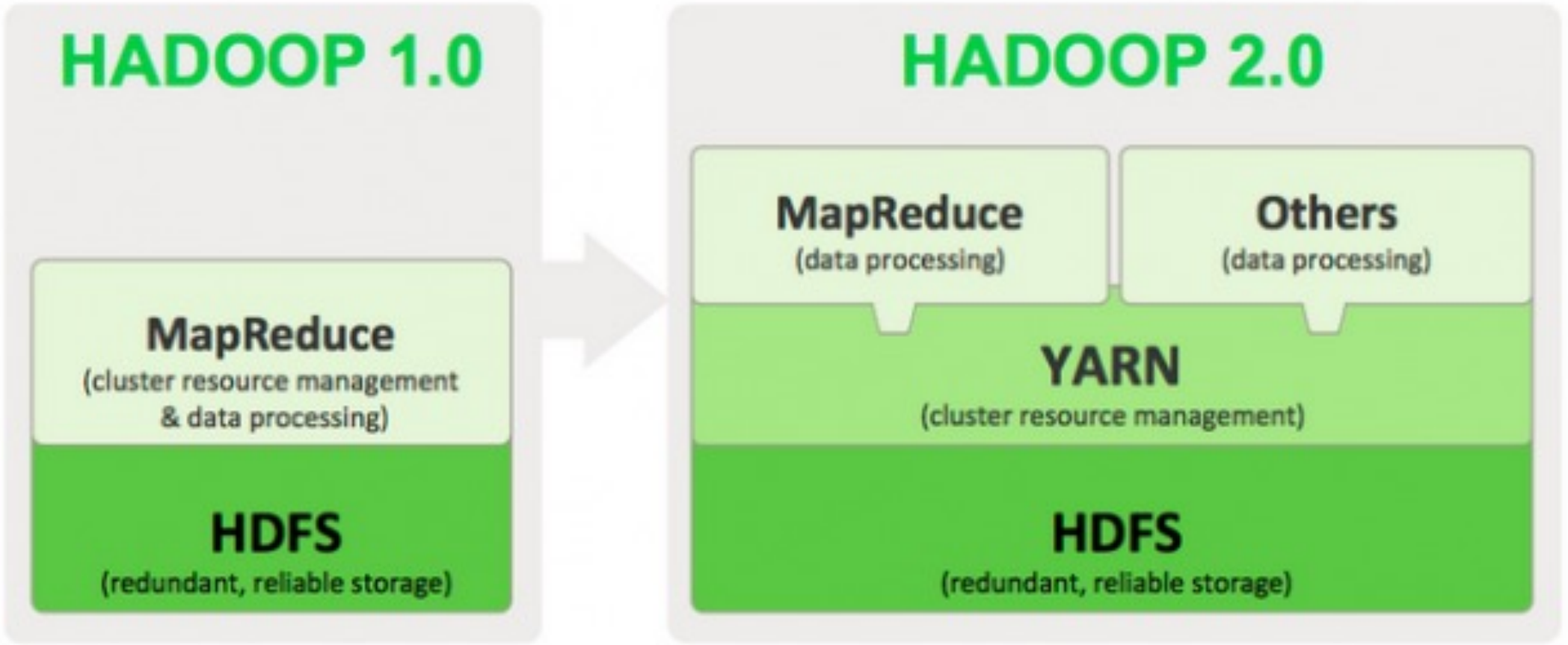


[Google distributed File System (GFS)]
[Google MapReduce]

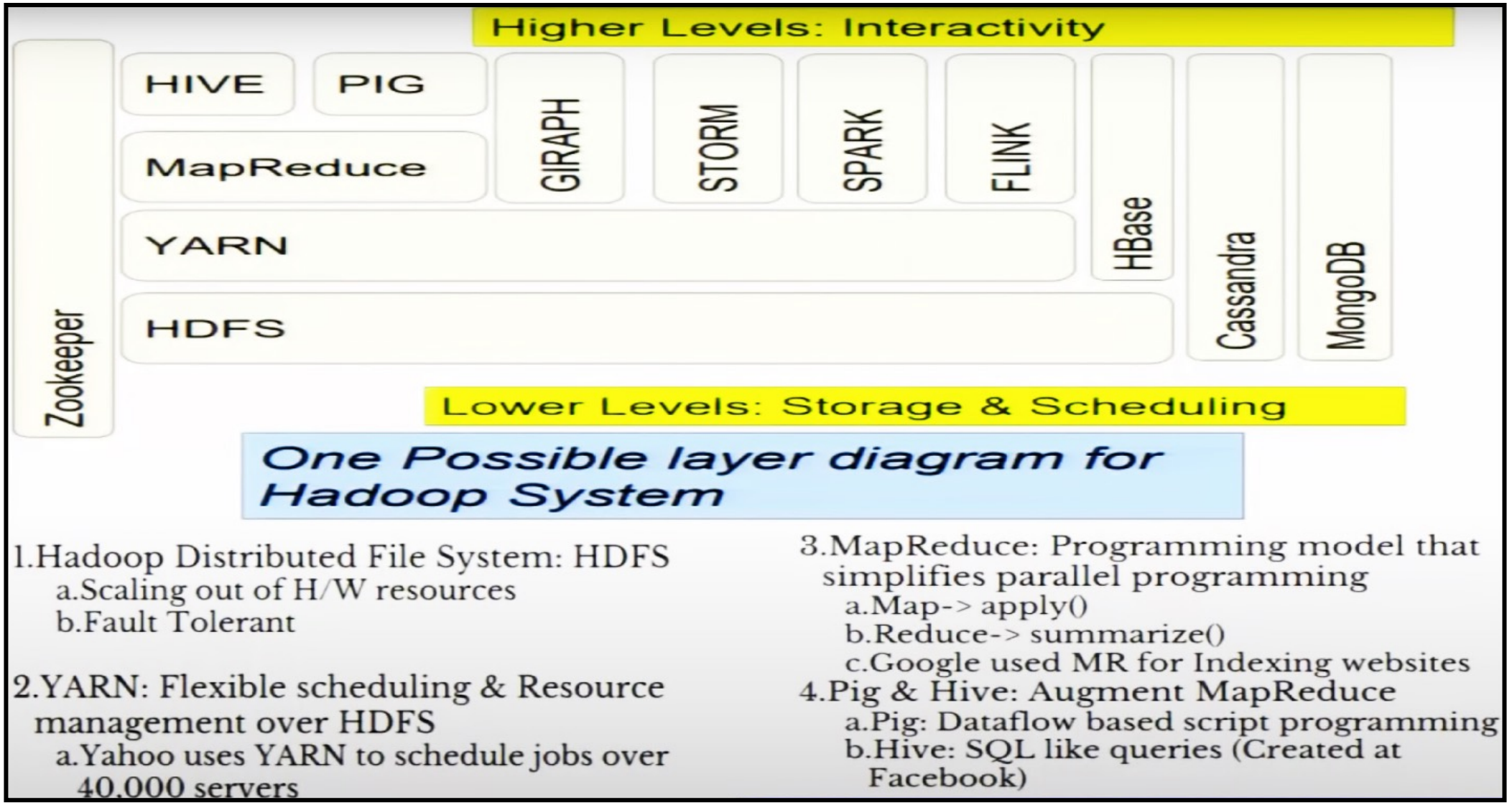
Apache Hadoop

- It is an open-source software developed as a project by Apache Software Foundation. **Doug Cutting** and **Michael J. Cafarella created Hadoop**. In the year 2008 Yahoo gave Hadoop to Apache Software Foundation.
- It has three main parts:
  - Hadoop Distributed File System (HDFS) is the storage of system of Hadoop which splits big data and distribute across many nodes in a cluster.
  - Yet Another Resource Negotiator (YARN) it simplifies the resource management.
  - MapReduce: Programming model that simplifies parallel programming.

Apache Hadoop



Hadoop Ecosystem



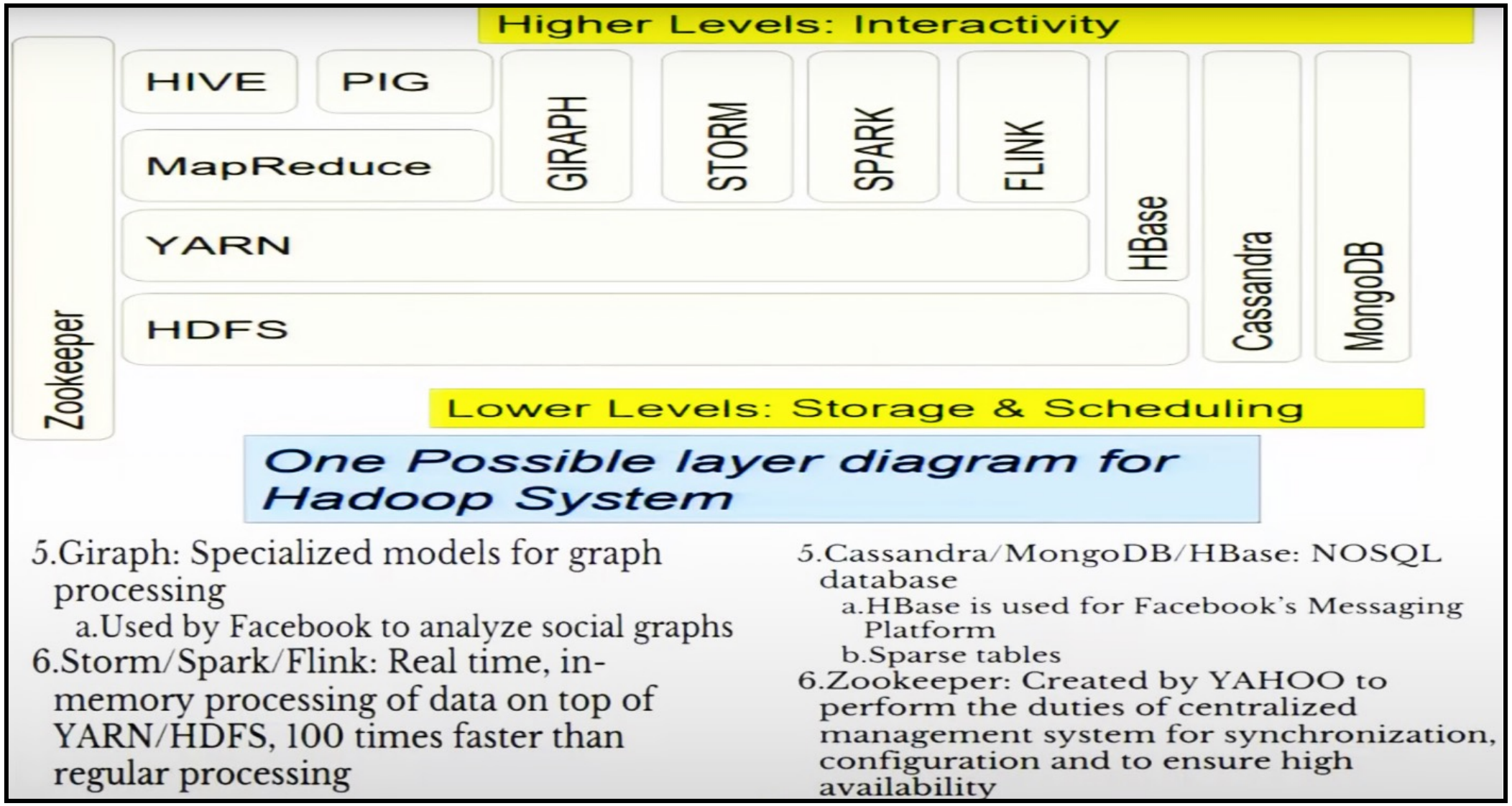
1.Hadoop Distributed File System: HDFS
   a.Scaling out of H/W resources
   b.Fault Tolerant

2.YARN: Flexible scheduling & Resource management over HDFS
   a.Yahoo uses YARN to schedule jobs over 40,000 servers

3.MapReduce: Programming model that simplifies parallel programming
   a.Map-> apply()
   b.Reduce-> summarize()
   c.Google used MR for Indexing websites

4.Pig & Hive: Augment MapReduce
   a.Pig: Dataflow based script programming
   b.Hive: SQL like queries (Created at Facebook)

Hadoop Ecosystem

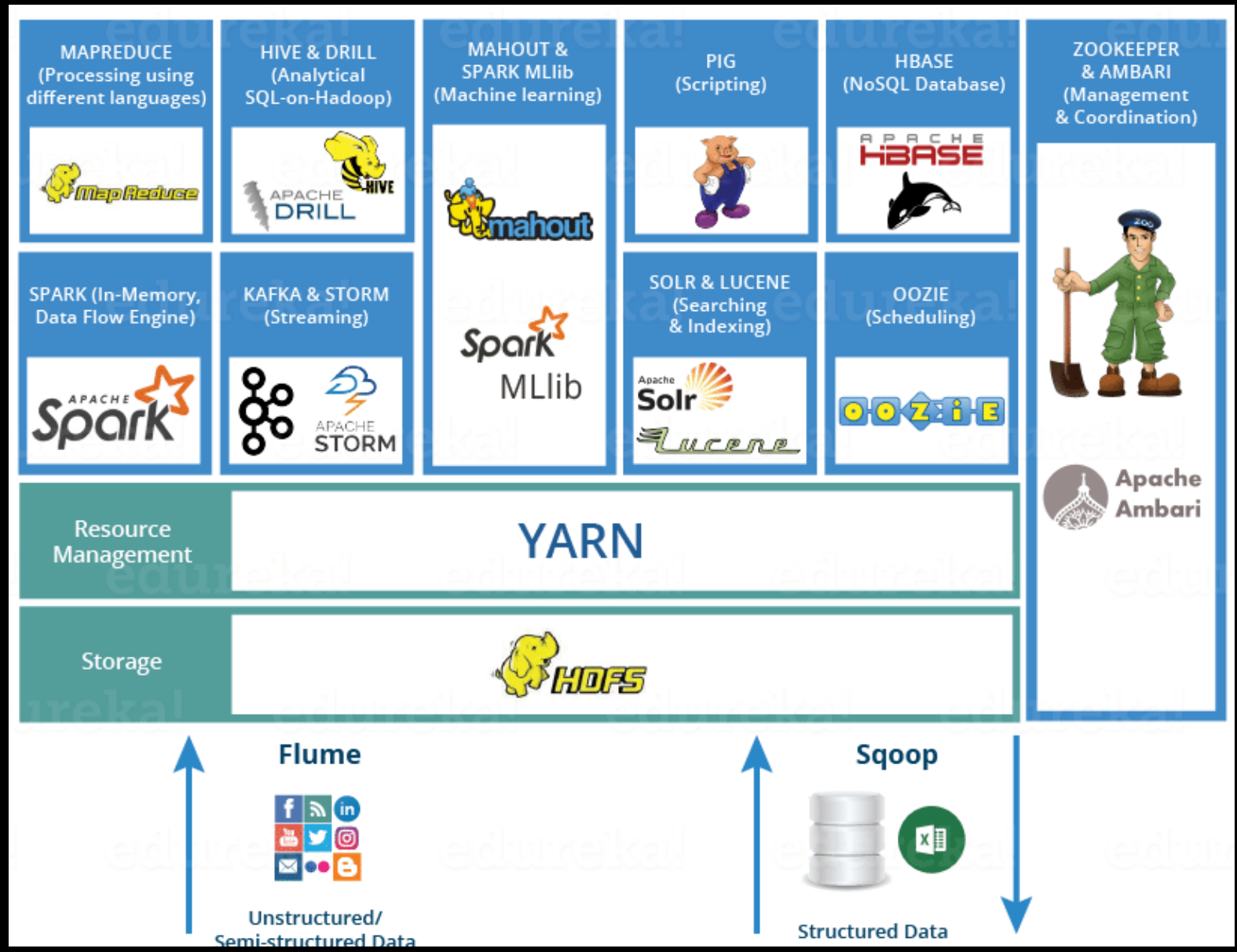


5.Giraph: Specialized models for graph processing
   a.Used by Facebook to analyze social graphs
6.Storm/Spark/Flink: Real time, in-memory processing of data on top of YARN/HDFS, 100 times faster than regular processing

5.Cassandra/MongoDB/HBase: NOSQL database
   a.HBase is used for Facebook's Messaging Platform
   b.Sparse tables
6.Zookeeper: Created by YAHOO to perform the duties of centralized management system for synchronization, configuration and to ensure high availability

Hadoop Ecosystem

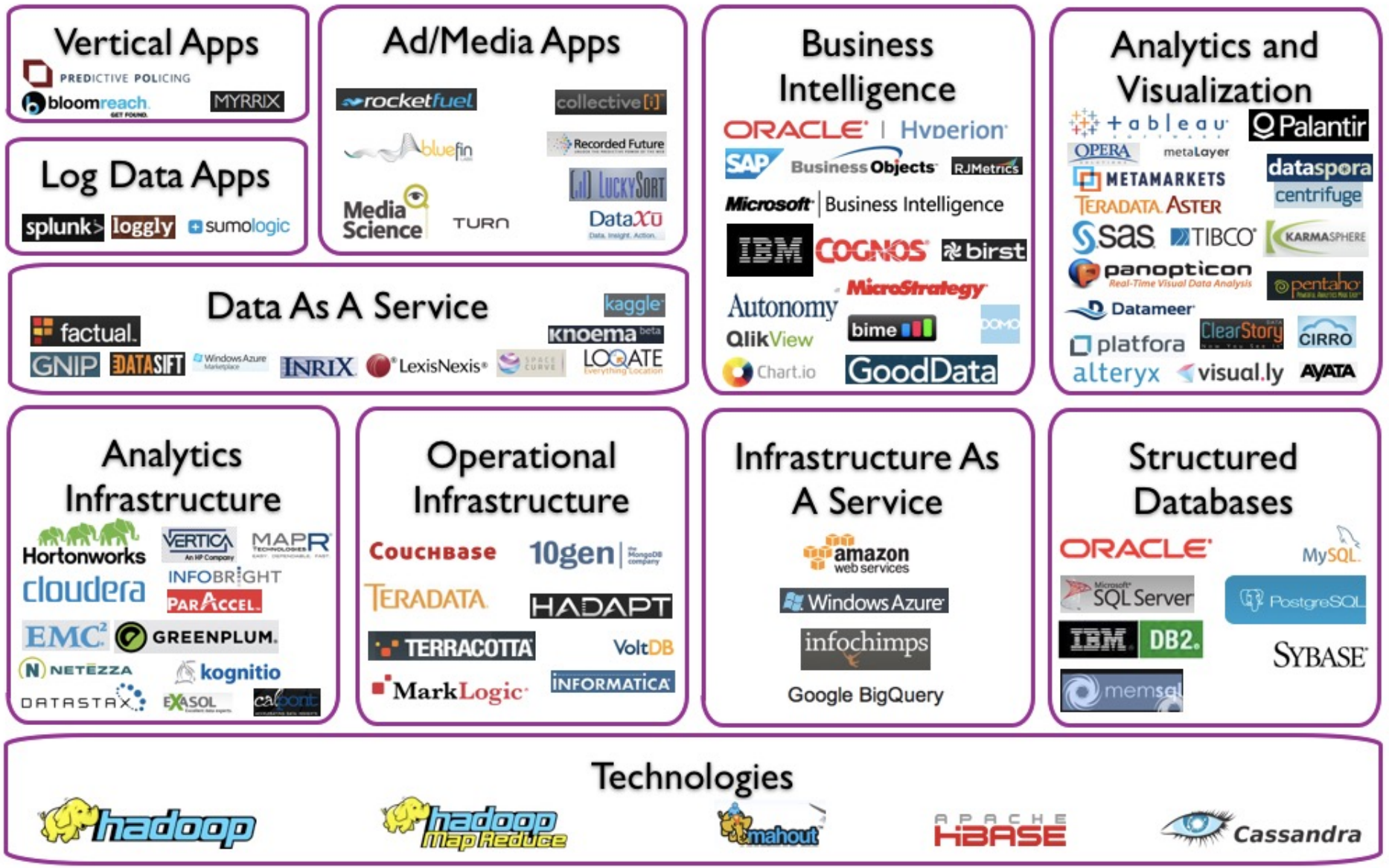
- **HDFS**> *Hadoop Distributed File System*
- **YARN**-> *Yet Another Resource Negotiator*
- **MapReduce**-> *Parallel programming model*
- **Spark** -> In-memory Data Processing
- **PIG, HIVE**-> *Data Processing Services using Query (SQL- like)*
- **HBase** -> *NoSQL Database*
- **Mahout, Spark MLlib** -> *Machine Learning*
- **Apache Drill** -> *SQL on Hadoop*
- **Zookeeper** -> *Managing Cluster*
- **Oozie** -> *Job Scheduling*
- **Flume, Sqoop** -> *Data Ingesting Services*
- **Solr & Lucene** -> *Searching & Indexing*
- **Ambari** -> *Provision, Monitor and Maintain cluster*



How many technologies do we have for big data?

- Big Data Landscape
- Big Data Glossary

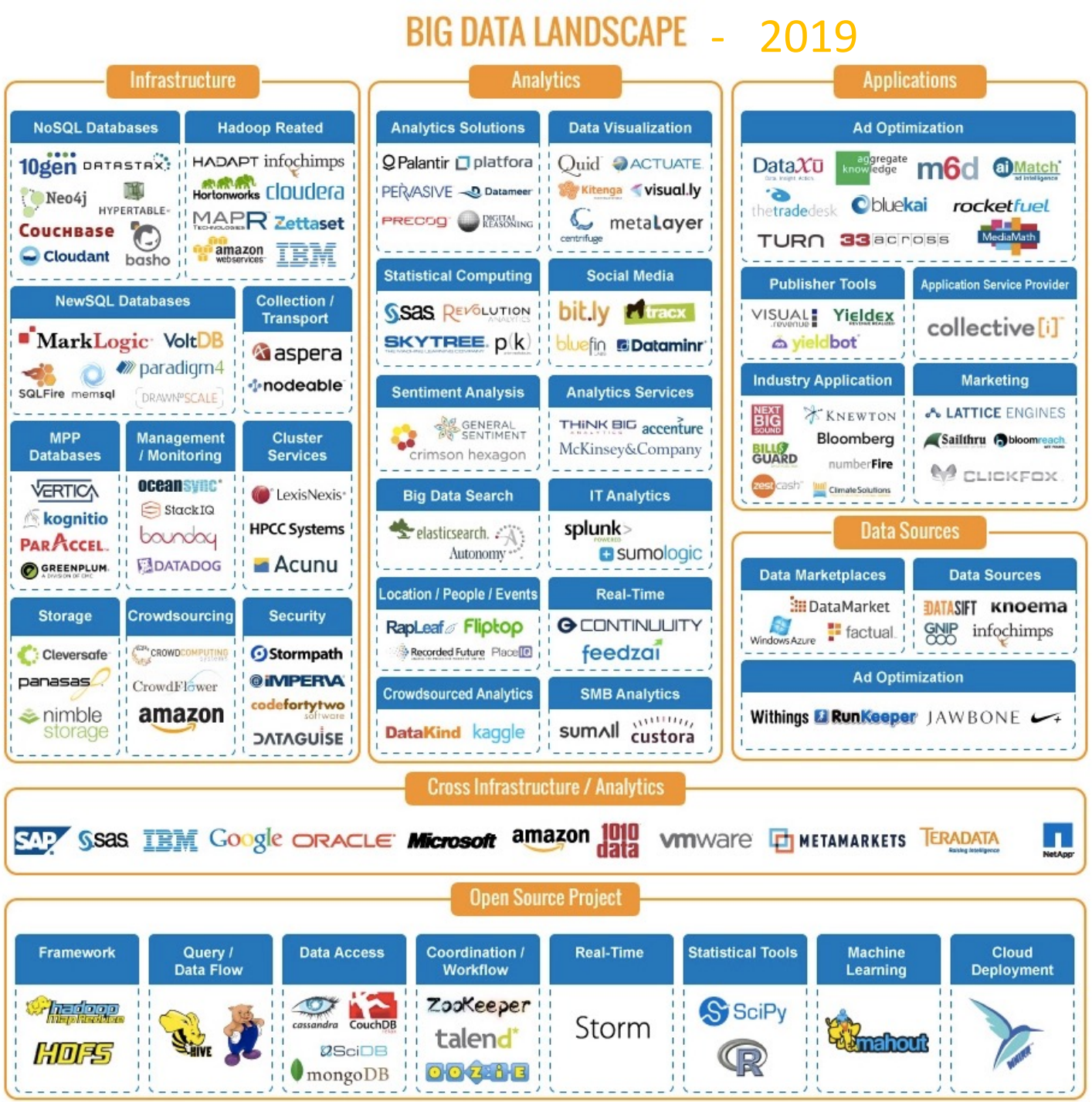
Big Data Landscape - 2012



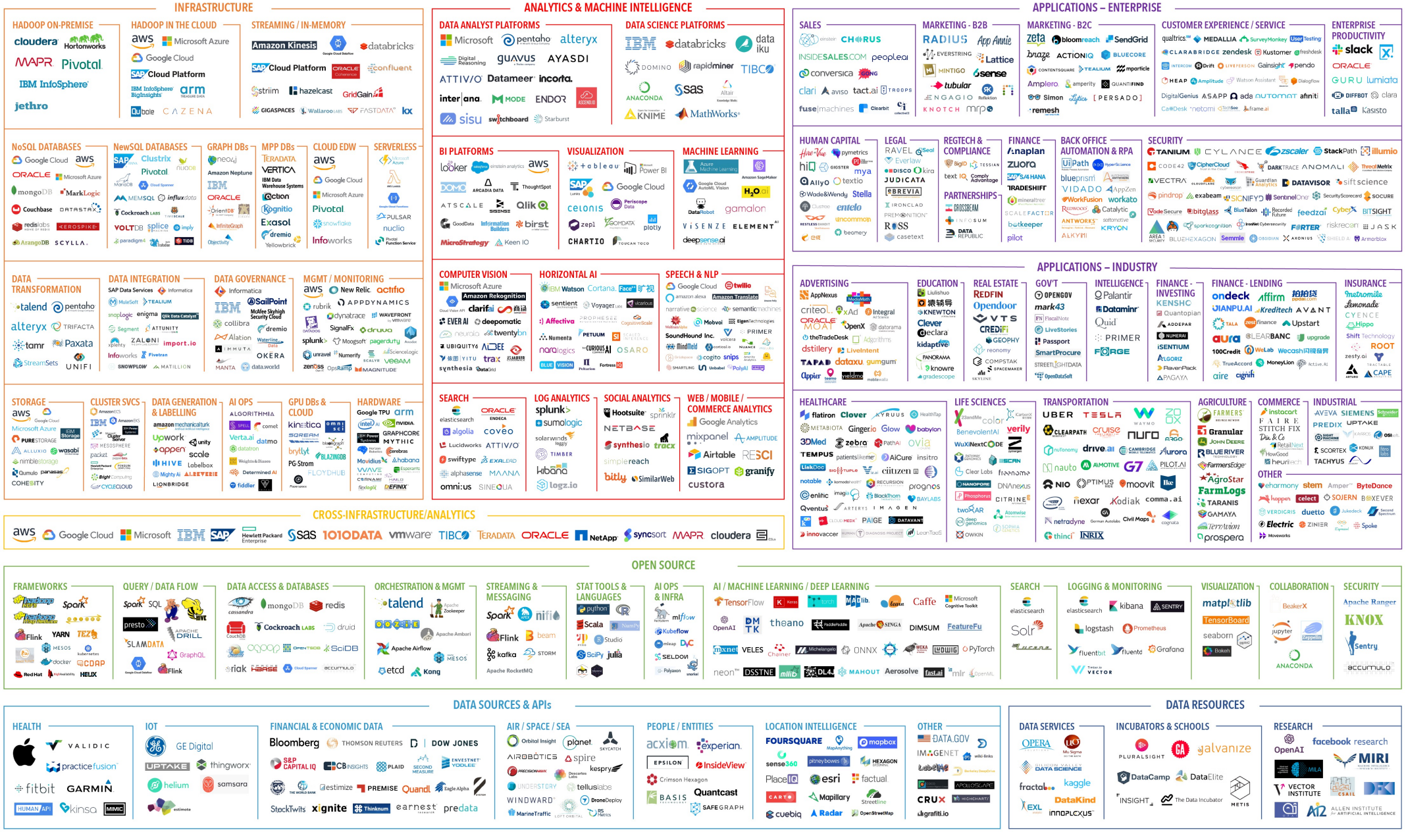
Copyright © 2012 Dave Feinleib

dave@vcdave.com

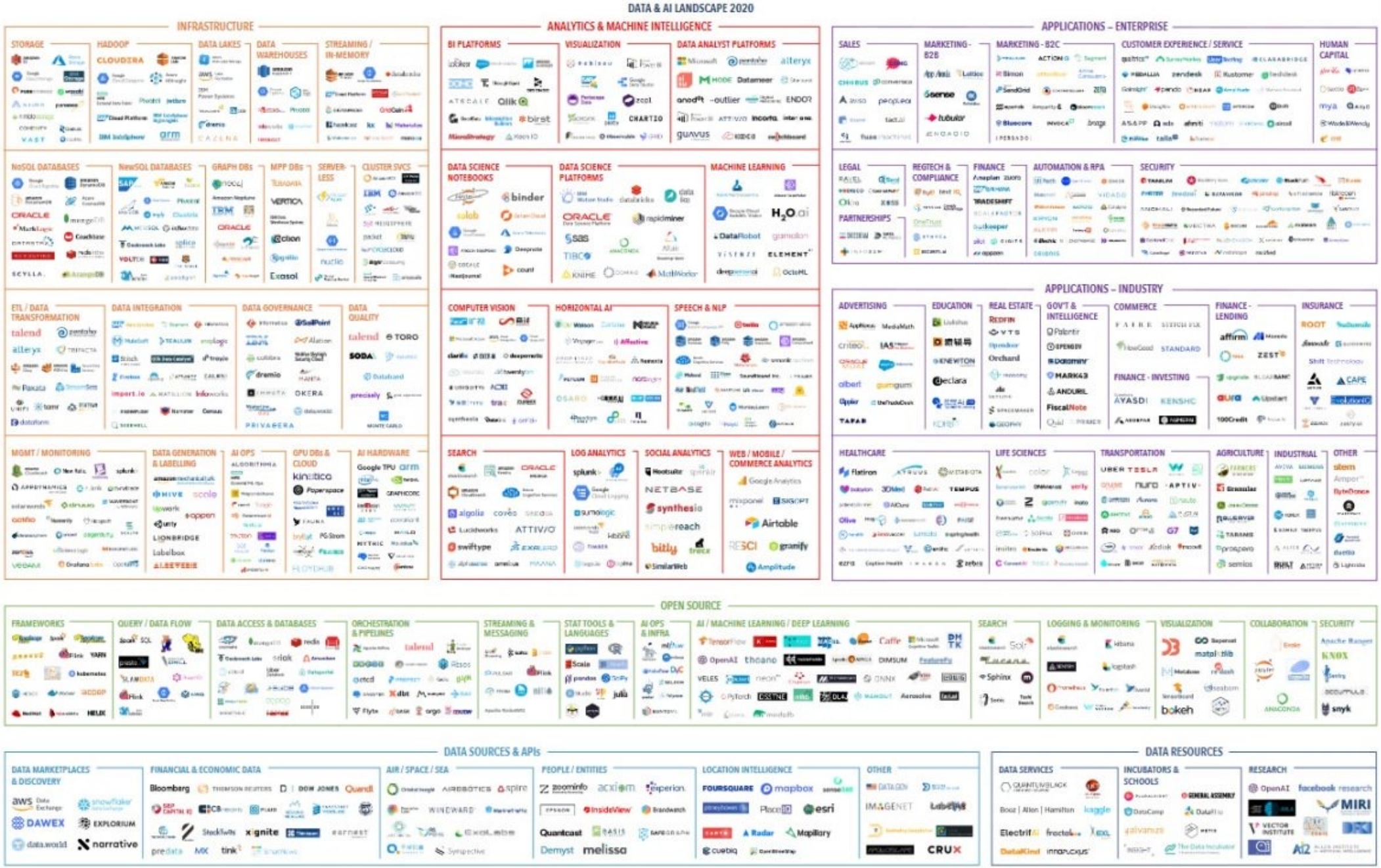
blogs.forbes.com/davefeinleib


BIG DATA LANDSCAPE - 2019
Infrastructure
NoSQL Databases
10gen
DATASTAX
Neo4j
HYPERTABLE
CouchBase
Cloudant
basho
Hadoop Reated
HADAPT
infochimps
Hortonworks
cloudera
MAPR
Zettaset
amazon webservices
IBM
NewSQL Databases
MarkLogic
VoltDB
paradigm4
SQLFire
memsql
DRAWNSCALE
Collection / Transport
aspera
nodeable
MPP Databases
VERTICA
kognitio
PARACCEL
GREENPLUM
Management / Monitoring
oceansync
StackIQ
boundary
DATADOG
Cluster Services
LexisNexis
HPCC Systems
Acunu
Storage
Cleversafe
panasas
nimble storage
Crowdsourcing
CROWDCOMPUTING
CrowdFlower
amazon
Security
Stormpath
IMPERVA
codefortytwo
DATAGUISE
Analytics
Analytics Solutions
Palantir
platfora
PERVASIVE
Datameer
PRECOG
DIGITAL REASONING
Data Visualization
Quid
ACTUATE
Kitenga
visual.ly
centrifuge
metaLayer
Statistical Computing
SAS
REVOLUTION ANALYTICS
SKYTREE
Social Media
bit.ly
tracx
bluefin
Dataminr
Sentiment Analysis
GENERAL SENTIMENT
crimson hexagon
Analytics Services
THINK BIG
accenture
McKinsey&Company
Big Data Search
elasticsearch.
Autonomy
IT Analytics
splunk>
sumologic
Location / People / Events
RapLeaf
Fliptop
Recorded Future
PlaceIQ
Real-Time
CONTINUITY
feedzai
Crowdsourced Analytics
DataKind
kaggle
SMB Analytics
sumAll
custora
Applications
Ad Optimization
DataXu
aggregate knowledge
m6d
aiMatch
thetradedesk
bluekai
rocketfuel
TURN
33across
MediaMath
Publisher Tools
VISUAL REVENUE
Yieldex
yieldbot
Application Service Provider
collective[i]
Industry Application
NEXT BIG SOUND
KNEWTON
Bloomberg
BILLGUARD
numberFire
zestcash
Climate Solutions
Marketing
LATTICE ENGINES
Sailthru
bloomreach
CLICKFOX
Data Sources
Data Marketplaces
DataMarket
Windows Azure
factual
Data Sources
DATASIFT
knoema
GNIP
infochimps
Ad Optimization
Withings
RunKeeper
JAWBONE
Cross Infrastructure / Analytics
SAP
SAS
IBM
Google
ORACLE
Microsoft
amazon
1010data
vmware
METAMARKETS
TERADATA
NetApp
Open Source Project
Framework
hadoop implementation
HDFS
Query / Data Flow
HIVE
Data Access
cassandra
CouchDB
SciDB
mongoDB
Coordination / Workflow
ZooKeeper
talend
OOZIE
Real-Time
Storm
Statistical Tools
SciPy
R
Machine Learning
mahout
Cloud Deployment

DATA & AI LANDSCAPE 2019



© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)


DATA & AI LANDSCAPE 2020
INFRASTRUCTURE
STORAGE
HADOOP
DATA LAKES
DATA WAREHOUSES
STREAMING / IN-MEMORY
NoSQL DATABASES
NewSQL DATABASES
GRAPH DBs
MPP DBs
SERVERLESS
CLUSTER SVCS
ETL / DATA TRANSFORMATION
DATA INTEGRATION
DATA GOVERNANCE
DATA QUALITY
MGMT / MONITORING
DATA GENERATION & LABELLING
AI OPS
GPU DBs & CLOUD
AI HARDWARE
ANALYTICS & MACHINE INTELLIGENCE
BI PLATFORMS
VISUALIZATION
DATA ANALYST PLATFORMS
DATA SCIENCE NOTEBOOKS
DATA SCIENCE PLATFORMS
MACHINE LEARNING
COMPUTER VISION
HORIZONTAL AI
SPEECH & NLP
SEARCH
LOG ANALYTICS
SOCIAL ANALYTICS
WEB / MOBILE / COMMERCE ANALYTICS
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER EXPERIENCE / SERVICE
HUMAN CAPITAL
LEGAL
REGTECH & COMPLIANCE
FINANCE
AUTOMATION & RPA
SECURITY
PARTNERSHIPS
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
REAL ESTATE
GOVT & INTELLIGENCE
COMMERCE
FINANCE - LENDING
INSURANCE
FINANCE - INVESTING
HEALTHCARE
LIFE SCIENCES
TRANSPORTATION
AGRICULTURE
INDUSTRIAL
OTHER
OPEN SOURCE
FRAMEWORKS
QUERY / DATA FLOW
DATA ACCESS & DATABASES
ORCHESTRATION & PIPELINES
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
AI OPS & INFRA
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOGGING & MONITORING
VISUALIZATION
COLLABORATION
SECURITY
DATA SOURCES & APIs
DATA MARKETPLACES & DISCOVERY
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
DATA SERVICES
INCUBATORS & SCHOOLS
RESEARCH
Version 1.0 - September 2020
© Matt Turck (@mattturck) & FirstMark (@firstmarkcap)
mattturck.com/data2020
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Big Data Glossary

Concepts as we will use throughout the course so, it's helpful to have specialized terminology available in a single place:

- **Big data**: Big data is an umbrella term for datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity, and variety. This term is also typically applied to technologies and strategies to work with this type of data.
- **Batch processing**: Batch processing is a computing strategy that involves processing data in large sets. This is typically ideal for non-time sensitive work that operates on very large sets of data. The process is started and at a later time, the results are returned by the system.
- **Cluster computing**: Clustered computing is the practice of pooling the resources of multiple machines and managing their collective capabilities to complete tasks. Computer clusters require a cluster management layer which handles communication between the individual nodes and coordinates work assignment.
- **Data lake**: Data lake is a term for a large repository of collected data in a relatively raw state. This is frequently used to refer to the data collected in a big data system which might be unstructured and frequently changing. This differs in spirit to data warehouses (defined below).
- **Data mining**: Data mining is a broad term for the practice of trying to find patterns in large sets of data. It is the process of trying to refine a mass of data into a more understandable and cohesive set of information.
- **Data warehouse**: Data warehouses are large, ordered repositories of data that can be used for analysis and reporting. In contrast to a *data lake*, a data warehouse is composed of data that has been cleaned, integrated with other sources, and is generally well-ordered. Data warehouses are often spoken about in relation to big data, but typically are components of more conventional systems.

Big Data Glossary

- **ETL**: ETL stands for extract, transform, and load. It refers to the process of taking raw data and preparing it for the system's use. This is traditionally a process associated with data warehouses, but characteristics of this process are also found in the ingestion pipelines of big data systems.
- **Hadoop**: Hadoop is an Apache project that was the early open-source success in big data. It consists of a distributed filesystem called HDFS, with a cluster management and resource scheduler on top called YARN (Yet Another Resource Negotiator). Batch processing capabilities are provided by the MapReduce computation engine. Other computational and analysis systems can be run alongside MapReduce in modern Hadoop deployments.
- **In-memory computing**: In-memory computing is a strategy that involves moving the working datasets entirely within a cluster's collective memory. Intermediate calculations are not written to disk and are instead held in memory. This gives in-memory computing systems like Apache Spark a huge advantage in speed over I/O bound systems like Hadoop's MapReduce.
- **Machine learning**: Machine learning is the study and practice of designing systems that can learn, adjust, and improve based on the data fed to them. This typically involves implementation of predictive and statistical algorithms that can continually zero in on "correct" behavior and insights as more data flows through the system.
- **Map reduce (big data algorithm)**: Map reduce (the big data algorithm, not Hadoop's MapReduce computation engine) is an algorithm for scheduling work on a computing cluster. The process involves splitting the problem set up (mapping it to different nodes) and computing over them to produce intermediate results, shuffling the results to align like sets, and then reducing the results by outputting a single value for each set.
- **NoSQL**: NoSQL is a broad term referring to databases designed outside of the traditional relational model. NoSQL databases have different trade-offs compared to relational databases, but are often well-suited for big data systems due to their flexibility and frequent distributed-first architecture.
- **Stream processing**: Stream processing is the practice of computing over individual data items as they move through a system. This allows for real-time analysis of the data being fed to the system and is useful for time-sensitive operations using high velocity metrics

Running Hadoop

Installing Hadoop in your machine

Hadoop Installation

- Option 1:
- https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html

- Option 2:
- Install virtual box and HDP
- https://hortonworks.com/wp-content/uploads/2016/02/Import_on_Vbox_3_1_2016.pdf

Option 1: Requirements

Necessary

Java >= 1.7

ssh

Linux OS (Ubuntu >= 14.04)
or

Windows

or

MacOS

Hadoop framework

Optional

Eclipse or

Other IDE

Hadoop Platforms

- Platforms: Unix and on Windows.
  - Linux: the only supported production platform.
  - Other variants of Unix, like Mac OS X: run Hadoop for development.
  - Windows + Cygwin: development platform (openssh)
- Java 7
  - Java >= 1.7.x (aka 7.0.x aka 7) is recommended for running Hadoop.

Java PATH Setup

- Set JAVA path
- Open the *.bashrc* file located in home directory
- gedit ~/.bashrc or vi ~/.bashrc
- Add below line at the end:
- Linux
  - export JAVA_HOME=/usr/lib/jvm/java-7-openjdk
- MacOS
  - export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_261.jdk/Contents/Home

.bash_profile

```
export PYSPARK_PYTHON=/usr/local/bin/python3
export PYSPARK_DRIVER_PYTHON=/usr/local/bin/python3/ipython

export PYTHONPATH=/usr/local/bin/python3

export
JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_261.jdk/Contents/Home

export SCALA_HOME=/Users/usr/work/scala-2.11.12
export PATH=$SCALA_HOME/bin:$PATH

export SPARK_HOME=/Users/usr/work/spark-2.4.7-bin-hadoop2.7
export PATH=$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH

export HADOOP_HOME=/Users/usr/work/hadoop-2.7.2
export PATH=$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH$
```

Hadoop Installation

- Download a stable version of Hadoop:
  - https://hadoop.apache.org/releases.html
- Untar the hadoop file:
  - Extract the contents of the Hadoop package to a location of your choice. Here, example shows: **/Usr/local/hadoop**.
  - $ cd /usr/local
  - $ sudo tar xvfz hadoop-2.7.2.tar.gz
  - $ sudo mv hadoop-2.7.2 hadoop
- JAVA_HOME at hadoop/conf/hadoop-env.sh:
  - Mac OS: /System/Library/Frameworks/JavaVM.framework/Versions/1.6.0/Home (/Library/Java/Home)
  - Linux: which java
- Environment Variables:
  - export PATH=$PATH:$HADOOP_HOME/bin

Installation & Configuration of SSH

- Hadoop requires SSH(Secure Shell) access to manage its nodes, i.e. remote machines plus your local machine if you want to use Hadoop on it.
- Install SSH using following command
- sudo apt-get install ssh
- First, we have to generate DSA an SSH key for user.

- ssh-keygen -t dsa -P'' -f ~/.ssh/id dsa
- cat ~/.ssh/id dsa.pub >>~/.ssh/authorized keys

Files to Configure

- The following are the files we need to configure
  - core-site.xml
  - hadoop-env.sh
  - mapred-site.xml
  - hdfs-site.xml

Formatting the HDFS filesystem via the NameNode

- The first step to starting up your Hadoop installation is
- Formatting the Hadoop file system
- We need to do this the first time you set up a Hadoop.
- Do not format a running Hadoop filesystem as you will lose all the data currently in HDFS
- To format the filesystem, run the command
- hadoop namenode -format

Starting single-node cluster

- Run these two commands:
- *hadoop1@hadoop1:/usr/local/hadoop$* start-all.sh
- or
- *hadoop1@hadoop1:/usr/local/hadoop$* start-dfs.sh && start-yarn.sh

This will startup a NameNode,SecondaryNameNode, DataNode, ResourceManager and a NodeManager on your machine.

Stopping your single-node cluster

Run the command to stop hadoop

hadoop1@hadoop1:/usr/local/hadoop$ `stop-all.sh or stop-dfs.sh and stop-yarn.sh`

To stop all the daemons running on your machine output will be like this.

stopping NodeManager

localhost: stopping ResourceManager

stopping NameNode

localhost: stopping DataNode

localhost: stopping SecondaryNameNode

Hadoop Installation Resources – Option 1

- For installing Hadoop on Mac, you can follow the instructions at this link:
  - (Preferred) Install Hadoop 3.3.0 on macOS - Kontext
  - or this link: Installing Hadoop on Mac. Pre-requisites | by Diwakar | Beer&Diapers.ai | Medium
- For Windows, the tutorial can be found at this link:
  - (Preferred) Install Hadoop 3.3.0 on Windows 10 Step by Step Guide - Kontext
  - or this link: How to install Hadoop 3.3.0 on Windows 10 | Easy step by step tutorial (Latest Version as of 2020) - YouTube

Option 2:

- Download virtual box:
- https://www.oracle.com/virtualization/technologies/vm/downloads/virtualbox-downloads.html

- Download HDP:
- https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html

References

- http://hadoop.apache.org/common/docs/r0.20.2/quickstart.html
- https://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
- https://hadoop.apache.org/docs/r1.2.1/single_node_setup.pdf
- https://medium.com/beeranddiapers/installing-hadoop-on-mac-a9a3649dbc4d
- https://cwiki.apache.org/confluence/display/HADOOP/Hadoop+Java+Versions