Harambee University



Faculty Technology and FBE

Department of Computer science and Accounting

Chapter 2 :- Data S

Chapter content

- 2.1 An Overview of Data Science
- 2.1.2. Data Processing Cycle
- 2.3 Data types and their represe
- 2.4 Data value Chain
- 2.5 Basic concepts of big data



Introduction

- In the previous chapter, the concept of the role of data for emerging technologies was discussed.
- **In this chapter, you are going to learn more about data science, data vs. information, data types and representation, data value chain, and basic concepts of big data.**

2.1. An Overview of 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
- As an academic discipline and profession, data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals

2.1.1. 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.).
- Whereas ***information*** is the **processed data on which decisions and actions are based.**

2.1.2. 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 3 steps –**

1. **input,**
2. **processing, and**
3. **output.**

These **three steps** constitute the data proce



Cont’d…

- **Input −** in this step, the input data is prepared in some convenient form for processing. The form will depend on the processing machine**.**
- **For example,** when electronic computers are used, the input data can be recorded on any one of the several types of storage medium, such as hard disk, CD, flash disk and so on.
- **Processing −** in this step, the input data is changed to produce data in a more useful form.

Cont’d…

- For example, interest can be calculated on deposit to a bank, or a summary of sales for the month can be calculated from the sales orders.
- **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. For example, output data may be payroll for employees.

2.3 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.

2.3.1. 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
- **Characters(char)-** is used to store a single character
- **Floating-point numbers(float)-** is used to store real numbers

Cont’d

- **Alphanumeric strings(string)-** used to store a combination of characters and numbers.
- A data type makes the values that expression, such as a variable or a function, might take.
- This data type defines the operations that can be done on the data, the meaning of the data, and the way values of that type can be stored.

2.3.2.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.**

Cont'd...

Fig. 2.2 below describes the three types of data and metadata.



Structured Data

- **Structured data** is data that adheres to **a pre-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. **Common examples of structured data** are Excel files or SQL databases. Each of these has structured rows and columns that can be sorted.

Semi-structured Data

- **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.
- Therefore, it is also known as a **self-describing structure. Examples of semi-structured data include** are forms of semi-structured data.

Unstructured Data

- **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.
- **Common examples of unstructured data include** audio, video files or No-SQL 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.
- In a set of photographs, for example, metadata could describe when and where the photos were taken.

Cont'd...

- The metadata then provides fields for dates and locations which, by themselves, can be considered structured data.
- Because of this reason, metadata is frequently used by Big Data solutions for initial analysis.

Activity 2.4

- *Discuss data types from programing and analytics perspectives.*
- *Compare metadata with structured, unstructured and semi-structured data*
- *Given at least one example of structured, unstructured and semi-structured data types*

2.4. 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 fo



Figure 2.3 Data Value Chain

2.4.1. 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.
- Data acquisition is one of the major big data challenges in terms of infrastructure requirements.
- The infrastructure required to support the acquisition of big data must deliver low, predictable latency in both capturing data and in executing queries; be able to handle very high transaction volumes, often in a distributed environment; and support flexible

2.4.2. Data Analysis

- It is concerned with making the raw data acquired amenable to use in decision-making as well as domain-specific usage.
- Data analysis 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.
- Related areas include data mining, business intelligence, and machine learning.

2.4.3. 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.
- Data curation processes can be categorized into different activities such as content creation, selection, classification, transformation, validation, and preservation.
- Data curation is performed by expert curators that are responsible for improving the accessibility and quality of data.

Cont’d

- **Data curators** (also known as scientific curators or data annotators) hold the responsibility of ensuring that data are trustworthy, discoverable, accessible, reusable and fit their purpose.
- A key trend for the duration of big data utilizes community and crowdsourcing approaches.

2.4.4. 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.

Cont’d

- 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 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.

2.4.5. 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.
- Data usage in business decision-making can enhance competitiveness through the reduction of costs, increased added value, or any other parameter that can be measured against existing performance criteria.

2.5. 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.
- In this section, we will talk about big data on a fundamental level and define common concepts you might come across. We will also take a high-level look at some of the processes and technologies currently being used in this

2.5.1. 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.

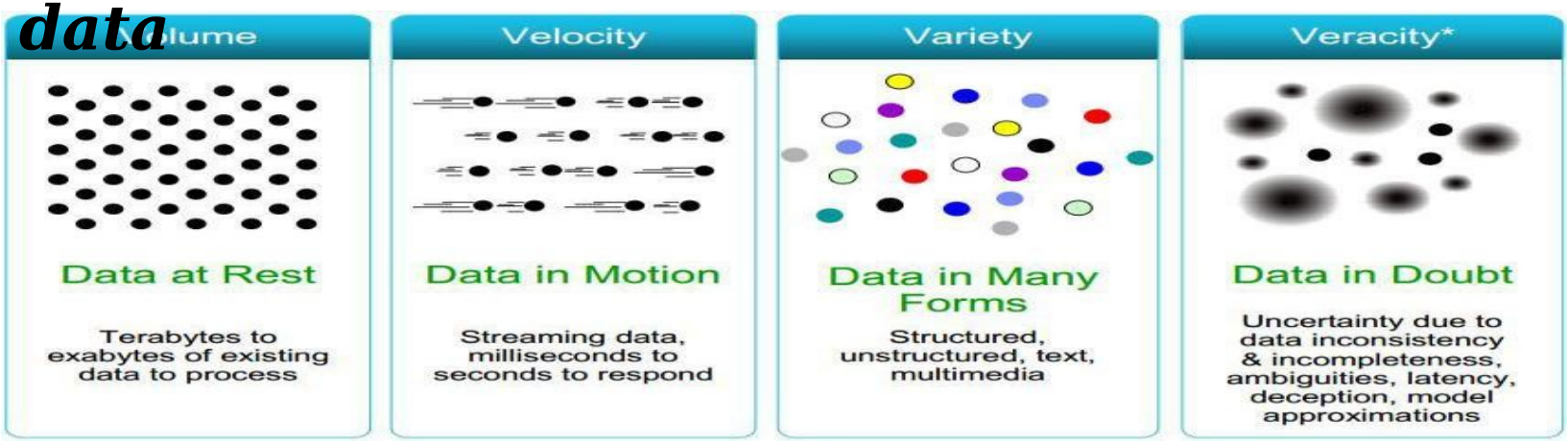
Cont’d

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.

Cont’d

Figure 2.4 Characteristics of big data



2.5.2. Clustered Computing and Hadoop Ecosystem

2.5.2.1.Clustered Computing

- ✓ Because of the qualities of big data, individual computers are often inadequate for handling the data at most stages. To better address the high storage and computational needs of big data, computer clusters are a better fit.
- ❑ 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.

Cont'd...

- **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.
- **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.

Cont’d

- Using clusters requires a solution for managing cluster membership, coordinating resource sharing, and scheduling actual work on individual nodes. Cluster membership and resource allocation can be handled by software like **Hadoop’s YARN** (which stands for Yet Another Resource Negotiator).
- The assembled computing cluster often acts as a foundation that other software interfaces with to process the data. The machines involved in the computing cluster are also typically involved with the management of a distributed storage system, which we will talk about when we discuss data persistence.

2.5.2.2.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.

Cont'd...

- 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:**

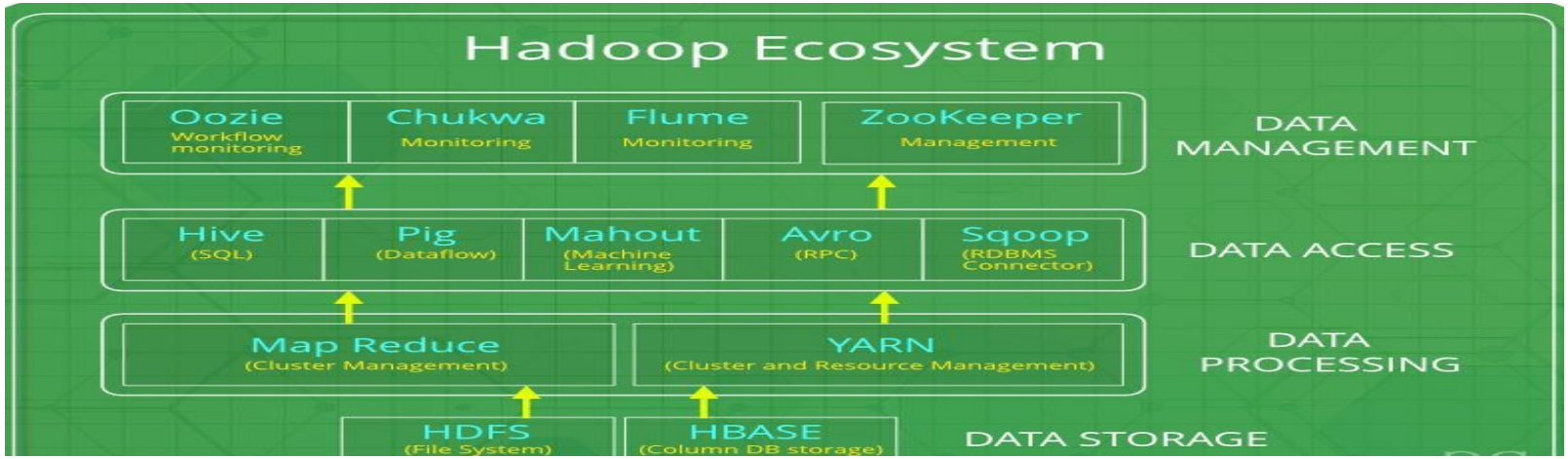
- **data management,**
- **access**
- **processing, and**
- **storage**.

Cont’d

- 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

Cont’d

- **Mahout, Spark MLLib:** Machine Learning **algorithm** libraries
- **Solar, Lucene:** Searching and Indexing
- **Zookeeper:** Managing cluster
- **Oozie:** Job Scheduling



2.5.3.Big Data Life Cycle with Hadoop

2.5.3.1. 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.

2.5.3.2. 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

2.5.3.3. 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.

2.5.3.4. 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.

Thank You!!

