Lisbon School of Economics & Management
Universidade de Lisboa

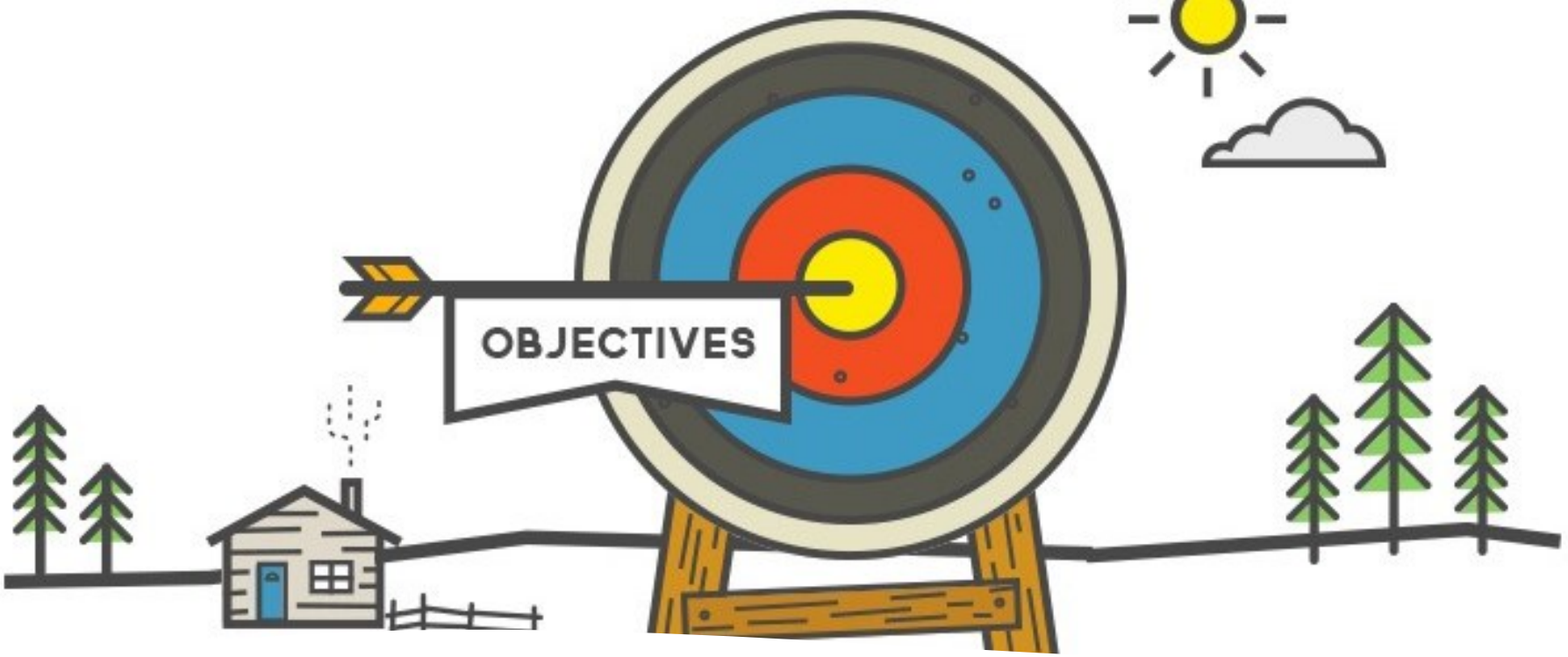
LISBOA
UNIVERSIDADE DE LISBOA

Big Data in Business and Sustainability

Prof. Carlos J. Costa, PhD

Session Overview

- What is Big Data
- The 5 Vs of Big Data
- Importance for Business
- Sustainability Applications
- Technologies: Hadoop, Spark, NoSQL
- Case Studies & Discussion



Learning Goals

- Understand Big Data concepts
- Identify the 5 characteristics (5Vs)
- Recognize business value
- Understand sustainability applications
- Explore key technologies used in Big Data

What is Big Data?

- Big Data refers to extremely large datasets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and business operations.

Why Big Data Matters

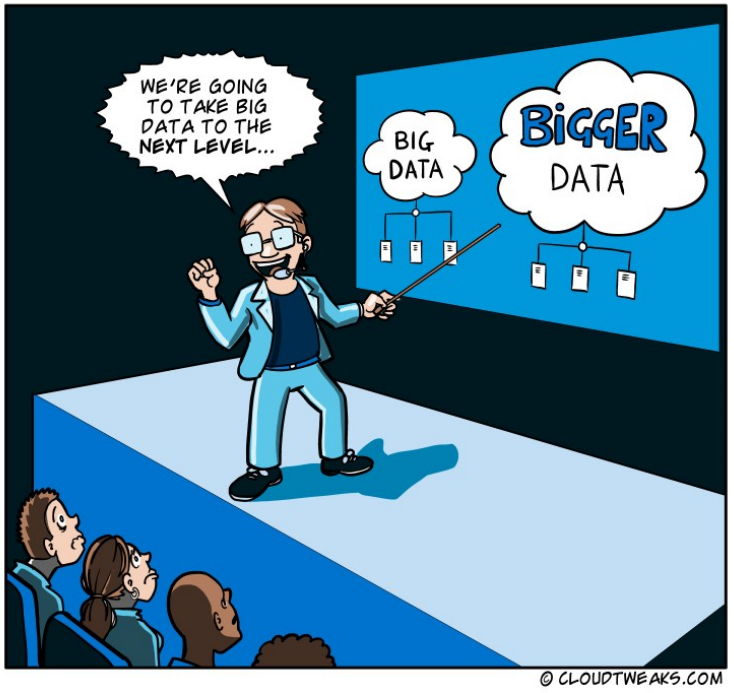
- Massive data generation
- Better decision making
- Competitive advantage
- Predictive insights
- Automation and efficiency

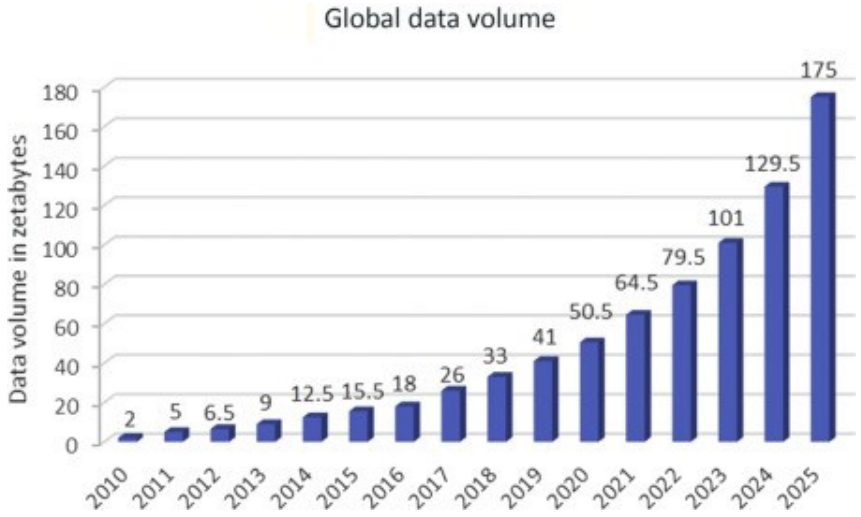
The 5 Vs of Big Data

Volume

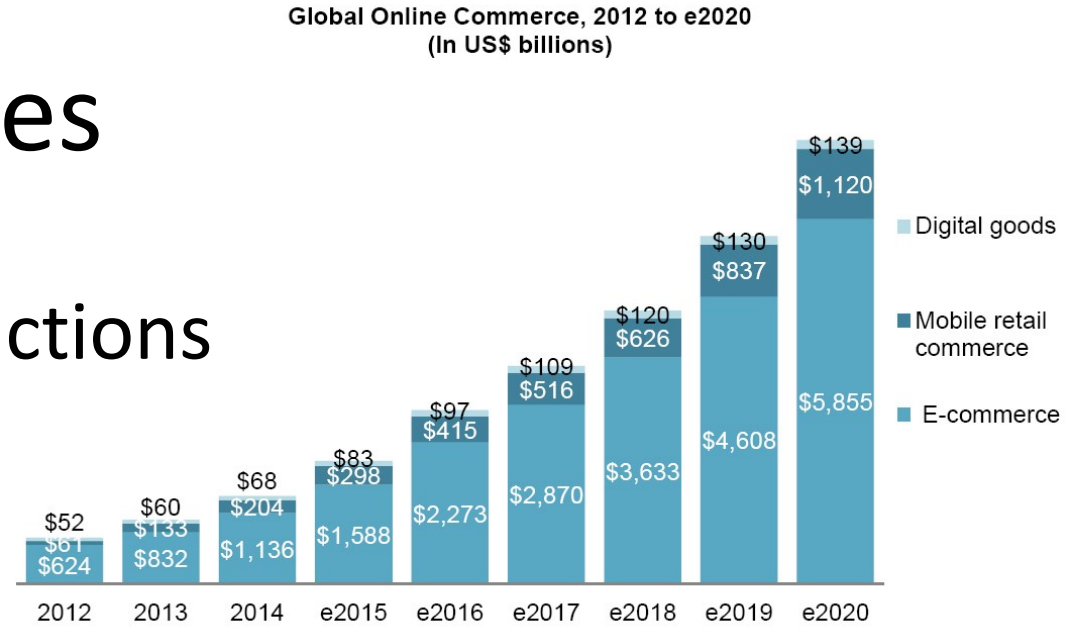
- Volume refers to the enormous amount of data generated every second from multiple sources such as social media, sensors, transactions, and devices.

1 Zettabyte = 10^{21}





Volume Examples



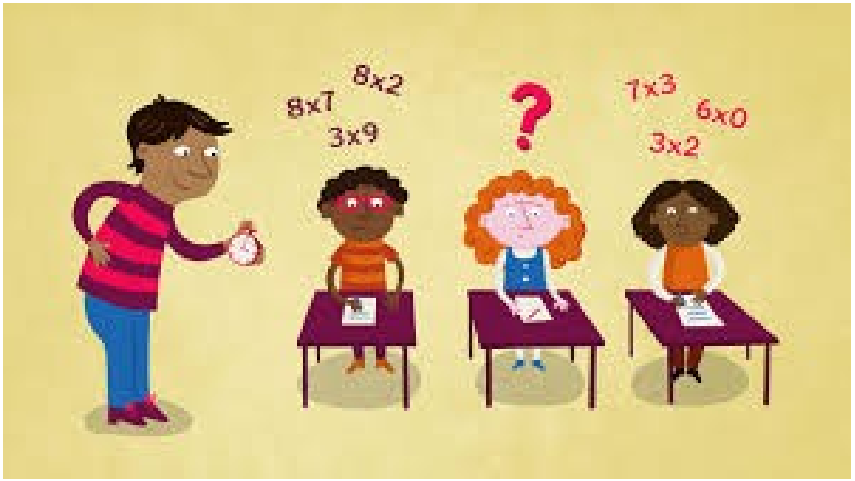
- E-commerce transactions
- Social media posts
- IoT sensor data

  The number of IoT devices is forecast to more than double from 19.8 billion in 2025 to over **40.6 billion by 2034**.

- Video streaming platforms

Velocity

- Velocity refers to the speed at which data is generated, processed, and analyzed.

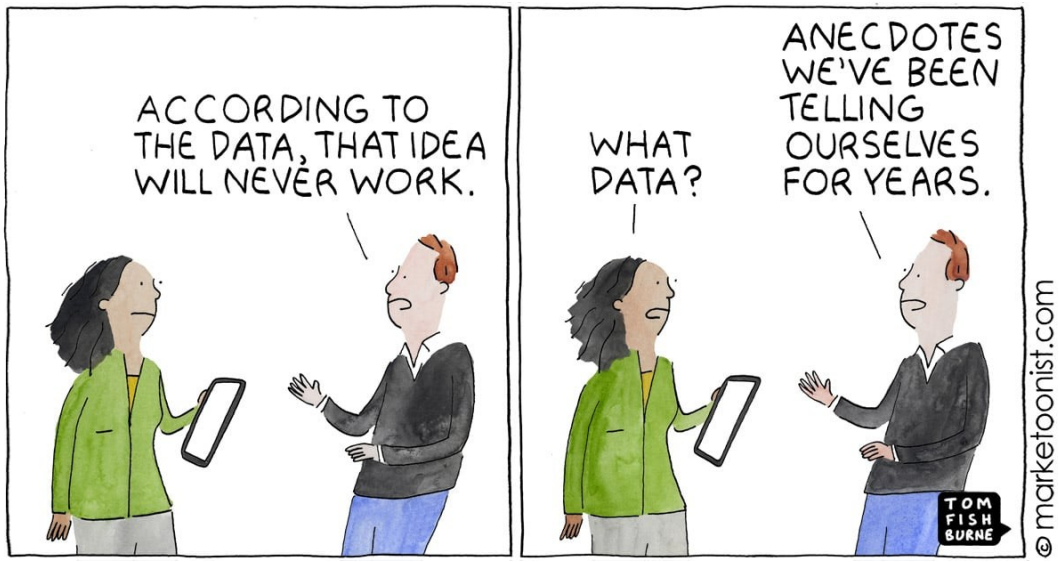


Velocity Examples

- Real-time stock market data
- Online recommendation systems
- Fraud detection systems
- IoT device monitoring

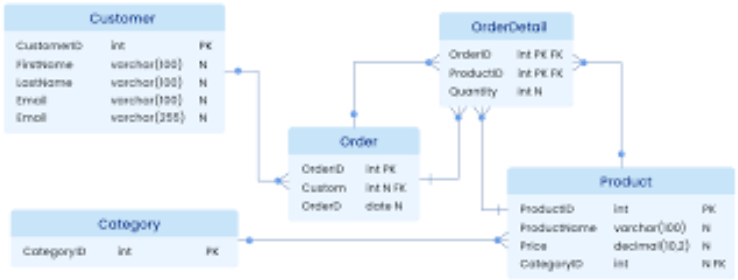
Variety

- Variety refers to the different types and formats of data: structured, semi-structured, and unstructured.



Variety Examples

Databases
(structured)



JSON/XML
(semi-structured)

JSON

```
{
  "name"
  "John Doe"
  "age"
  "New York"
}
```

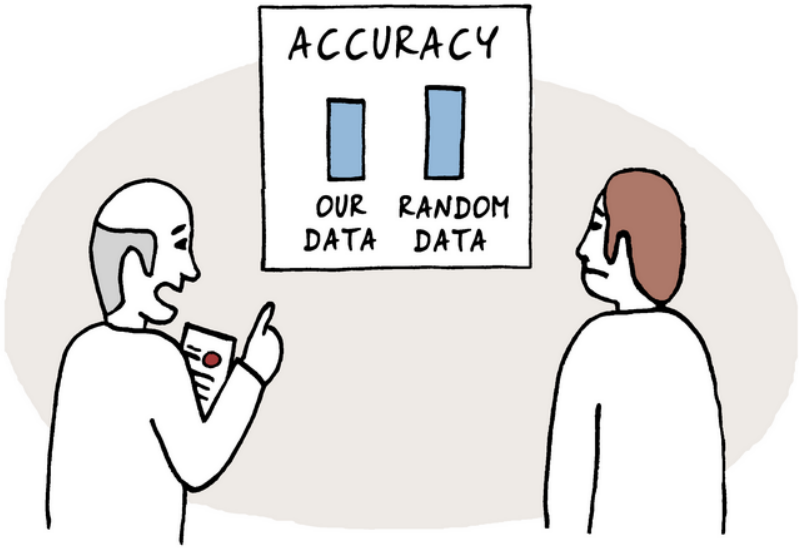
XML

```
<person>
  <name>John Doe
  <age>30
  <city>New York
</person>
```

Videos, images, emails
(unstructured)

Veracity

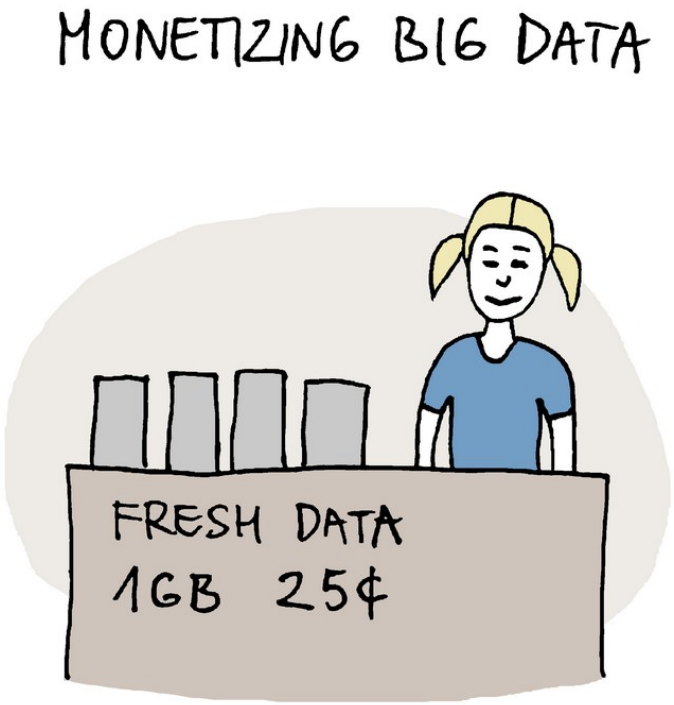
- Veracity refers to the
  - reliability,
  - accuracy, and
  - quality of data.



Challenges of Veracity

- Incomplete data
- Inconsistent sources
- Noise in datasets
- Bias and misinformation

Value



- Value refers to the meaningful insights and business benefits extracted from Big Data.

Examples of Value

- Customer insights
- Predictive maintenance
- Personalized marketing
- Risk analysis

Big Data in Business

- Organizations use Big Data to
  - optimize operations,
  - improve customer experiences, and
  - develop new products and services.

Business Applications

- Customer analytics
- Supply chain optimization
- Fraud detection
- Marketing optimization

Competitive Advantage

- Companies that effectively use Big Data can make faster and smarter decisions than competitors.

Big Data and Sustainability

- Big Data supports sustainable development by improving resource efficiency and enabling better environmental monitoring.

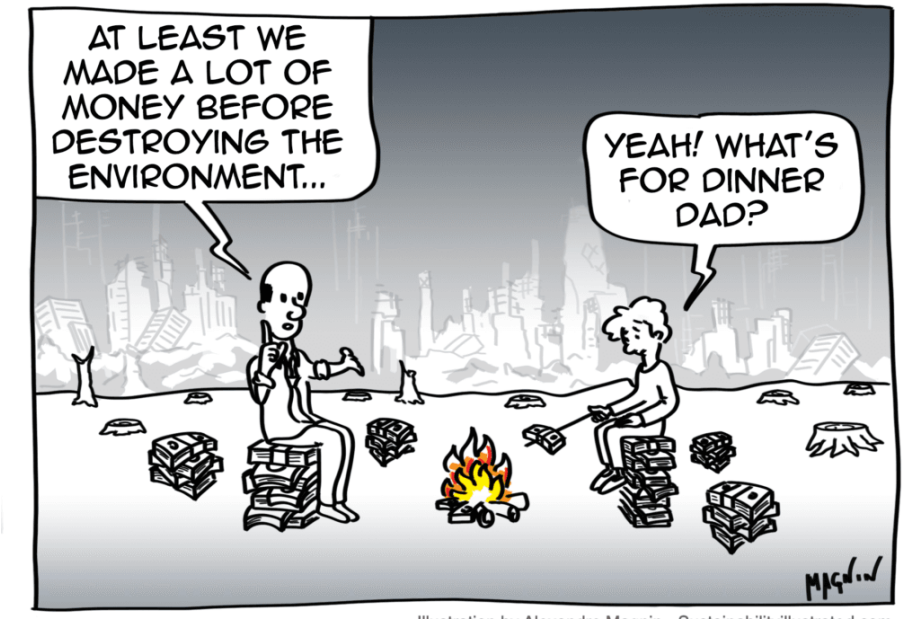


Illustration by Alexandre Magnin - Sustainabilityillustrated.com

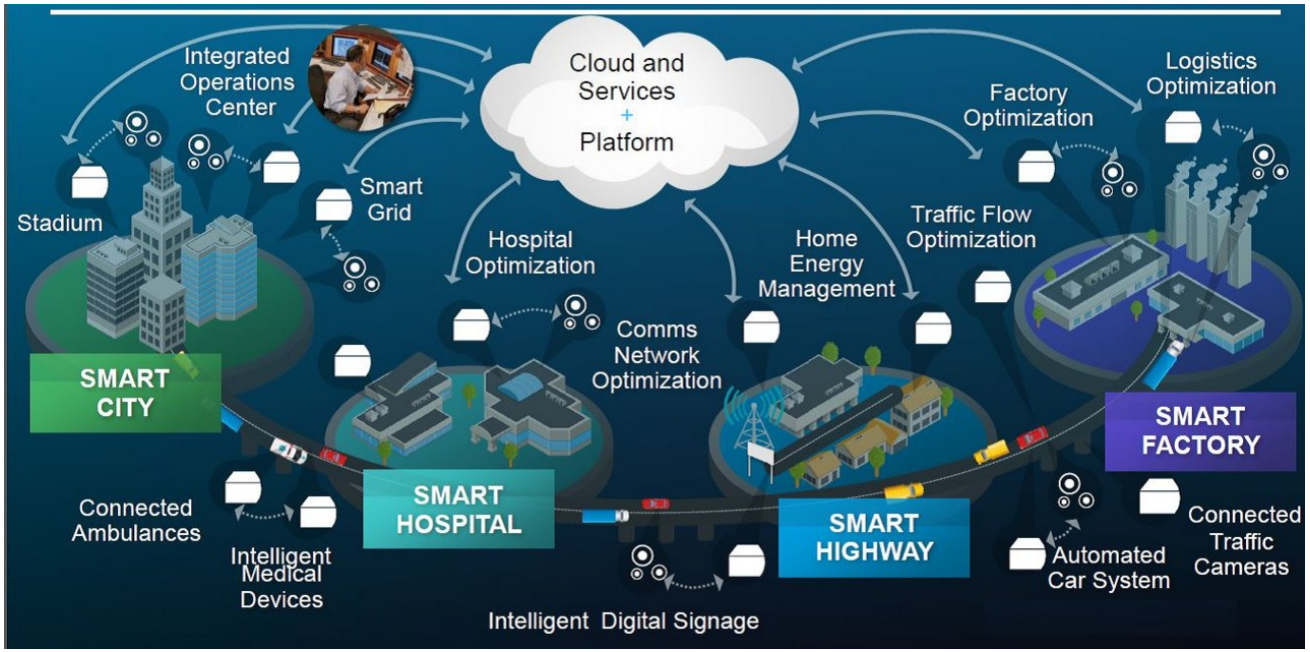
Sustainability Applications

- Smart cities
- Energy optimization
- Climate monitoring
- Sustainable agriculture

Illustration by Alexandre Magnin - Sustainabilityillustrated.com

Example: Smart Cities

- Data from sensors helps manage traffic, energy use, and waste management efficiently.



Example: Energy Sector

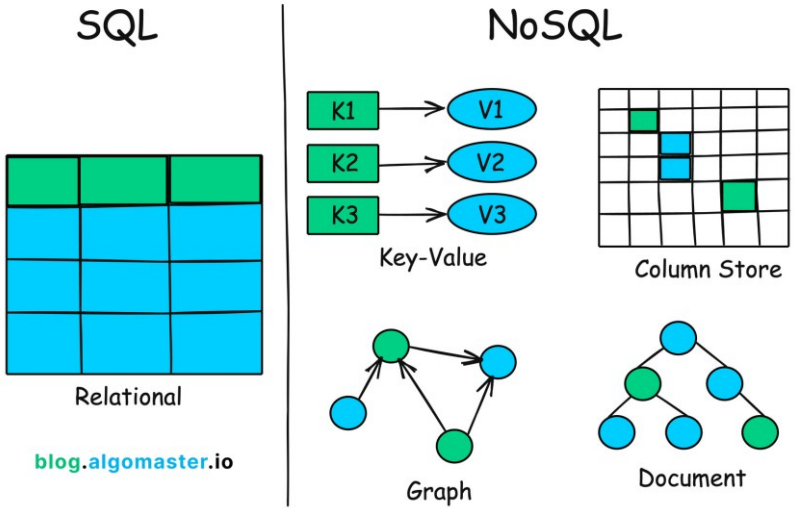
- Big Data enables prediction of energy demand and optimization of renewable energy usage.

Big Data Architecture Overview

- Typical architecture includes:
  - data sources,
  - storage systems,
  - processing frameworks, and
  - analytics tools.

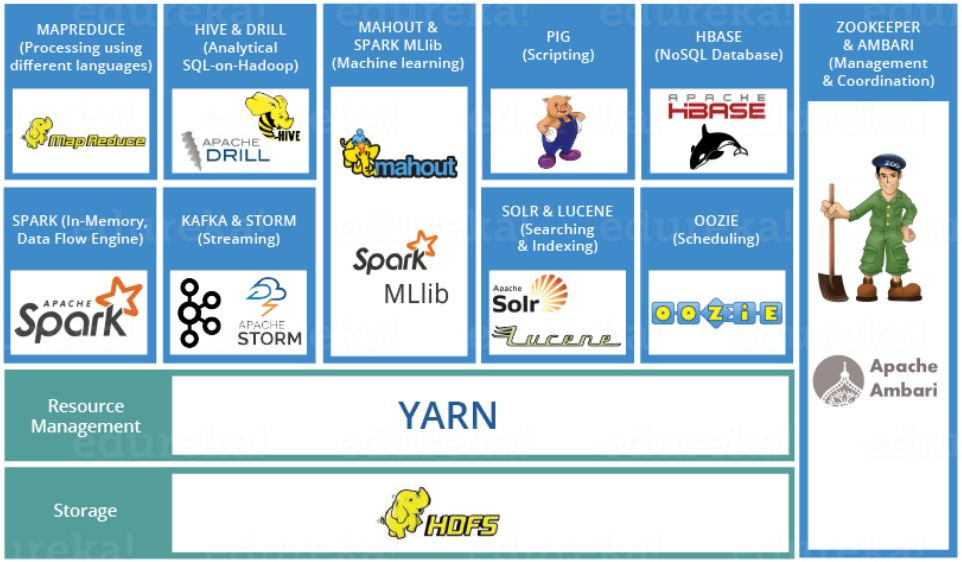
Key Technologies in Big Data

- Hadoop
- Apache Spark
- NoSQL Databases



Hadoop Overview

- Hadoop is an open-source framework that allows distributed storage and processing of large datasets across clusters of computers.

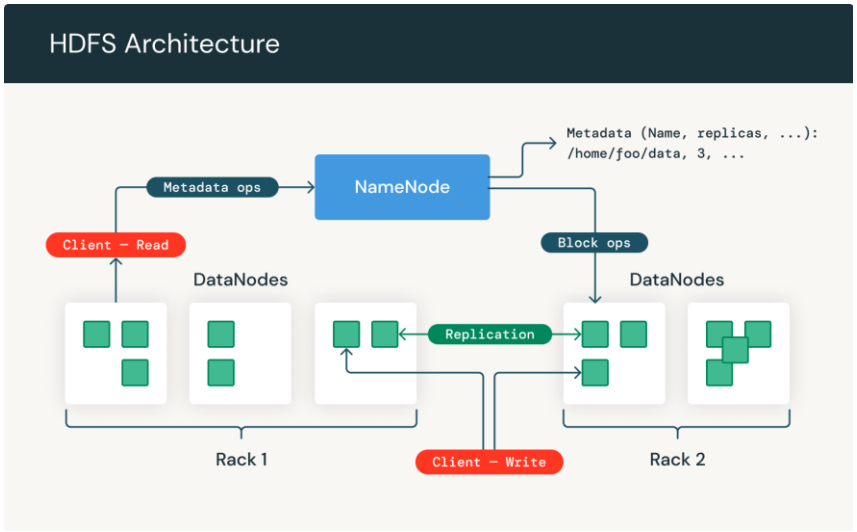


Hadoop Components

- HDFS (Hadoop Distributed File System)
- MapReduce
- Apache Hadoop YARN
- https://hadoop.apache.org/

HDFS (Hadoop Distributed File System)

- the primary storage system used by Hadoop applications.
- open source framework
- works by rapidly transferring data between nodes.
- often used by companies who need to handle and store big data.
- is a key component of many Hadoop systems, as it provides a means for managing big data, as well as supporting big data analytics.



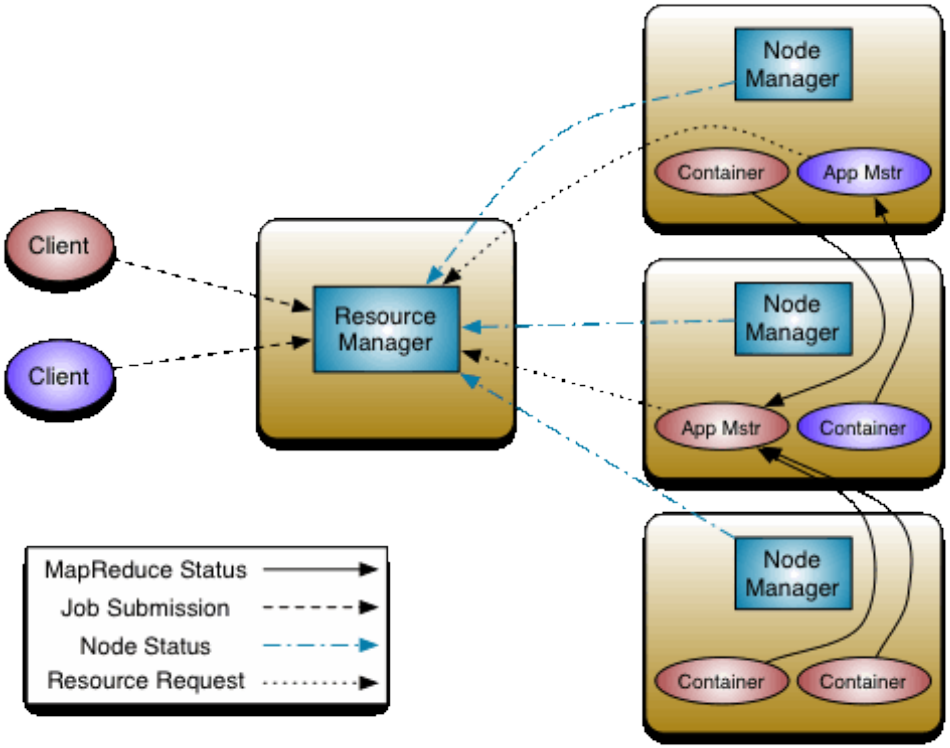
MapReduce

- MapReduce: A programming model for processing large datasets
- Big Data Processing: Designed to handle massive amounts of data
- Parallel Processing: Tasks are executed simultaneously
- Distributed Computing: Work is divided across multiple machines
- Cluster Environment: Runs on a group of interconnected computers

https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

Apache Hadoop YARN

- Apache Hadoop YARN Concept: Separates resource management from job scheduling and monitoring
- ResourceManager (RM): Global authority that allocates and manages cluster resources
- NodeManager (NM): Runs on each machine and manages containers and monitors resources (CPU, memory, disk, network)
- ApplicationMaster (AM): Created per application to request resources and manage task execution
- Applications: Can be a single job or a DAG (Directed Acyclic Graph) of jobs
- Framework Structure: ResourceManager + NodeManager form the cluster resource management framework



https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/YARN.html

Apache Spark Overview

- Apache Spark is a fast distributed data processing engine designed for large-scale data analytics.

Spark Advantages

- Faster than Hadoop MapReduce
- Real-time processing
- In-memory computation
- Supports machine learning

NoSQL Databases

- NoSQL databases are designed to store and manage unstructured or semi-structured data at scale.

Types of NoSQL

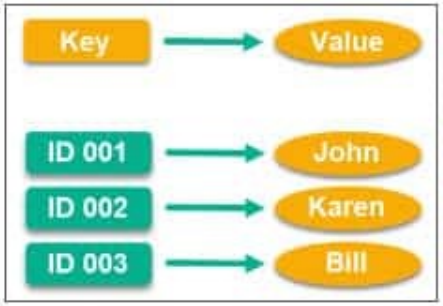
SQL Databases

Table

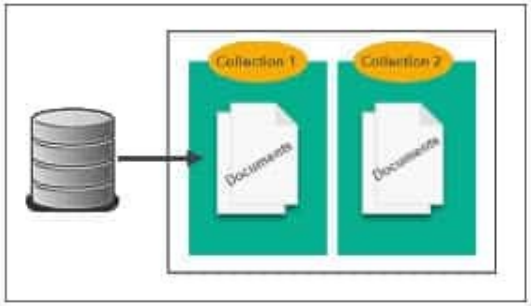
ID	Name	Grade	GPA
001	John	Senior	4.00
002	Karen	Freshman	3.67
003	Bill	Junior	3.33

NoSQL Databases

Key-value

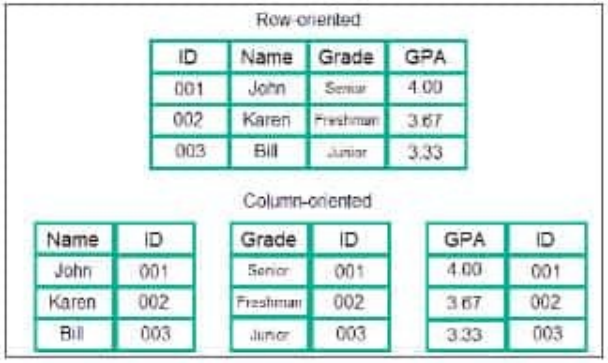


Document



Graph

Wide-column



NoSQL Examples

- MongoDB
- Cassandra
- Redis
- Neo4j

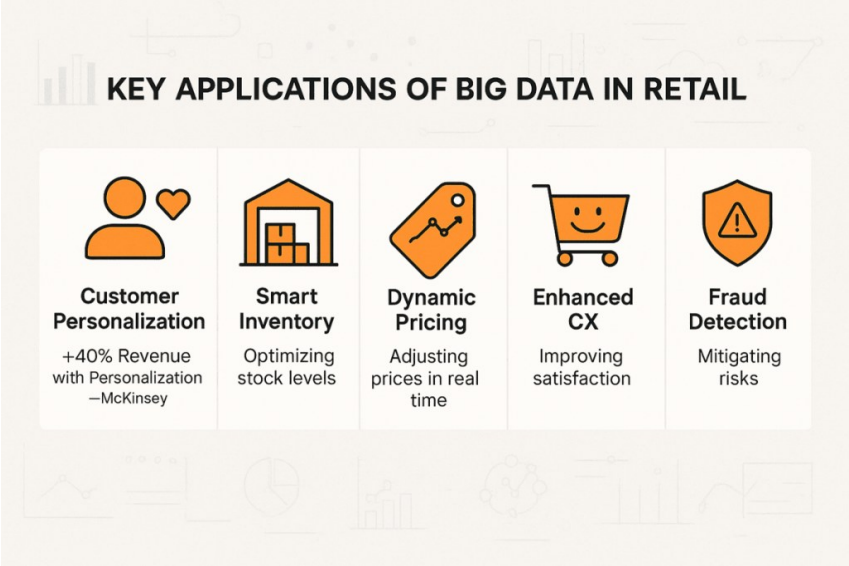
Type	Example
Key-Value Store	 
Wide Column Store	 
Document Store	 
Graph Store	 

Big Data Processing Workflow

1. Data Collection
2. Data Storage
3. Data Processing
4. Data Analysis
5. Visualization

Case Study Example

- Retail companies analyze purchasing patterns to personalize recommendations and improve sales.



- https://data.folio3.com/blog/big-data-in-retail-industry/

Challenges of Big Data

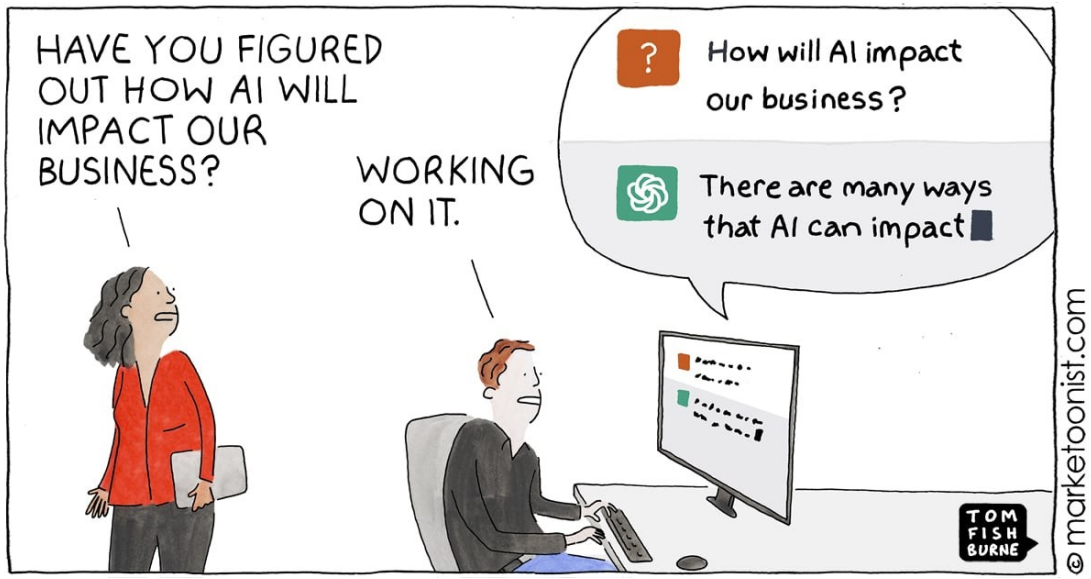
- Data privacy
- Security
- Infrastructure costs
- Data governance

"We have a VP of Records Management, but we don't know who it is because nobody can locate the file."

"I'm sure there are better ways to disguise sensitive information, but we don't have a big budget."

Future of Big Data

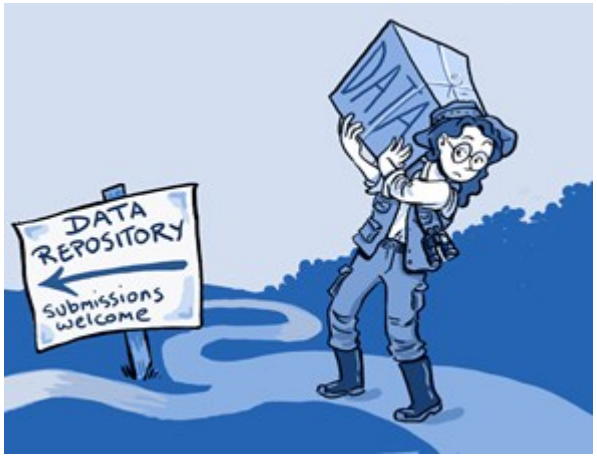
- AI integration

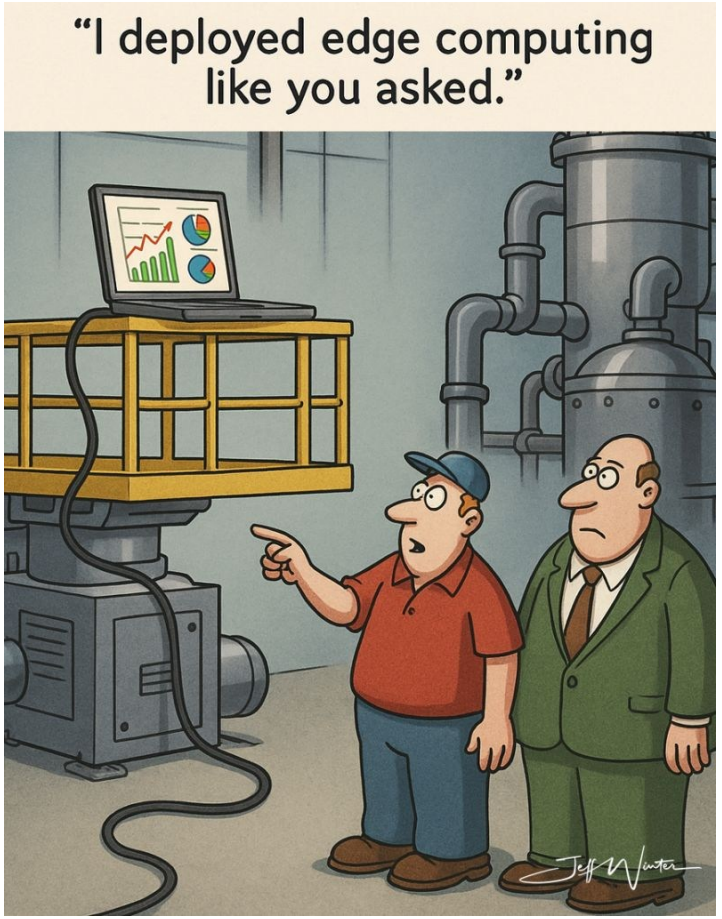


Real-time analytics

Future of Big Data

- Edge computing
- Increased automation





Discussion Activity

- Discuss examples of Big Data usage in your industry or daily life.

Summary

- Big Data defined by the 5Vs
- Critical for business innovation
- Supports sustainability
- Powered by technologies like Hadoop, Spark, and NoSQL

Questions

- Thank you!
- Questions and discussion.