
DATA
SCIENCE

DATA SCIENCE

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured, semi-structured and unstructured data. Data science is much more than simply analyzing data. It offers a range of roles and requires a range of skills.

What are data and information?

Data can be defined as a representation of facts, concepts, or instructions in a formalized manner, which should be suitable for communication, interpretation, or processing, by human or electronic machines. It can be described as unprocessed facts and figures. It is represented with the help of characters such as alphabets (A-Z, a-z), digits (0-9) or special characters (+, -, /, *, <,>, =, etc.).

What are data and information?

Information is the processed data on which decisions and actions are based. It is data that has been processed into a form that is meaningful to the recipient and is of real or perceived value in the current or the prospective action or decision of recipient.

Data Processing Cycle

Data processing is the re-structuring or re-ordering of data by people or machines to increase their usefulness and add values for a particular purpose. Data processing consists of the following basic steps - input, processing, and output. These three steps constitute the data processing cycle.

Data Processing Cycle

- **Input** − in this step, the input data is prepared in some convenient form for processing. The form will depend on the processing machine.
- **Processing** − in this step, the input data is changed to produce data in a more useful form.
- **Output** − at this stage, the result of the proceeding processing step is collected. The particular form of the output data depends on the use of the data.

Data types and their representation

Data types can be described from diverse perspectives. In computer science and computer programming, for instance, a data type is simply an attribute of data that tells the compiler or interpreter how the programmer intends to use the data.

Data types from Computer Programming Perspective

Common data types include:

- **Integers(int)** - is used to store whole numbers, mathematically known as integers.

- **Booleans(bool)** - is used to represent restricted to one of two values: true or false.

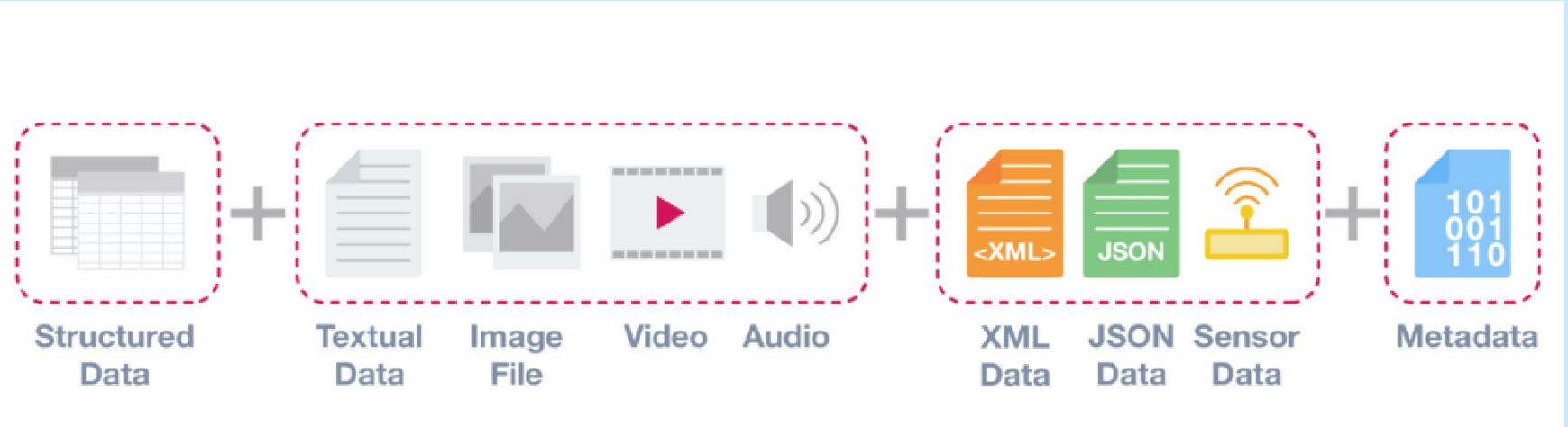
Data types from Computer Programming Perspective

- **Characters(char)** - is used to store a single character.
- **Floating-point numbers(float)** - is used to store real numbers.
- **Alphanumeric strings(string)** - used to store a combination of characters and numbers.

DATA TYPES FROM DATA ANALYTICS PERSPECTIVE

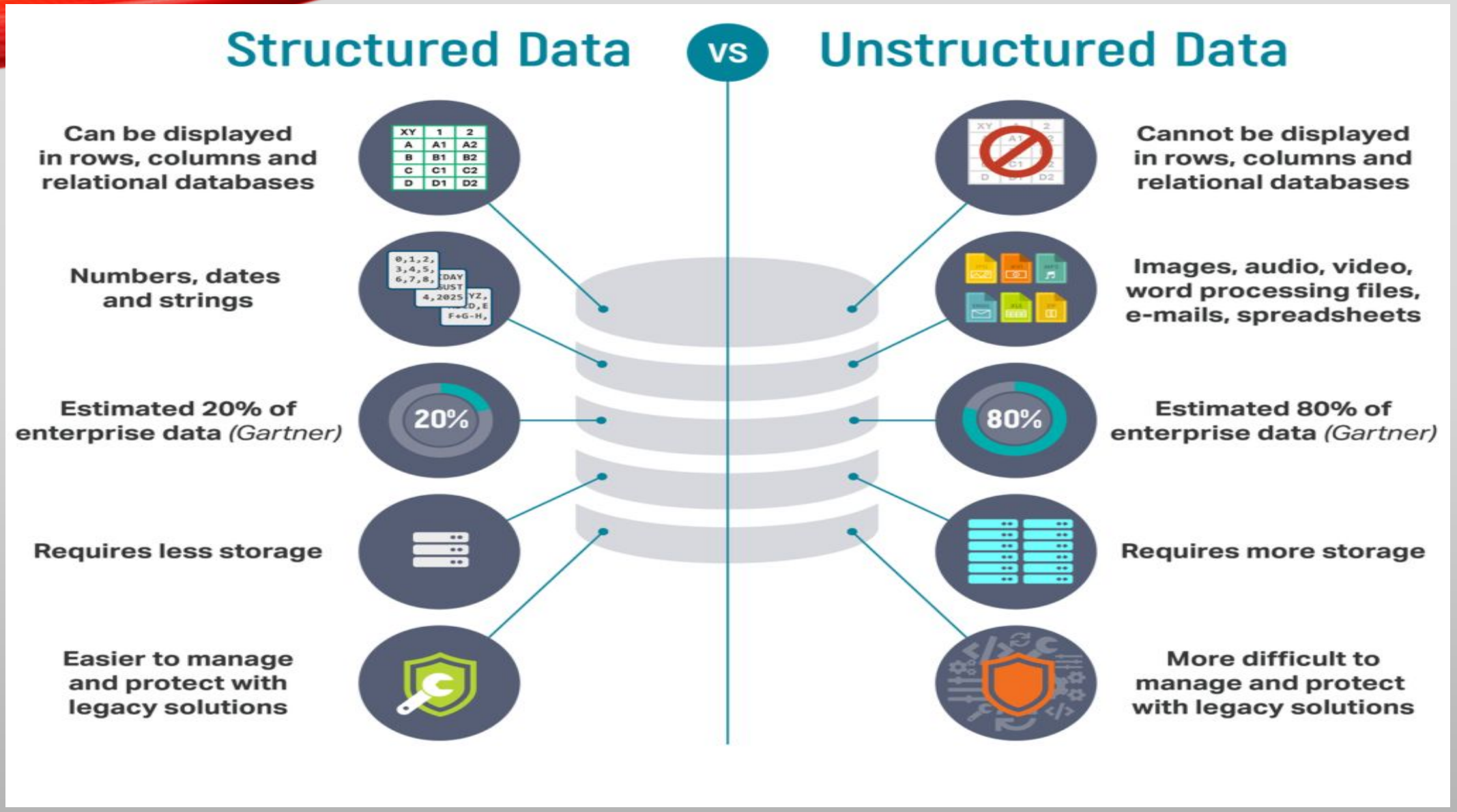
Data types from Data Analytics perspective

From a data analytics point of view, it is important to understand that there are three common types of data types or structures: **Structured**, **Semi-structured**, and **Unstructured data types**.



Structured Data

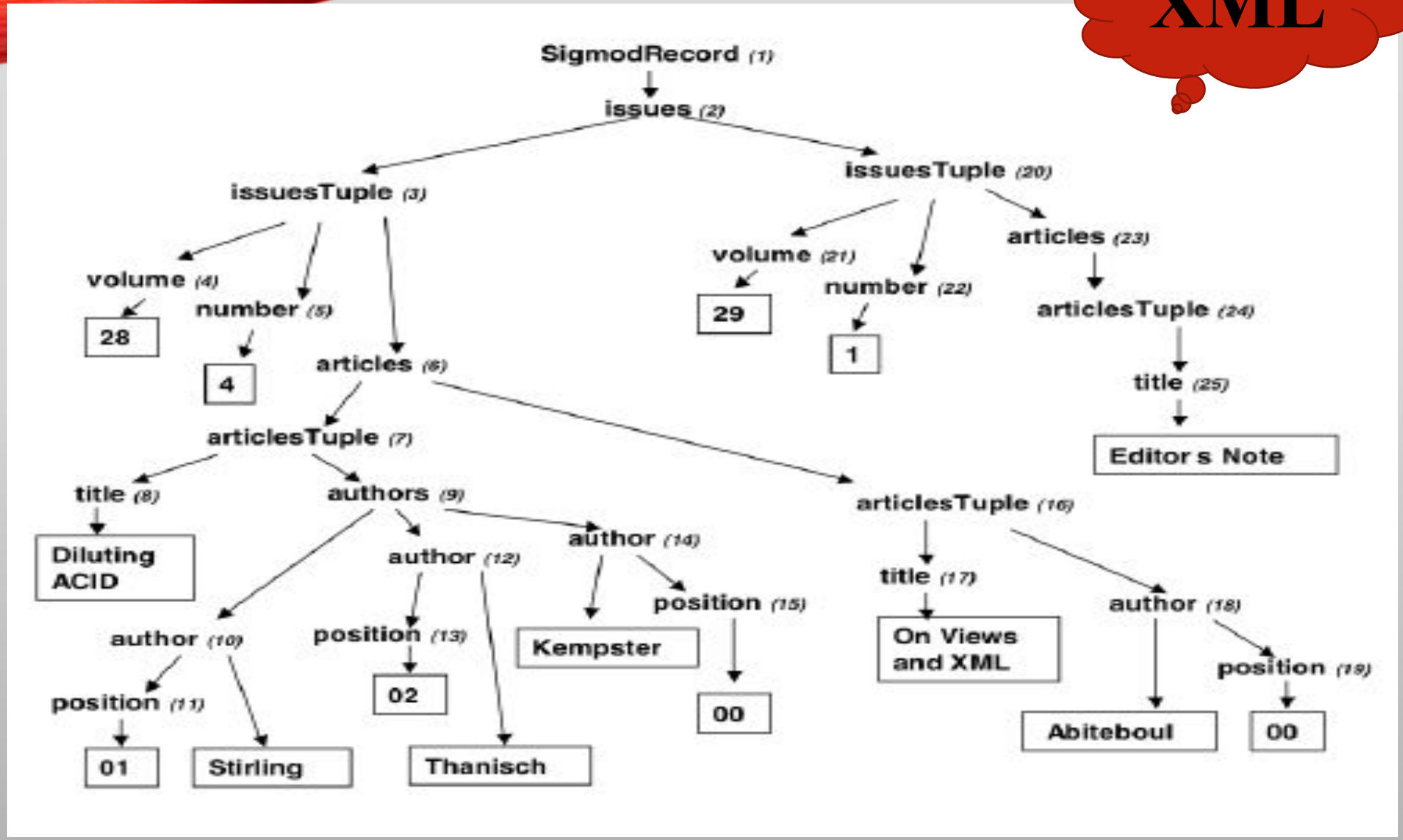
- is data that adheres to a per-defined data model and is therefore straightforward to analyze. Structured data conforms to a tabular format with a relationship between the different rows and columns.


Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
Requires less storage
Easier to manage and protect with legacy solutions
20%
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
Requires more storage
More difficult to manage and protect with legacy solutions
80%

Semi-structured Data

- is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless, contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.

XML


SigmodRecord (1)
issues (2)
issuesTuple (3)
issuesTuple (20)
volume (4)
number (5)
articles (6)
volume (21)
number (22)
articles (23)
28
4
29
1
articlesTuple (24)
title (25)
Editor s Note
articlesTuple (7)
title (8)
authors (9)
articlesTuple (16)
Diluting ACID
author (14)
author (12)
title (17)
author (18)
author (10)
position (13)
position (15)
Kempster
On Views and XML
position (11)
position (19)
02
00
Abiteboul
00
01
Stirling
Thanisch

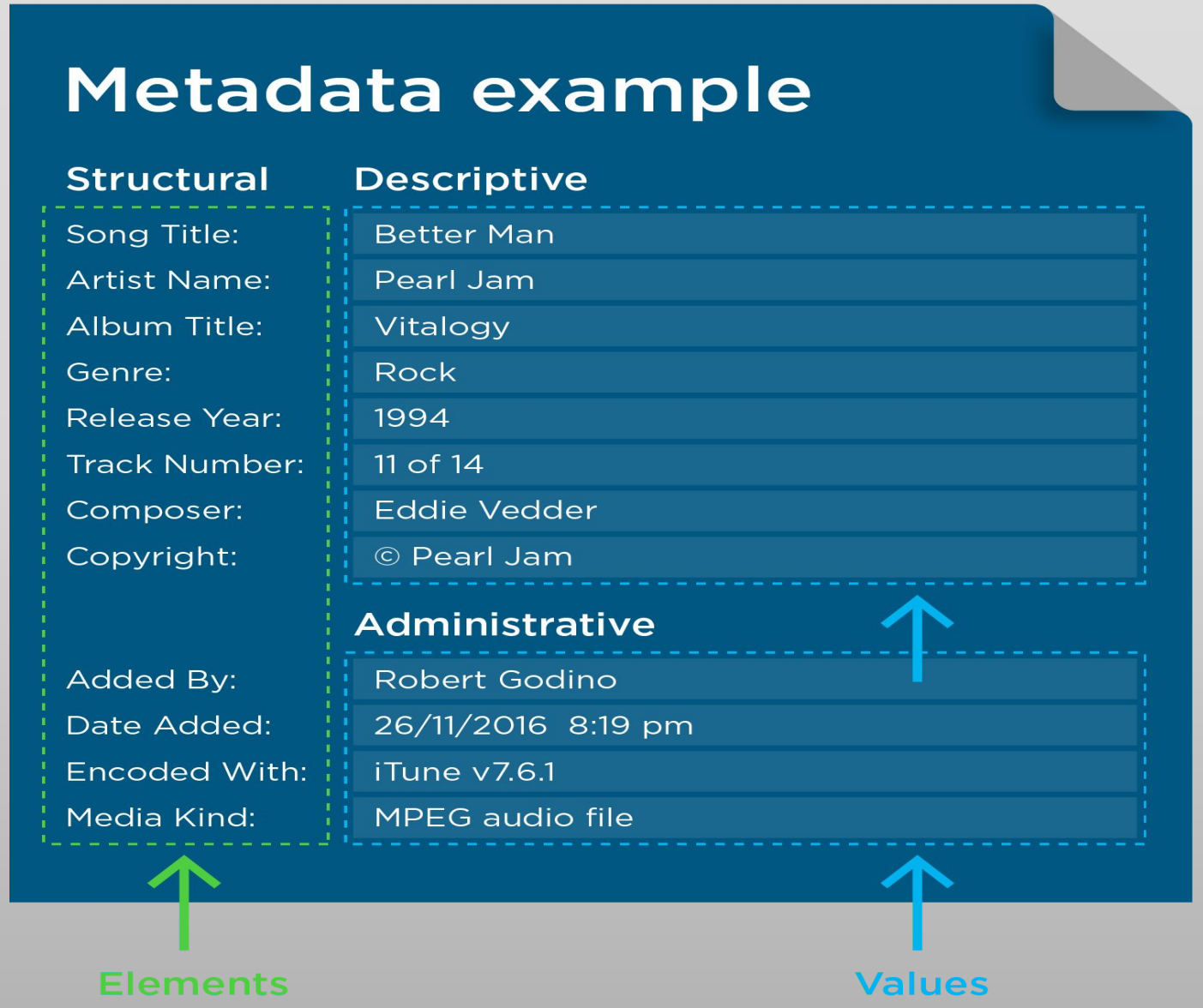
Unstructured Data

- is information that either does not have a predefined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structured databases.

Metadata-Data about Data

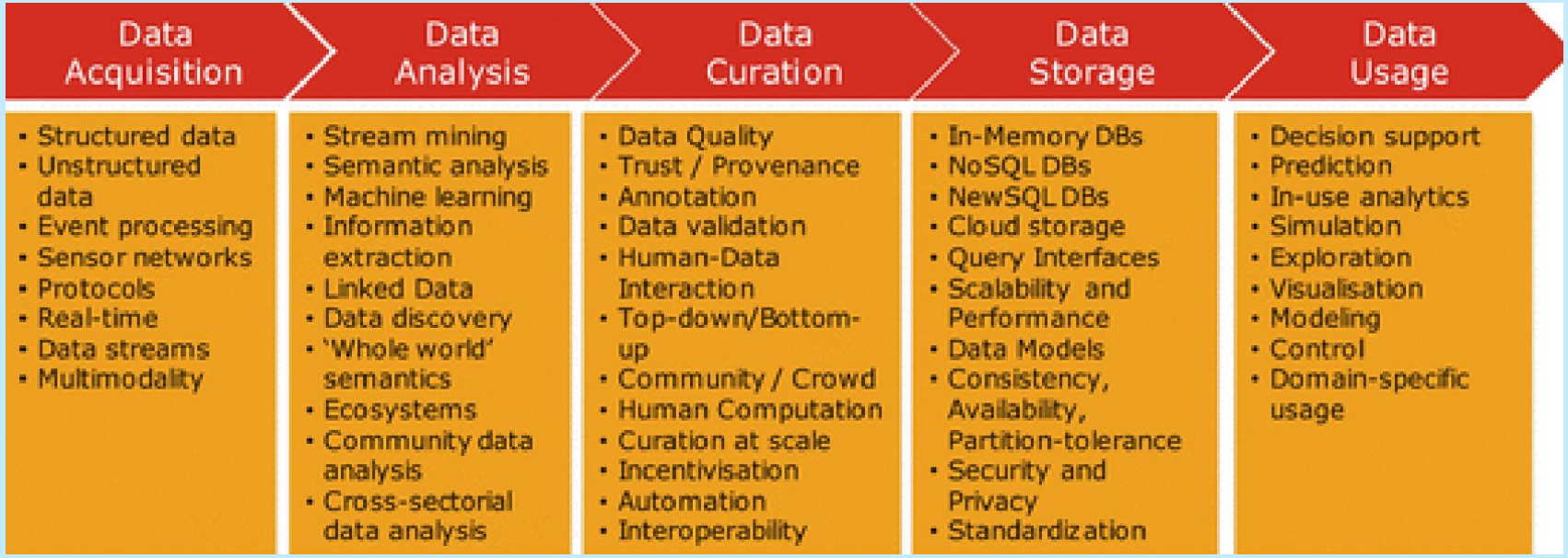
The last category of data type is metadata. From a technical point of view, this is not a separate data structure, but it is one of the most important elements for Big Data analysis and big data solutions. Metadata is data about data. It provides additional information about a specific set of data.

Metadata example



DATA VALUE CHAIN

The Data Value Chain is introduced to describe the information flow within a big data system as a series of steps needed to generate value and useful insights from data. The Big Data Value Chain identifies the following key high-level activities:



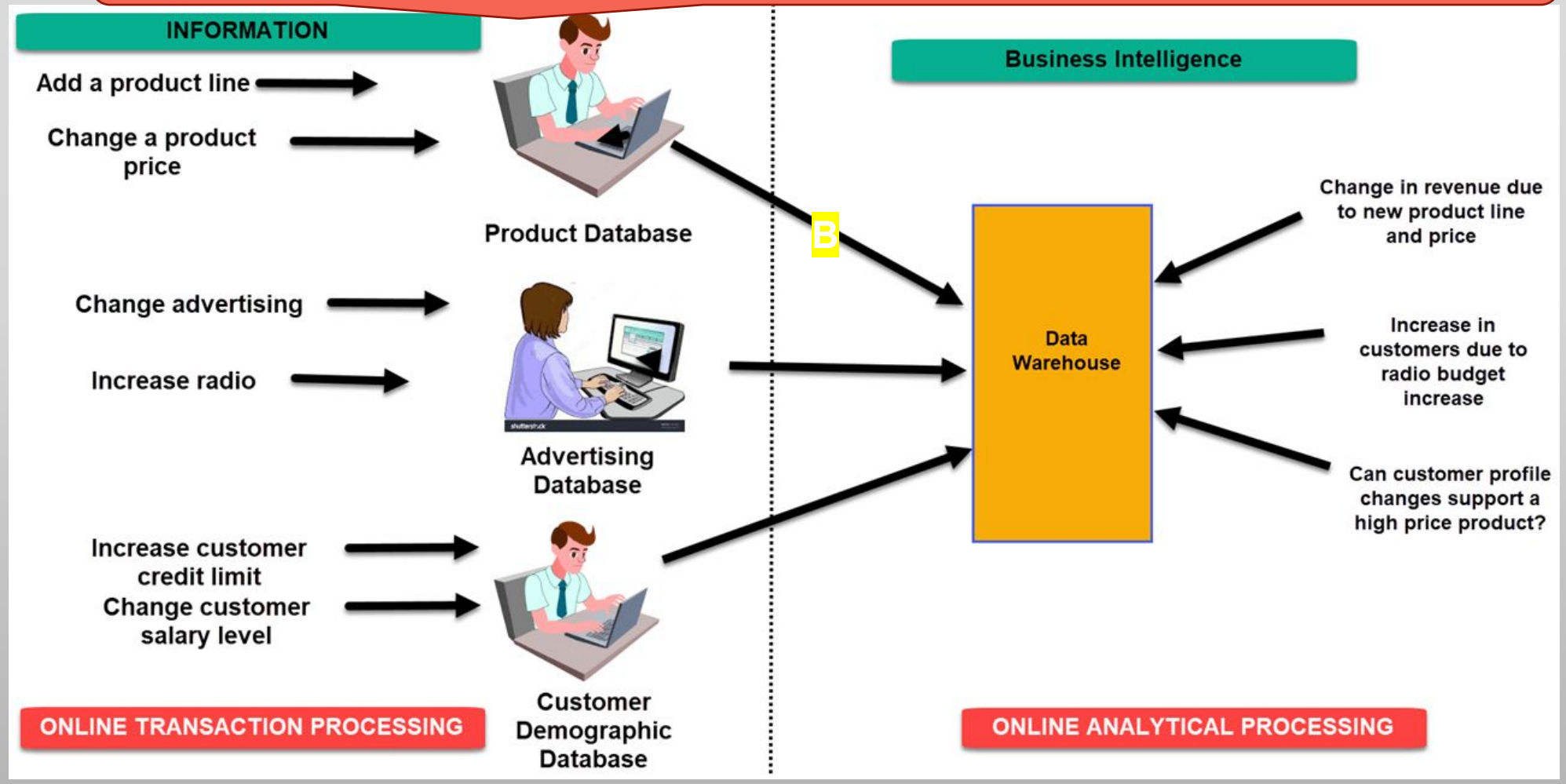
Data Acquisition

It is the process of gathering, filtering, and cleaning data before it is put in a data warehouse or any other storage solution on which data analysis can be carried out. It is one of the major big data challenges in terms of infrastructure requirements.

Data Analysis

It is concerned with making the raw data acquired amenable to use in decision-making as well as domain-specific usage. It also involves exploring, transforming, and modeling data with the goal of highlighting relevant data, synthesizing and extracting useful hidden information with high potential from a business point of view.

- **Business Intelligence an example area in Data Analysis**



Data Curation

It is the active management of data over its life cycle to ensure it meets the necessary data quality requirements for its effective usage.

It is performed by experts called **data curators** (also known as scientific curators or data annotators), they hold the responsibility of ensuring that data are trustworthy, discoverable, accessible, reusable and fit their purpose.

Data curation processes can be categorized into different activities such as:

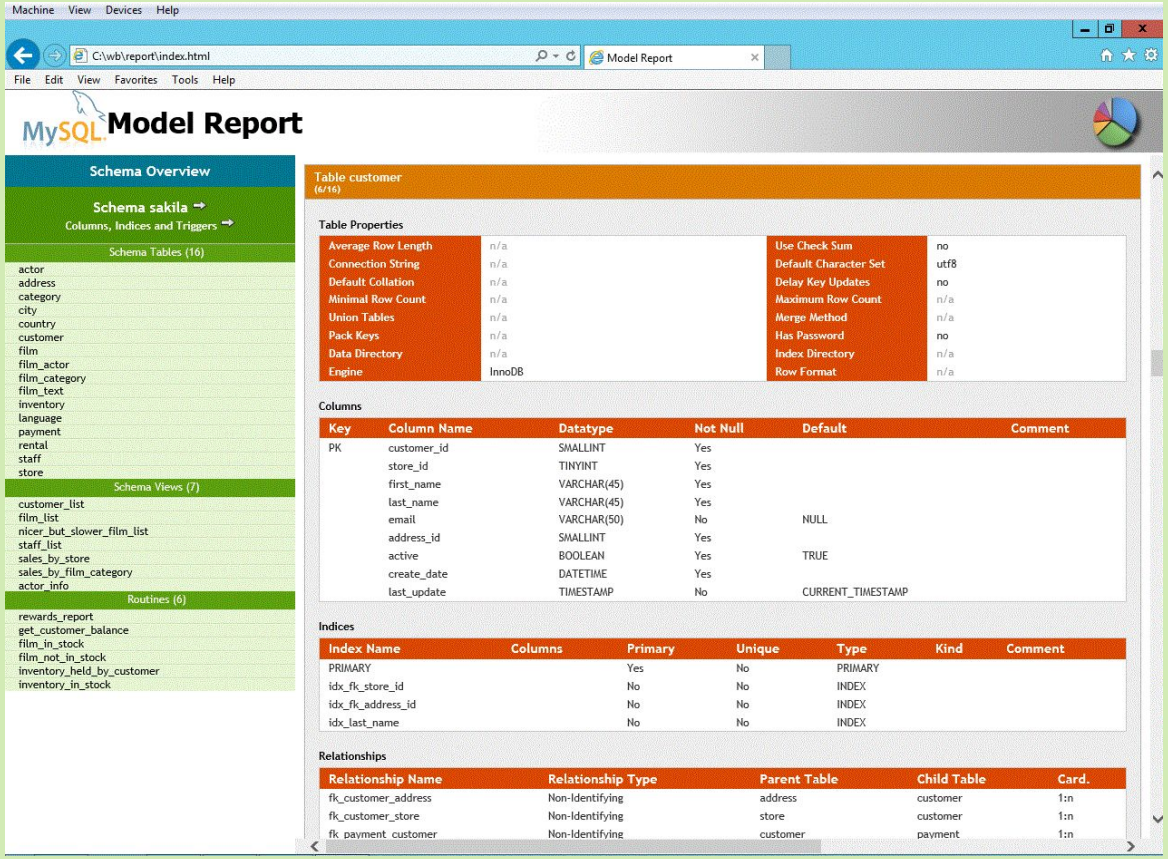
-content creation
-selection,
-classification,
-transformation
-validation
-preservation

Data Storage

It is the persistence and management of data in a scalable way that satisfies the needs of applications that require fast access to the data.

Relational Database Management Systems (RDBMS) have been the main, and almost unique, a solution to the storage paradigm for nearly 40 years.

- **MySQL** is one of the most well-known RDBMSs.



However, the **ACID (Atomicity, Consistency, Isolation, and Durability)** properties that guarantee database transactions lack flexibility with regard to schema changes and the performance and fault tolerance when data volumes and complexity grow, making them unsuitable 28 for big data scenarios.

NoSQL technologies have been designed with the scalability goal in mind and present a wide range of solutions based on alternative data models.

Data Usage

It covers the data-driven business activities that need access to data, its analysis, and the tools needed to integrate the data analysis within the business activity.

Basic Concepts of Big Data

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing have greatly expanded in recent years.

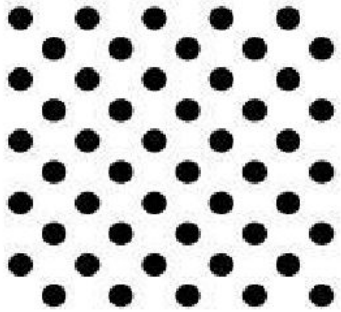
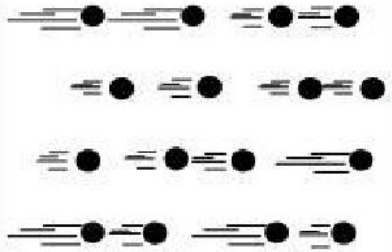
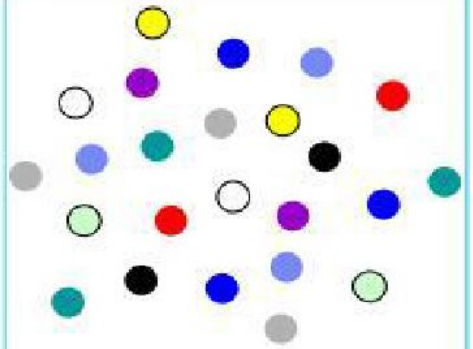
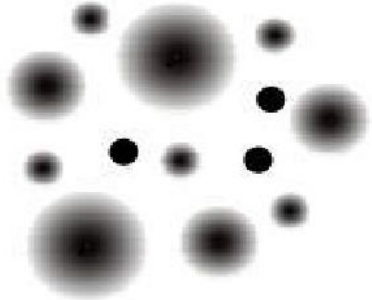
What Is Big Data?

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.

In this context, a “large dataset” means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization. Big data is characterized by 3V and more:

Big data is characterized by 3V and more:

- **Volume**: large amounts of data Zeta bytes/Massive datasets
- **Velocity**: Data is live streaming or in motion
- **Variety**: data comes in many different forms from diverse sources
- **Veracity**: can we trust the data? How accurate is it? etc.

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

CLUSTERED COMPUTING AND HADOOP ECOSYSTEM

Clustered Computing

- Cluster computing is a form of computing in which a group of computers are linked together so that they can act like a single entity.
- It is the technique of linking two or more computers into a network (usually through a local area network) in order to take advantage of the parallel processing power of those computers.

Big data clustering software combines the resources of many smaller machines, seeking to provide a number of benefits:

- **Resource Pooling**: Combining the available storage space to hold data is a clear benefit, but CPU and memory pooling are also extremely important. Processing large datasets requires large amounts of all three of these resources.
- **High Availability**: Clusters can provide varying levels of fault tolerance and availability guarantees to prevent hardware or software failures from affecting access to data and processing. This becomes increasingly important as we continue to emphasize the importance of real-time analytics.

30

- **Easy Scalability**: Clusters make it easy to scale horizontally by adding additional machines to the group. This means the system can react to changes in resource requirements without expanding the physical resources on a machine.

Hadoop and its Ecosystem

Hadoop is an open-source framework intended to make interaction with big data easier. It is a framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models. It is inspired by a technical document published by Google.

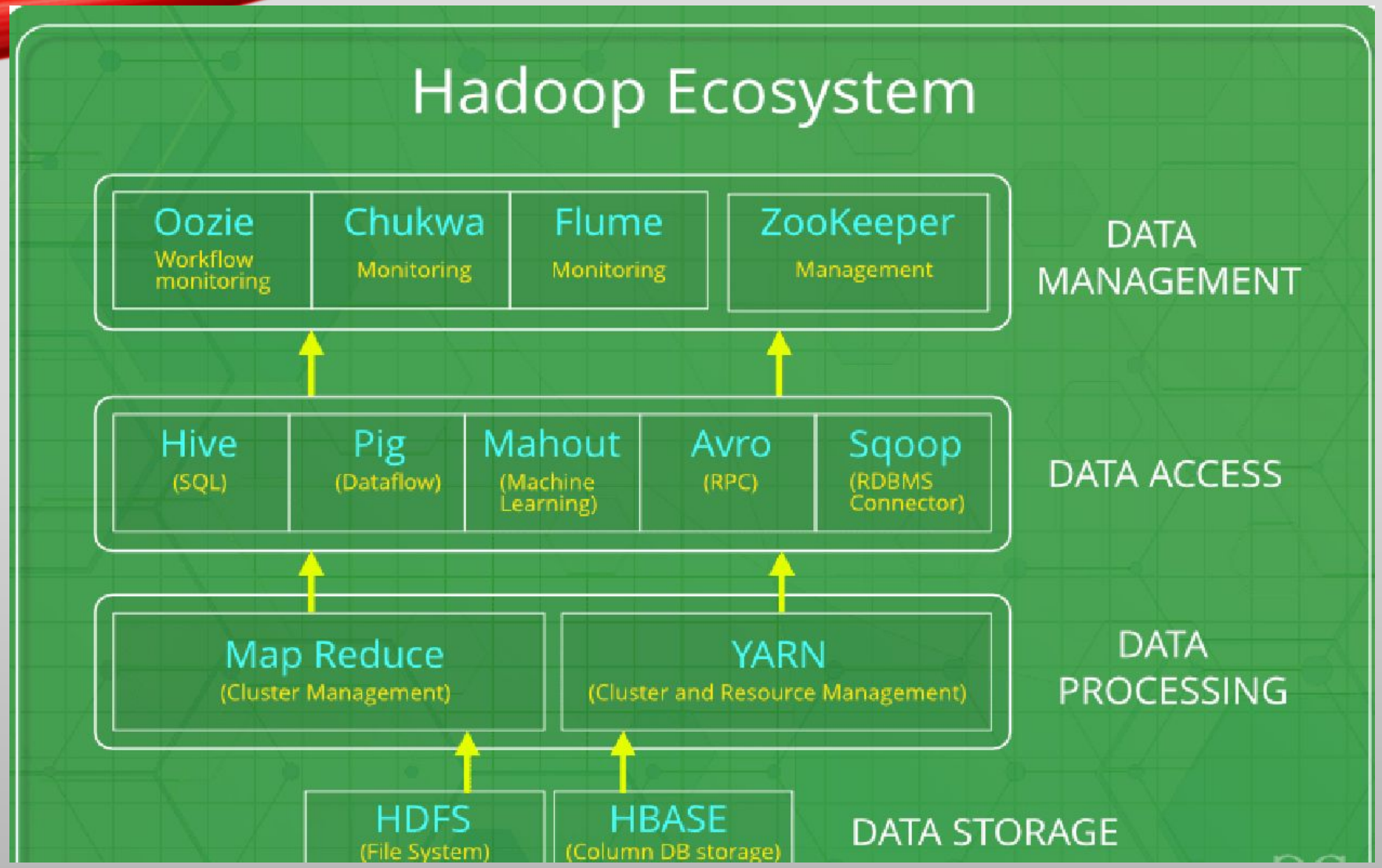
The four key characteristics of Hadoop are:

- **Economical:** Its systems are highly economical as ordinary computers can be used for data processing.
- **Reliable:** It is reliable as it stores copies of the data on different machines and is resistant to hardware failure.
- **Scalable:** It is easily scalable both, horizontally and vertically. A few extra nodes help in scaling up the framework.

- **Flexible:** It is flexible and you can store as much structured and unstructured data as you need to and decide to use them later.

Hadoop has an ecosystem that has evolved from its four core components: data management, access, processing, and storage. It is continuously growing to meet the needs of Big Data. It comprises the following components and many others:

- **HDFS:** Hadoop Distributed File System
- YARN: Yet Another Resource Negotiator
- **MapReduce:** Programming based Data Processing
- **Spark:** In-Memory data processing
- **PIG, HIVE:** Query-based processing of data services
- **HBase:** NoSQL Database
- Mahout, Spark MLLib: Machine Learning algorithm libraries
- **Solar, Lucene:** Searching and Indexing
- **Zookeeper:** Managing cluster
- **Oozie:** Job Scheduling


Hadoop Ecosystem
Oozie
Workflow monitoring
Chukwa
Monitoring
Flume
Monitoring
ZooKeeper
Management
DATA MANAGEMENT
Hive
(SQL)
Pig
(Dataflow)
Mahout
(Machine Learning)
Avro
(RPC)
Sqoop
(RDBMS Connector)
DATA ACCESS
Map Reduce
(Cluster Management)
YARN
(Cluster and Resource Management)
DATA PROCESSING
HDFS
(File System)
HBASE
(Column DB storage)
DATA STORAGE

History of Hadoop

- Hadoop was created by Daug Cutting who had created the Apache Lucene (Text Search), which is origin in Apache Nutch (Open source search Engine). Hadoop is a part of Apache Lucene Project. Actually Apache Nutch was started in 2002 for working crawler and search.
- In January 2008, Hadoop was made its own top-level project at Apache for, confirming success, By this time, Hadoop was being used by many other companies such as Yahoo!, Facebook, etc.
- In April 2008, Hadoop broke a world record to become the fastest system to sort a terabyte of data.
- Yahoo take test in which To process ITB of data (1024 columns)
  oracle – 3 $^1/_2\, day$
  Teradata – 4 $^1/_2\, day$
  Netezza – 2 hour 50 min
  Hadoop – 3.4 min

Big Data Life Cycle with Hadoop

Ingesting data into the system

The first stage of Big Data processing is Ingest. The data is ingested or transferred to Hadoop from various sources such as relational databases, systems, or local files. Sqoop transfers data from RDBMS to HDFS, whereas Flume transfers event data.

Big Data Life Cycle with Hadoop

Processing the data in storage

The second stage is Processing. In this stage, the data is stored and processed. The data is stored in the distributed file system, HDFS, and the NoSQL distributed data, HBase. Spark and MapReduce perform data processing.

Big Data Life Cycle with Hadoop

Computing and analyzing data

The third stage is to Analyze. Here, the data is analyzed by processing frameworks such as Pig, Hive, and Impala. Pig converts the data using a map and reduce and then analyzes it. Hive is also based on the map and reduce programming and is most suitable for structured data.

Big Data Life Cycle with Hadoop

Visualizing the results

The fourth stage is Access, which is performed by tools such as Hue and Cloudera Search. In this stage, the analyzed data can be accessed by users.

GROUP 4

NECALY JAGUNOS
VEAH CABULAO
ROSENDA LUMANTAS
CLARESE TINAJA


THE END