

02

Introduction to Big Data

Kalyani Selvarajah
School of Computer Science
University of Windsor



Advanced Database Topics
COMP 8157-01/02/03
Fall 2023

Today's Agenda

Evolution of Data

Introduction to Big data

Big Data Analytics (BDA)

Introduction to Hadoop



Announcement

- Project Proposal Submission:
 - September 24th/25th/26th (choose the right date as per the section)



Introductory Questions

Why the amount of data increasing tremendously?

Why RDBMS is not suitable for Big Data?

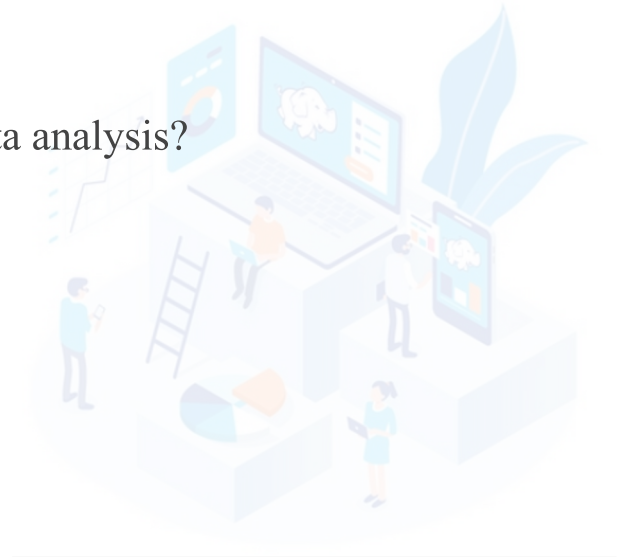
What are the information collected through web logs for Big Data analysis?

Define the term Big Data?

Why Big Data Analysis is helpful?

What are the issues with traditional ETL in Big Data?

What is the solution for Big Data Analysis ?



Big Data: Word Cloud

Join: <https://www.menti.com/>



Evolution of Data



Source: <https://www.shutterstock.com/search/3d+iot>

IOT devices

Social medias

Ecommerce websites

Banking & Finance

Media & Entertainment

Healthcare

Transportation

...

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devices - are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

3.9bn

people use emails

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

4TB

of data produced by a connected tent

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

PwC

2013

44ZB

2020

DEMYSIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
b bit	0 or 1	1/8 of a byte
B byte	8 bits	1 byte
KB kilobyte	1,000 bytes	1,000 bytes
MB megabyte	1,000 ² bytes	1,000,000 bytes
GB gigabyte	1,000 ³ bytes	1,000,000,000 bytes
TB terabyte	1,000 ⁴ bytes	1,000,000,000,000 bytes
PB petabyte	1,000 ⁵ bytes	1,000,000,000,000,000 bytes
EB exabyte	1,000 ⁶ bytes	1,000,000,000,000,000,000 bytes
ZB zettabyte	1,000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB yottabyte	1,000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

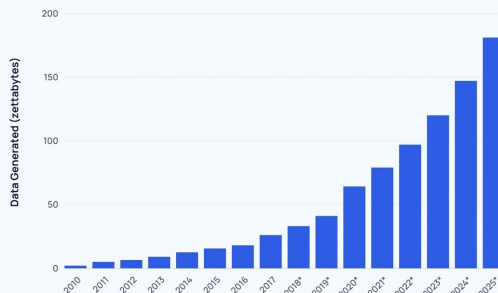
*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

Global Data Generated Annually



463EB

of data will be created every day by 2025

idc

95m

photos and videos are shared on Instagram

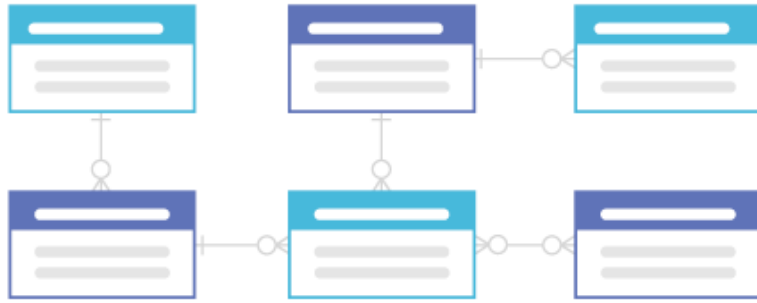
Instagram Business

28PB

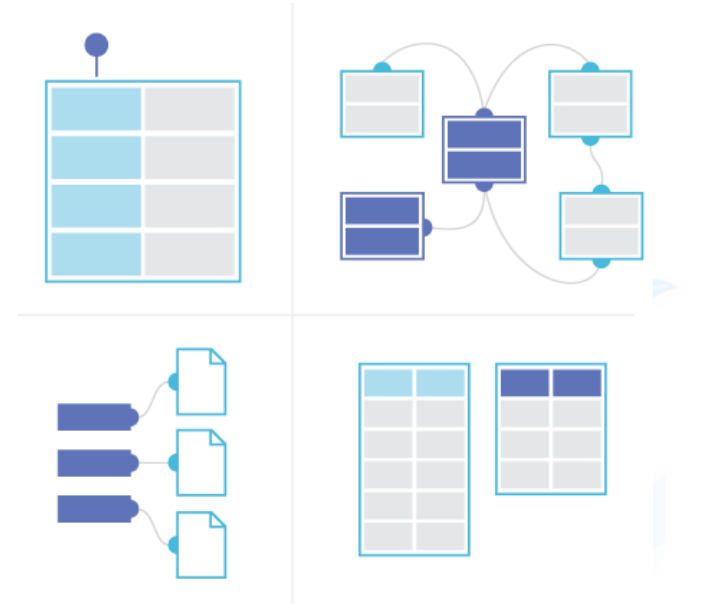
to be generated from wearable devices by 2020

Statista

Relational Data to Big Data



RDBMS



Non-Relational Database

Drawback with RDBMS for Large dataset

Scalability & Complexity: Processing large data may fail

Can't store unstructured data

High cost.

Important Information in the Web Logs

What did you view

When did you view

What did you purchase

Your shipping address, phone number etc.

Your review information


How many days did it take to write your review?

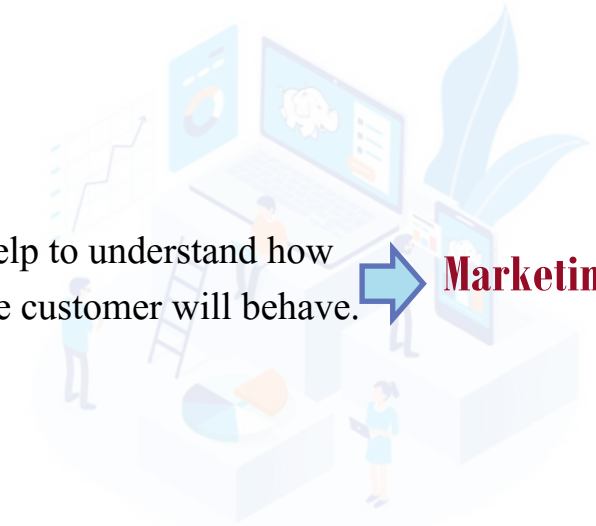
How are you likely to click a link in the Amazon email given its content?

Prime member information

If you are selling, your selling history information and buyer review information to rank you in a product search.

Your website access information for user browsing pattern.

Help to understand how the customer will behave.  **Marketing**



Debates of Big Data Implication

PROS

- Many advocates declare Big Data to be a new rock star.
- It will become the new epistemologies in science.
- Big Data would revolutionize our way of thinking, working, and living.
- Big Data will be a source of new economic value and innovation.
- Data can speak for itself, and we should let the data speak.
- Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities and perform research.



Debates of Big Data Implication

PROS

- Many advocates declare Big Data to be a new rock star.
- It will become the new epistemologies in science.
- Big Data would revolutionize our way of thinking, working, and living.
- Big Data will be a source of new economic value and innovation.
- Data can speak for itself, and we should let the data speak.
- Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities and perform research.

- Big Data is inconclusive, overstated, exaggerated, and misinformed by the media.
- Data cannot speak for itself.
- Never judge a decision by its outcome —bias.
- Extraordinary Popular Delusion and the Madness of Crowd.
- The hype overtaken reality and there was little time to think about.
- Big Data were “literally hard” and “expensive”.
- The size of data should fit the research question being asked; in some cases, small is best.

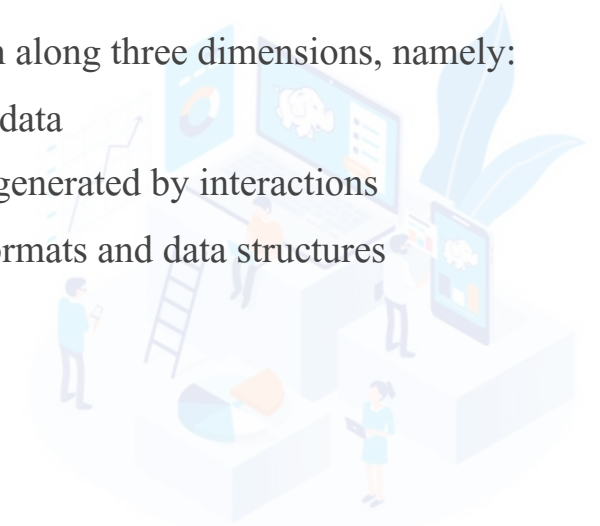
CONS

Historical Interpretation Of Big Data

Douglas Laney's 3Vs:

He noticed that due to surging of e-commerce activities, data has grown along three dimensions, namely:

1. Volume: means the incoming data stream and cumulative volume of data
2. Velocity: represents the pace of data used to support interaction and generated by interactions
3. Variety: signifies the variety of incompatible and inconsistent data formats and data structures



Historical Interpretation Of Big Data

IBM — 4Vs:

IBM added another attribute or “V” for “Veracity” on the top of Douglas Laney’s 3Vs notation:

1. Volume stands for the scale of data
2. Velocity denotes the analysis of streaming data
3. Variety indicates different forms of data
4. **Veracity** implies the uncertainty of data

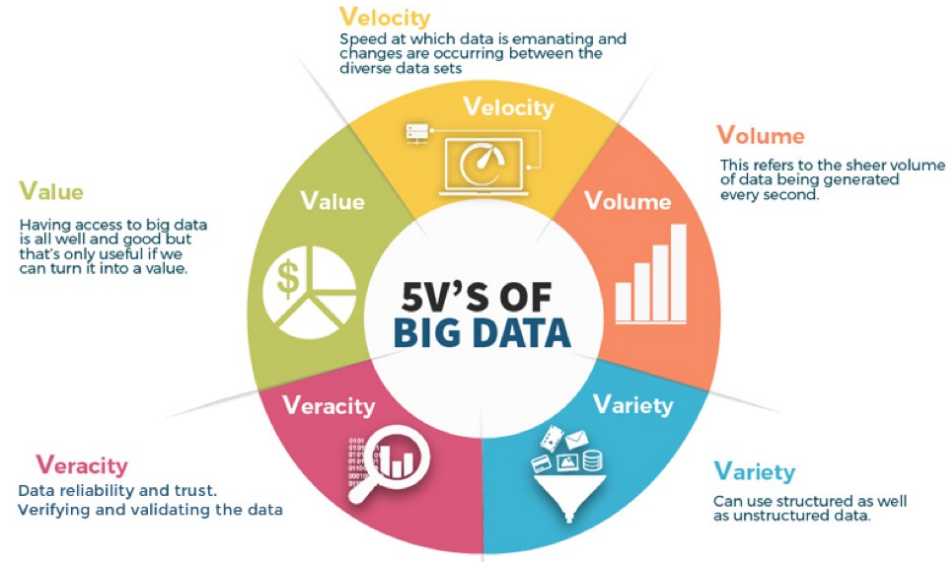


Historical Interpretation Of Big Data

Yuri Demchenko's 5Vs:

Yuri added the value dimension along with the IBM 4Vs' definition in 2013:

1. Volume stands for the scale of data
2. Velocity denotes the analysis of streaming data
3. Variety indicates different forms of data
4. Veracity implies the uncertainty of data
5. **Value** refers to how useful the data is in decision making.

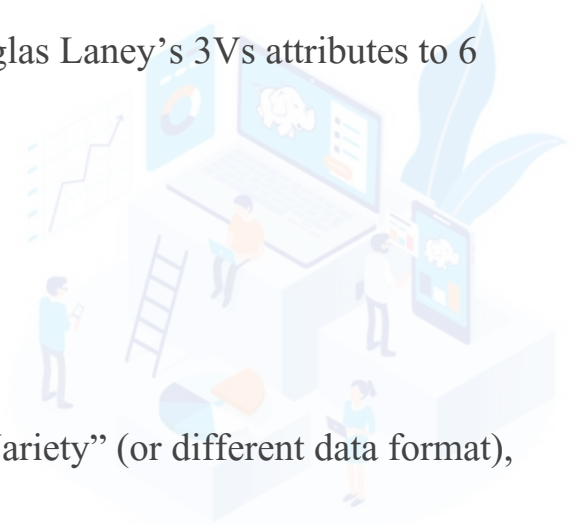


Historical Interpretation Of Big Data

Microsoft — 6Vs:

For the sake of maximizing the business value, Microsoft extended Douglas Laney's 3Vs attributes to 6 Vs:

1. Volume stands for scale of data
2. Velocity denotes the analysis of streaming data
3. Variety indicates different forms of data
4. Veracity focuses on trustworthiness of data sources
5. Variability refers to the complexity of data set. In comparison with “Variety” (or different data format), it means the number of variables in data sets
6. **Visibility** emphasizes that you need to have a full picture of data in order to make informative decision

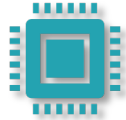


7 Types Definitions of Big Data



The original big data (3Vs)

Referred to Douglas Laney's volume, velocity, and variety, or 3Vs. It has been widely cited since 2001. Many have tried to extend the number of Vs, such as 4Vs, 5Vs, 6Vs ... up to 11Vs



Big Data as technology

Oriented by new technology development, such as MapReduce, bulk synchronous parallel (BSP — Hama), resilient distributed datasets (RDD, Spark), and Lambda architecture (Flink)



Big Data as application

Emphasizes different applications based on different types of big data. Barry defined it as application of process-mediated data, human-sourced information, and machine-generated data. Shaun focused on analyzing transactions, interactions, and observation of data.



Big Data as signals

Another type of application-oriented definition, but it focuses on timing rather than the type of data. It looks for a foresight of data or new “signal” pattern in dataset



Big Data as opportunity

“Big data as analyzing data that was previously ignored because of technology limitations.” It highlights many potential opportunities by revisiting the collected or archived datasets when new technologies are variable.



Big Data as metaphor

It defines Big Data as a human thinking process. It elevates BDA to the new level, which means BDS is not a type of analytic tool rather it is an extension of human brain

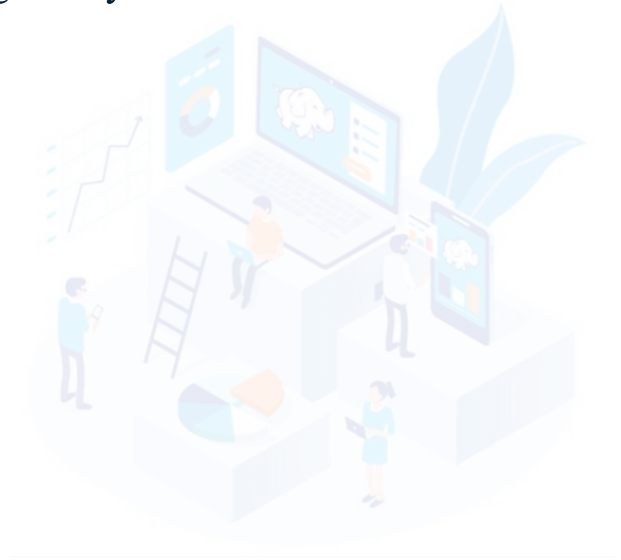


Big Data as new term for old stuff

Means the new bottle (relabel the new term “big data”) for old wine (BI, data mining, or other traditional data analytic activities). It is one of the most cynical ways to define big data.

Discussion:

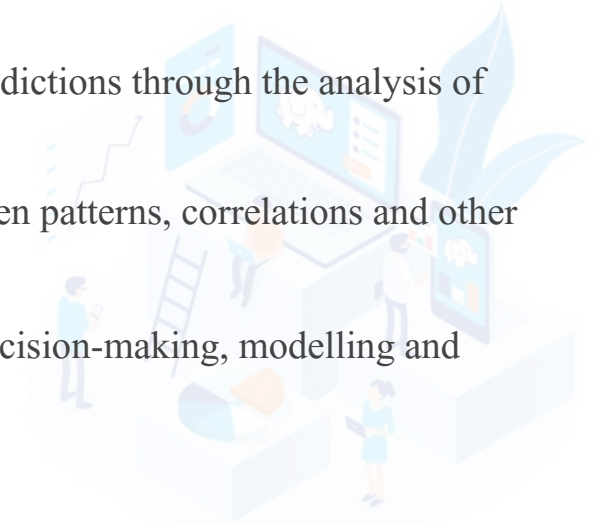
- The Impact of Big Data on Modern Society – positively or negatively



Big Data Analytics (BDA)

Objective of BDA is actually to seek for business intelligence (BI).

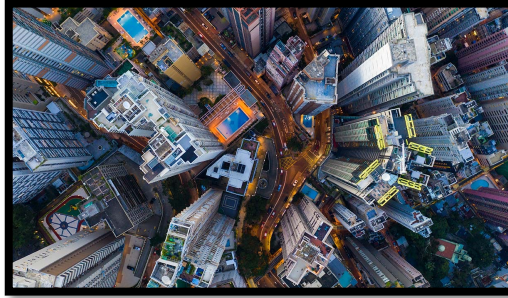
- It enables decision makers to make the right decisions based on predictions through the analysis of available data.
- Big data analytics examines large amounts of data to discover hidden patterns, correlations and other insights.
- With big data analytics, you can ultimately fuel better and faster decision-making, modelling and predicting of future outcomes and enhanced business intelligence.



Examples of big data analytics in industries:



Healthcare big data analytics drive quicker responses to emerging diseases and improve direct patient care, the customer experience, and administrative, insurance and payment processing.



Financial analytics improve customer targeting using customer analytics. Businesses can make better informed underwriting decisions and provide better claims management while mitigating risk and fraud.



Communications service providers (CSPs) can use big data analytics to optimize network monitoring, management and performance to help mitigate risk and reduce costs. They can also use analytics to improve customer targeting and service.

Type of Big Data Analytics (BDA)

1. **Descriptive Analytics:** This type of analytics focuses on summarizing historical data to provide a clear understanding of past events and trends. It helps organizations answer questions like "**WHAT** happened?" and often involves the use of data visualization tools to present data in a meaningful way.
2. **Diagnostic Analytics:** Diagnostic analytics goes a step further by examining historical data to understand **WHY** certain events or patterns occurred. It helps in identifying the root causes of specific outcomes or issues.
3. **Predictive Analytics:** Predictive analytics uses historical data and statistical algorithms to make predictions about future events or trends. It answers questions like "What is likely to happen in the future?" and is valuable for forecasting and risk assessment.
4. **Prescriptive Analytics:** Prescriptive analytics takes the insights from descriptive, diagnostic, and predictive analytics and goes a step further by providing recommendations on what actions to take to optimize outcomes. It helps in answering questions like "What should we do to achieve a desired outcome?"

Big Data Analytics and Machine Learning

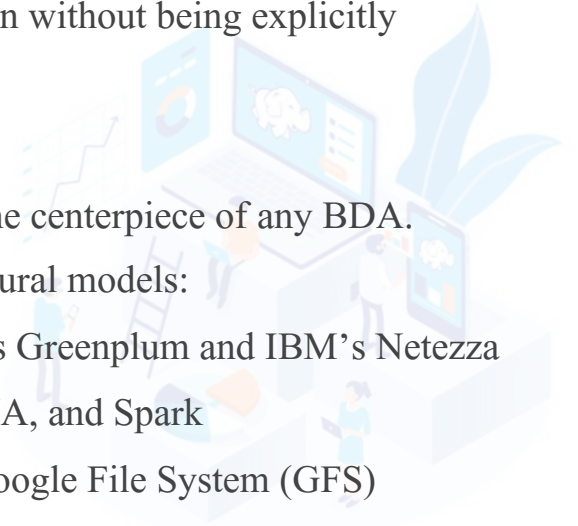
Machine Learning:

“The field of study that gives computers (or machines) that ability to learn without being explicitly programmed”

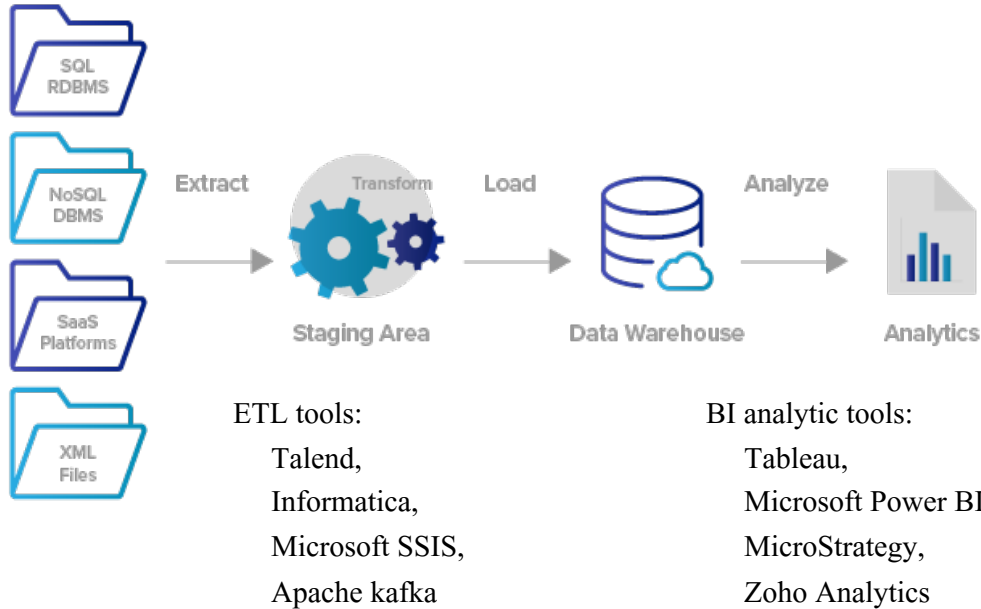
Machine Learning (ML) underpins the BDA implementation → ML is the centerpiece of any BDA.

In terms of computational support to BDA, there are four major architectural models:

1. Massively parallel processing database system: For example, EMC's Greenplum and IBM's Netezza
2. In-memory database systems, such as Oracle Exalytics, SAP's HANA, and Spark
3. MapReduce processing model and platforms such as Hadoop and Google File System (GFS)
4. Bulk Synchronous Parallel (BSP) systems such as Apache HAMA and Giraph



Issues with Traditional ETL in Big Data



Why ETL?

Data is scattered across different locations.
Data is stored in different type of sources.
Volume of data keeps on increasing.
Data can be structured, semi-structured or unstructured.

Issues with Big Data:

Difficult to handle increasing data size by ETL
Doesn't helpful for a real time purpose.
Datawarehouse is very costly

Questions Reflect the Bottom Line Of BI

1. How to store massive data (such as in PB or EB scale currently) or information in the available resources
2. How to access these massive data or information quickly
3. How to work with datasets in variety formats: structured, semi-structured, and unstructured
4. How to process these datasets in a full scalable, fault tolerant, and flexible manner
5. How to extract BI interactively and cost-effectively

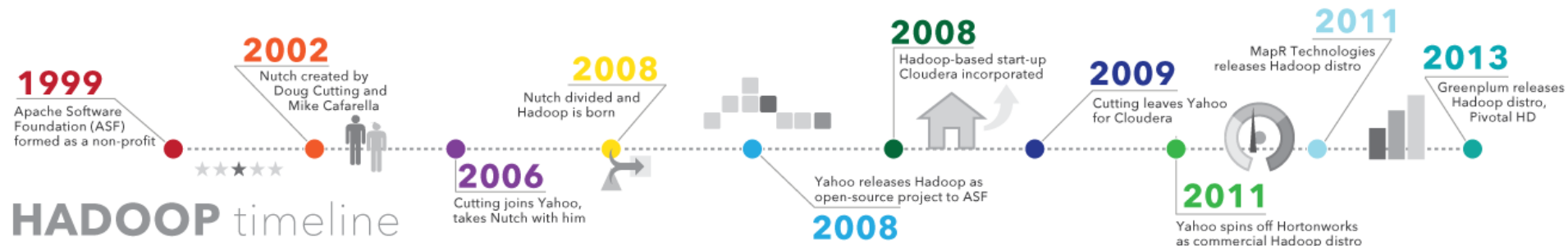




Hadoop

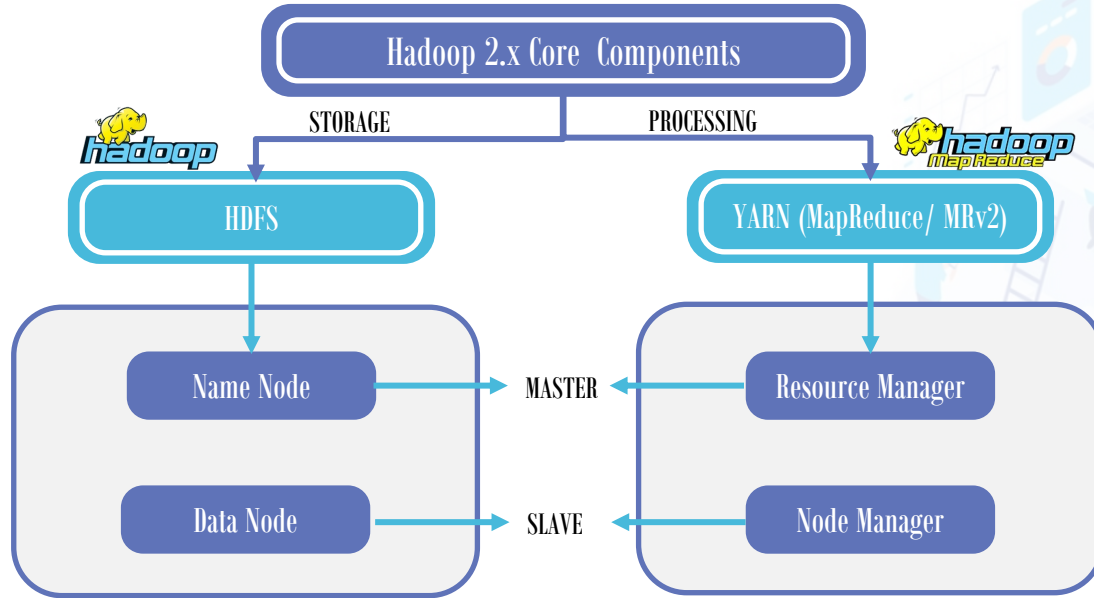
- Hadoop was started by Doug Cutting and Mike Cafarella to support two other well-known projects, Lucene and Nutch
- Hadoop has been inspired by Google's File System (GFS) which was detailed in a paper by released by Google in 2003
- Hadoop, originally called Nutch Distributed File System (ND FS) split from Nutch in 2006 to become a sub-project of Lucene. At this point it was renamed to Hadoop.
- They gave it to Apache as an open-source project. → **Apache Hadoop**
- The commercial distribution for Hadoop:
 - Cloudera
 - Hortonworks
 - MapR
- The public cloud for Hadoop:
 - AWS
 - Windows Azure HDInsight
 - Google Dataproc**

Brief history of Hadoop

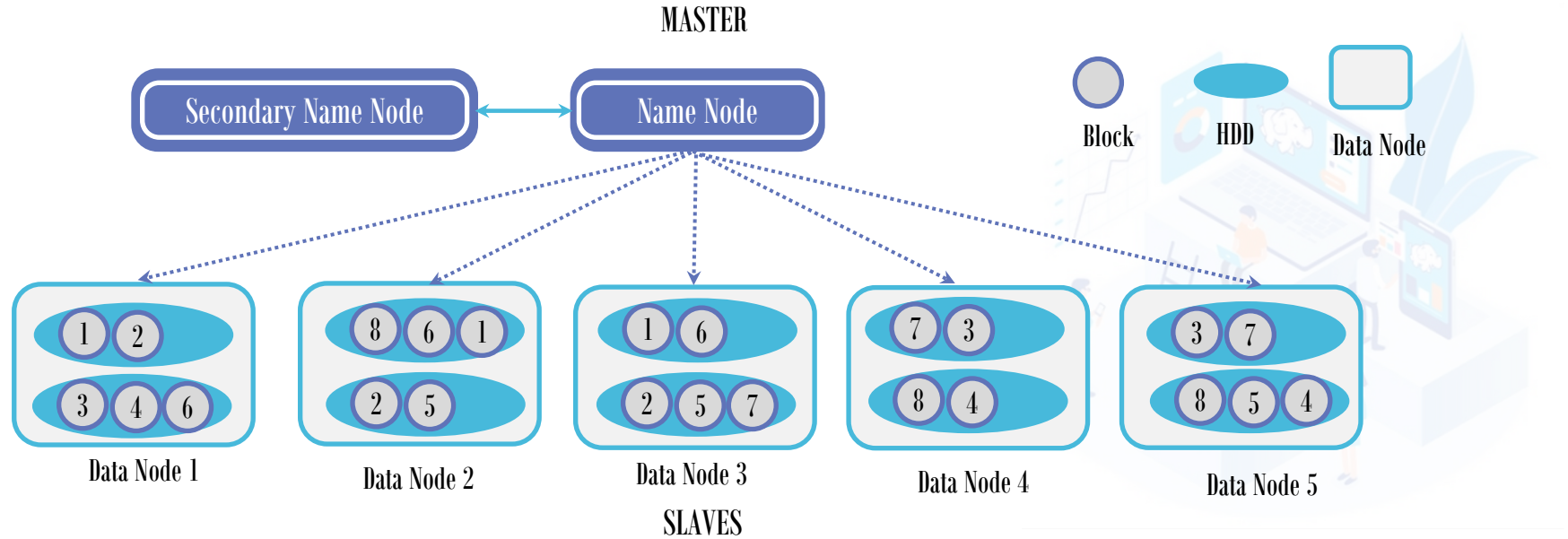


Main components of Hadoop

- HDFS: Hadoop Distributed File structure
- MapReduce



HDFS: Hadoop Distributed File structure

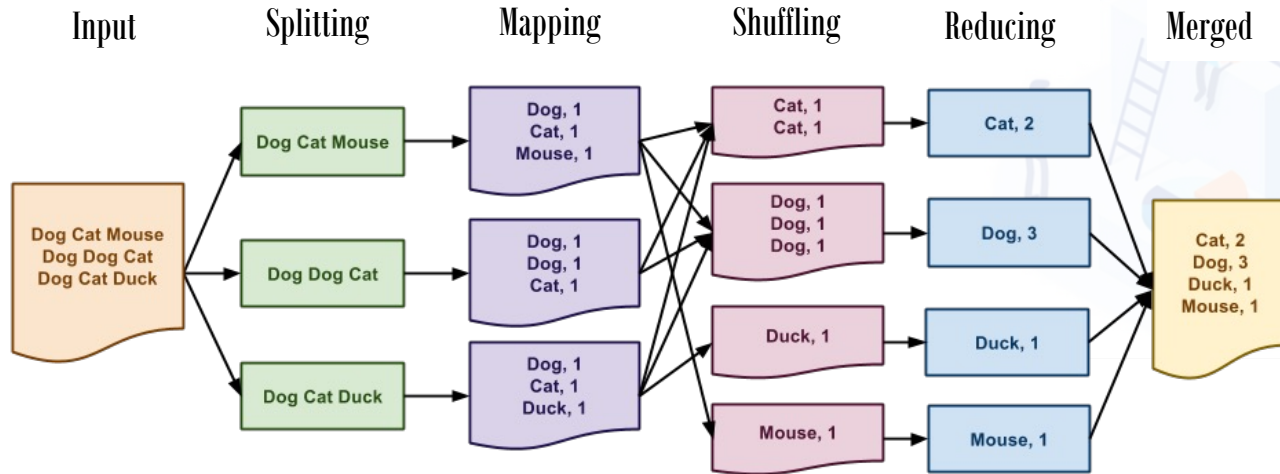


MapReduce

MapReduce is a programming model used to process large dataset workloads.

The basic strategy of MapReduce is to divide and conquer.

A major advantage of MapReduce is its capability of shared-nothing data processing.



SPARK



Spark was developed by the UC Berkeley RAD Lab (now called as AMP Lab).

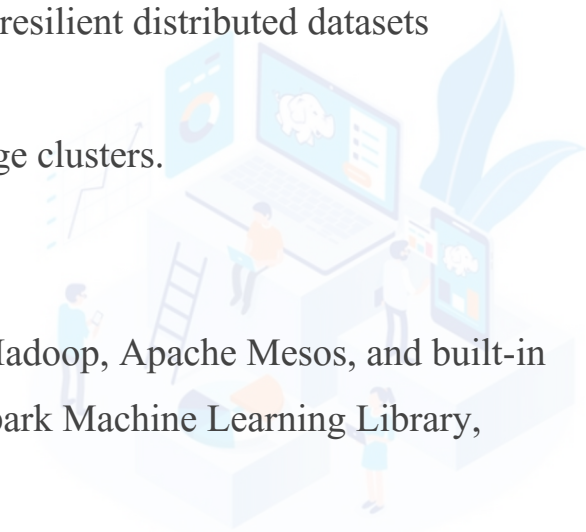
It intends to replace a MapReduce model with a better solution- adopts resilient distributed datasets (RDDs) in memory computation (micro batch) technique.

Spark is a fast- and general-purpose computation platform based on large clusters.

Open source: Apache Spark

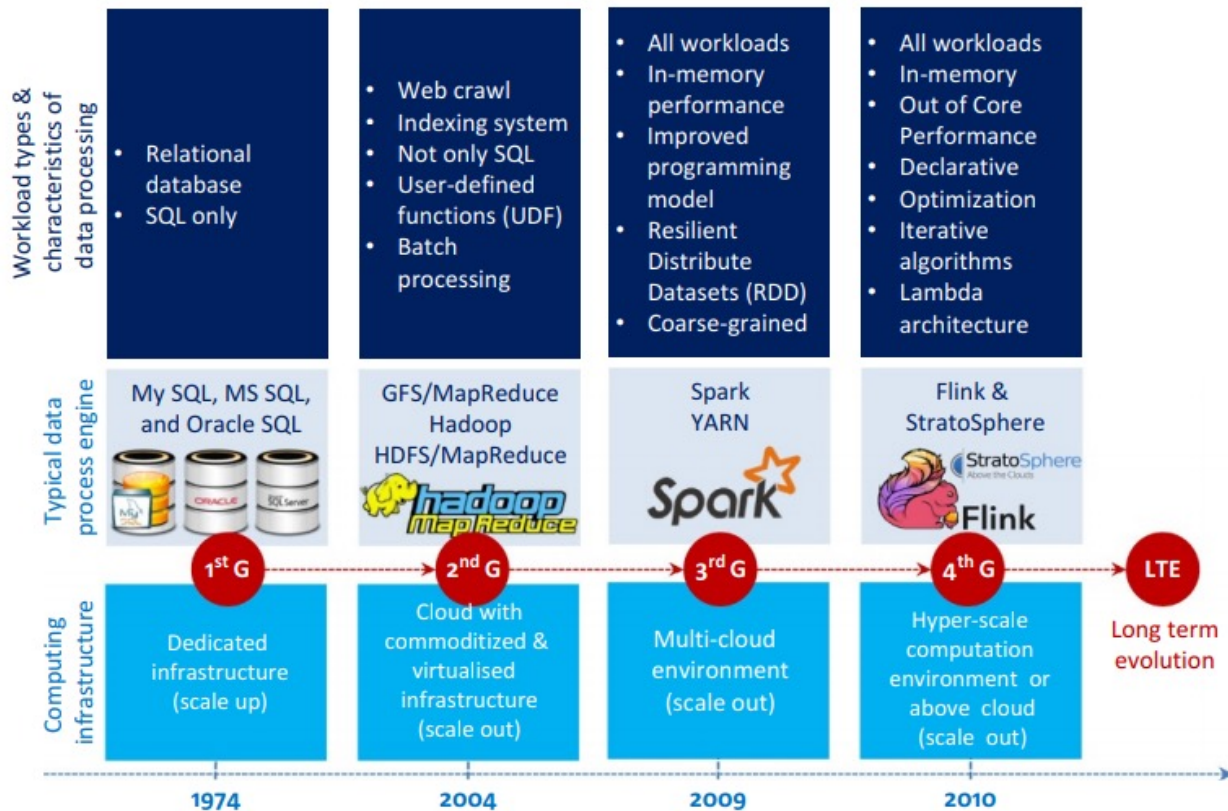
Spark consists of seven major elements:

Spark core of data engine, Spark cluster manager (includes Hadoop, Apache Mesos, and built-in Standalone cluster manger), Spark SQL, Spark streaming, Spark Machine Learning Library, Spark GraphX, and Spark programming tools.

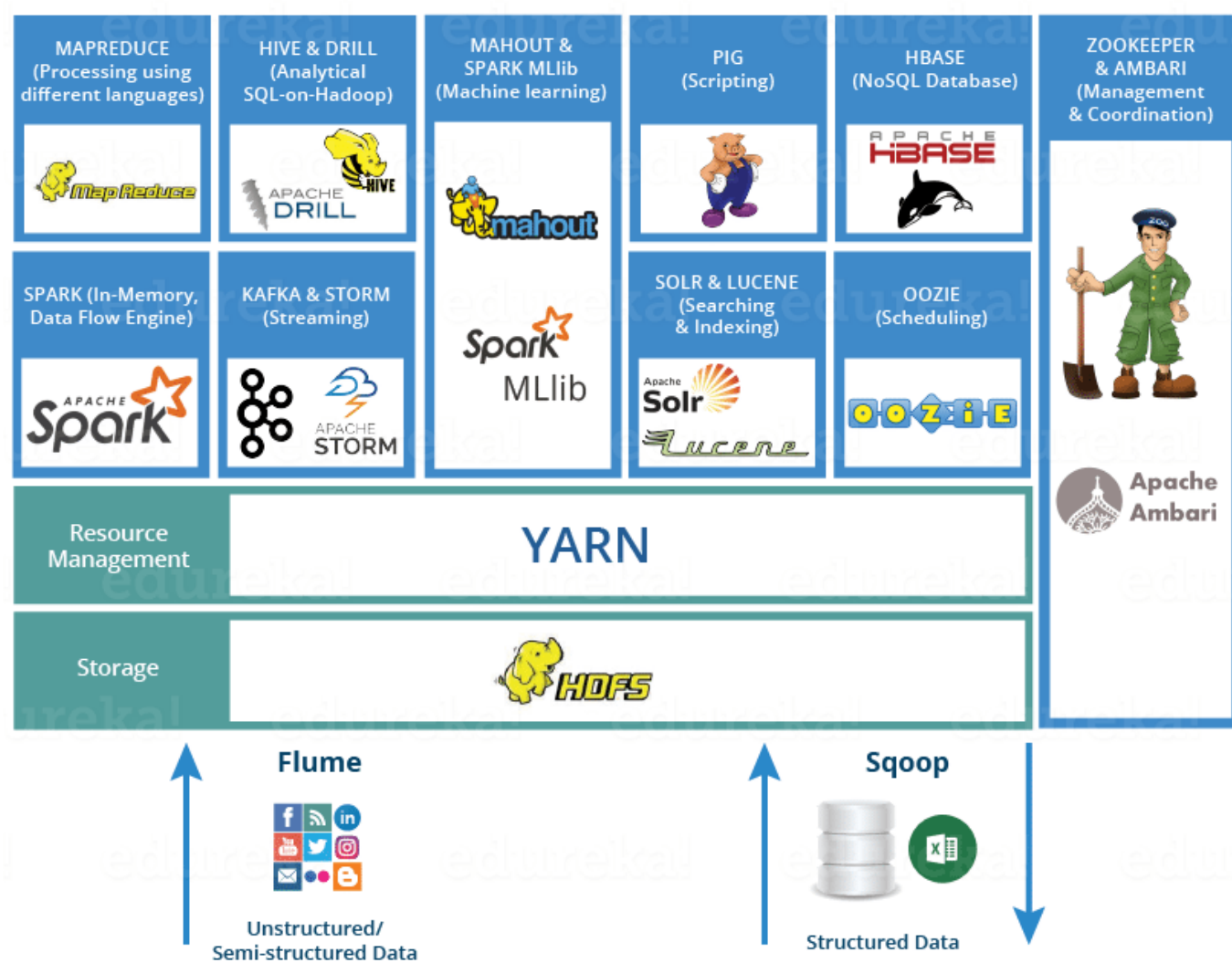


Other Data Process Engines

Microsoft Dryad,
Storm,
Tez,
Flink, and
CIEL



Apache Hadoop Ecosystem



Learning Hadoop through LinkedIn learning

- <https://www.linkedin.com/learning/learning-hadoop-2/getting-started-with-hadoop?u=56973065>



Summary

Evolution of Data

Relational Data to Big Data

Debates of Big Data Implication

Historical Interpretation Of Big Data

Big Data Analytics (BDA)

Big Data Analytics and Machine Learning

Issues with Traditional ETL in Big Data

Hadoop (HDFS & Map Reduce) help to handle Big data efficiently.

Spark is a powerful open-source unified analytics engine.

