

Addis Ababa University

School of Commerce

Introduction to Emerging
Technologies

Chapter Two- Data science

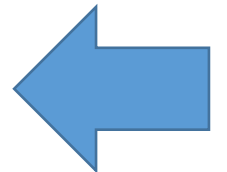
Contents

- Learning outcomes
- An overview of data science
- Data Vs information
- Data processing cycle
- Data types and their representation
- Data value chain
- Basic concepts of big data
- Hadoop ecosystem
- Review questions

Learning outcomes

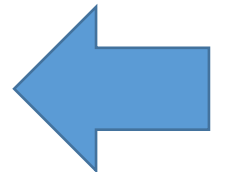
After the successfully completing this chapter, the students can

- Differentiate data and information
- Explain data processing life cycle
- Differentiate different data types from diverse perspectives
- Explain the data value chain
- Explain the basics of big data
- Analyze Hadoop ecosystem components and their use in big data



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 Vs Information

Data:

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

...Data Vs Information

Information:

- Processed data on which decisions and actions are based.
- 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.
- Interpreted data; created from organized, structured, and processed data in a particular context.

...Data Vs Information

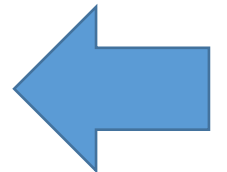
- ♦ Data – a collection of facts made up of text, numbers and dates:

Murray 35000 7/18/86

- ♦ Information - the meaning given to data in the way it is interpreted:

Mr. Murray is a sales person whose annual salary is \$35,000 and whose hire date is July 18, 1986.

Source: internet



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. It has three steps.

Input:

- Data preparation in convenient form for processing. The form will depend on the processing machine.
- For example, when electronic computers are used for data processing, the input data can be recorded on hard disk, CD, flash disk and so on.



Source: Introduction to emerging technology module page 23

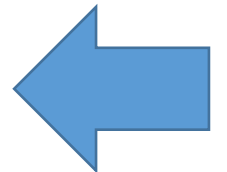
Data Processing Cycle

Processing:

- The input data is changed to produce data in a more useful form.
- 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:

- The result of the 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.

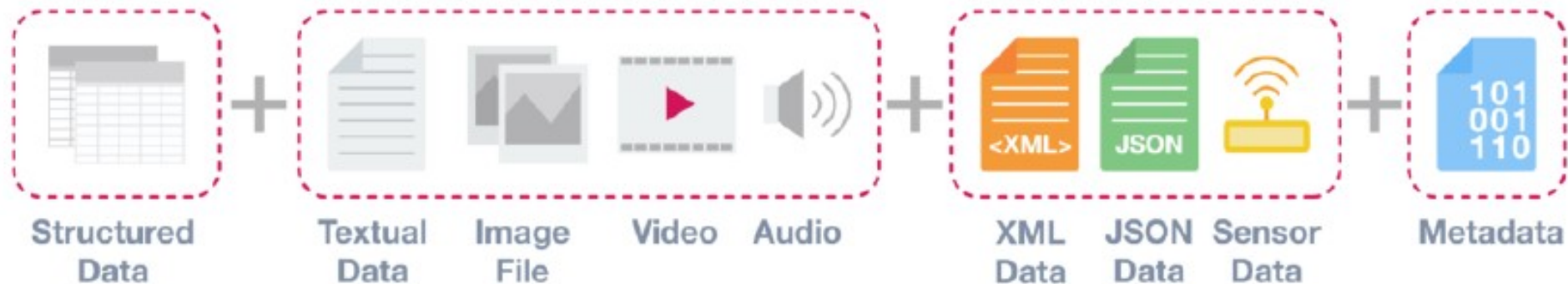


Data types and their representation

1. **Data types from Computer programming perspective:** 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.

E.G int, bool, char, float, double, string

2. **Data types from Data Analytics perspective:** there are three common types of data types or structures: Structured, Semi-structured, and Unstructured data types.



Source: Introduction to emerging technology module page 25

Structured Data

- It conforms to a tabular format with a relationship between the different rows and columns.
- Examples of structured data are Excel files or SQL databases. Each of these has structured rows and columns that can be sorted.

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Source: internet

Semi-structured data

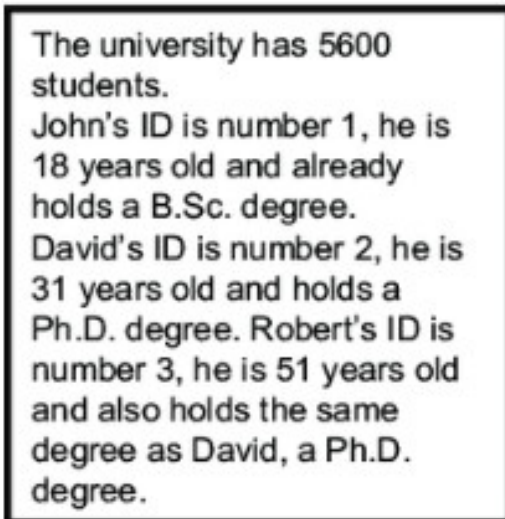
- It is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless, contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.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

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

Source: internet

Unstructured Data

- It 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 which results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structured databases.
- Examples of unstructured data include audio, video files or No-SQL databases.

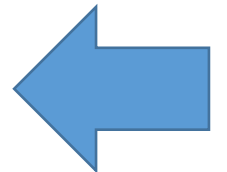


The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Source: internet

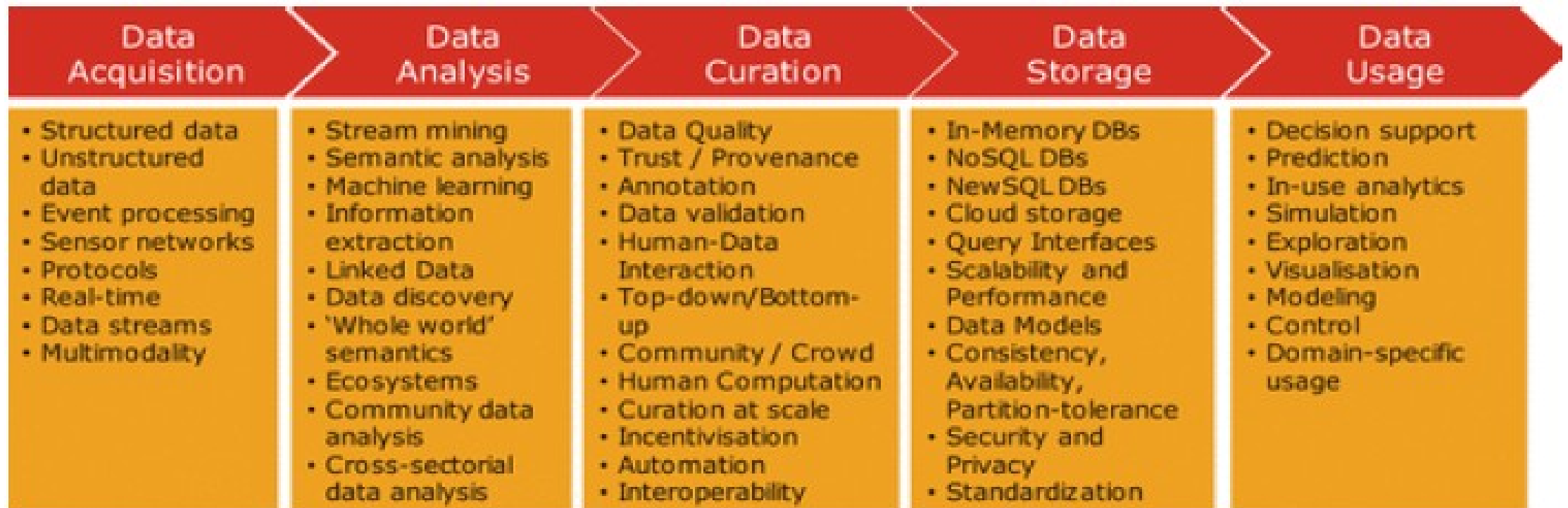
Metadata – Data about Data

- It 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.
- Example, In a set of photographs, metadata could describe when and where the photos were taken.



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.



Source: Introduction to emerging technology module page 26

...Data value Chain

Data Acquisition:

- 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.
- One of the major big data challenges in terms of infrastructure requirements because the infrastructure 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 and dynamic data structures.

...Data value Chain

Data Analysis:

- Concerned with making the raw data acquired amenable to use in decision-making as well as domain-specific usage.
- 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.

...Data value Chain

Data Curation:

- The active management of data over its life cycle to ensure it meets the necessary data quality requirements for its effective usage.
- Its 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.
- Data curators hold the responsibility of ensuring that data are trustworthy, discoverable, accessible, reusable and fit their purpose.

...Data value Chain

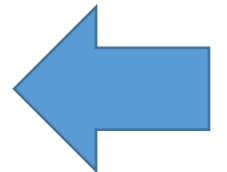
Data Storage:

- 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. However, the ACID (Atomicity, Consistency, Isolation, and Durability) properties that guarantee database transactions lack flexibility with regard to schema changes
- NoSQL technologies have been designed with the scalability goal in mind and present a wide range of solutions based on alternative data models.

...Data value Chain

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.



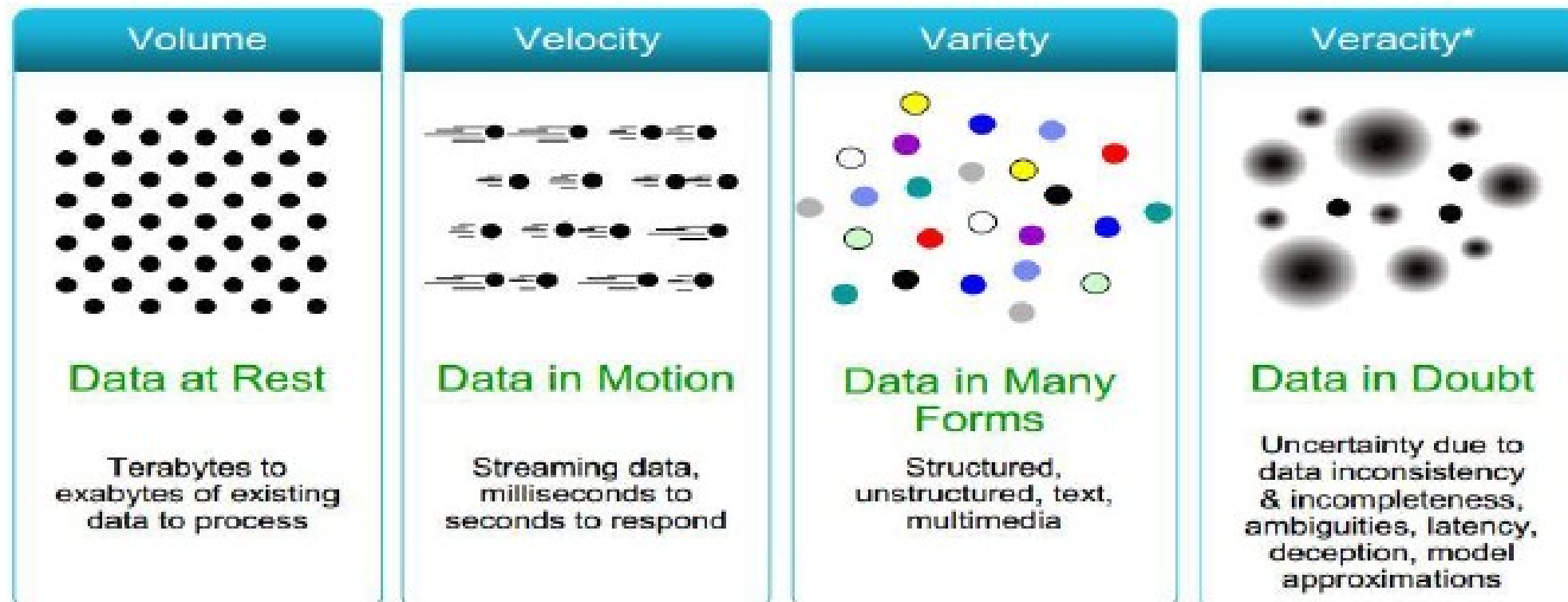
Basic concepts of big data

- **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.
- E.g.



Basic concepts of big data

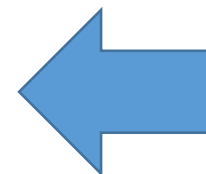
Big data is characterized by 3V and more:



Source: Introduction to emerging technology module page 29

Clustered Computing

- Individual computers are often inadequate for handling the big data at most stages.
- To address the high storage and computational needs of big data, computer clusters are needed.
- Big data clustering software combines the resources of many smaller machines, seeking to provide a number of benefits:
 - **Resource Pooling** ⇨ combine available storage space, CPU, ...
 - **High Availability** ⇨ fault tolerance and availability
 - **Easy Scalability** ⇨ expansion in resource requirement without expanding the physical resources on the machine
- The good example of clustering software is **Hadoop's YARN**



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.
- Gives the massive data storage facility, enormous computational power and the ability to handle different virtually limitless jobs or tasks.
- The four key characteristics of Hadoop are:
 - Economical ⇨ ordinary computers can be used for data processing
 - Reliable ⇨ stores copies of data on different machines (resistant to HW failure)
 - Scalable ⇨ expand horizontally or vertically by adding few extra nodes
 - Flexible ⇨ store as much structured and unstructured data as you need

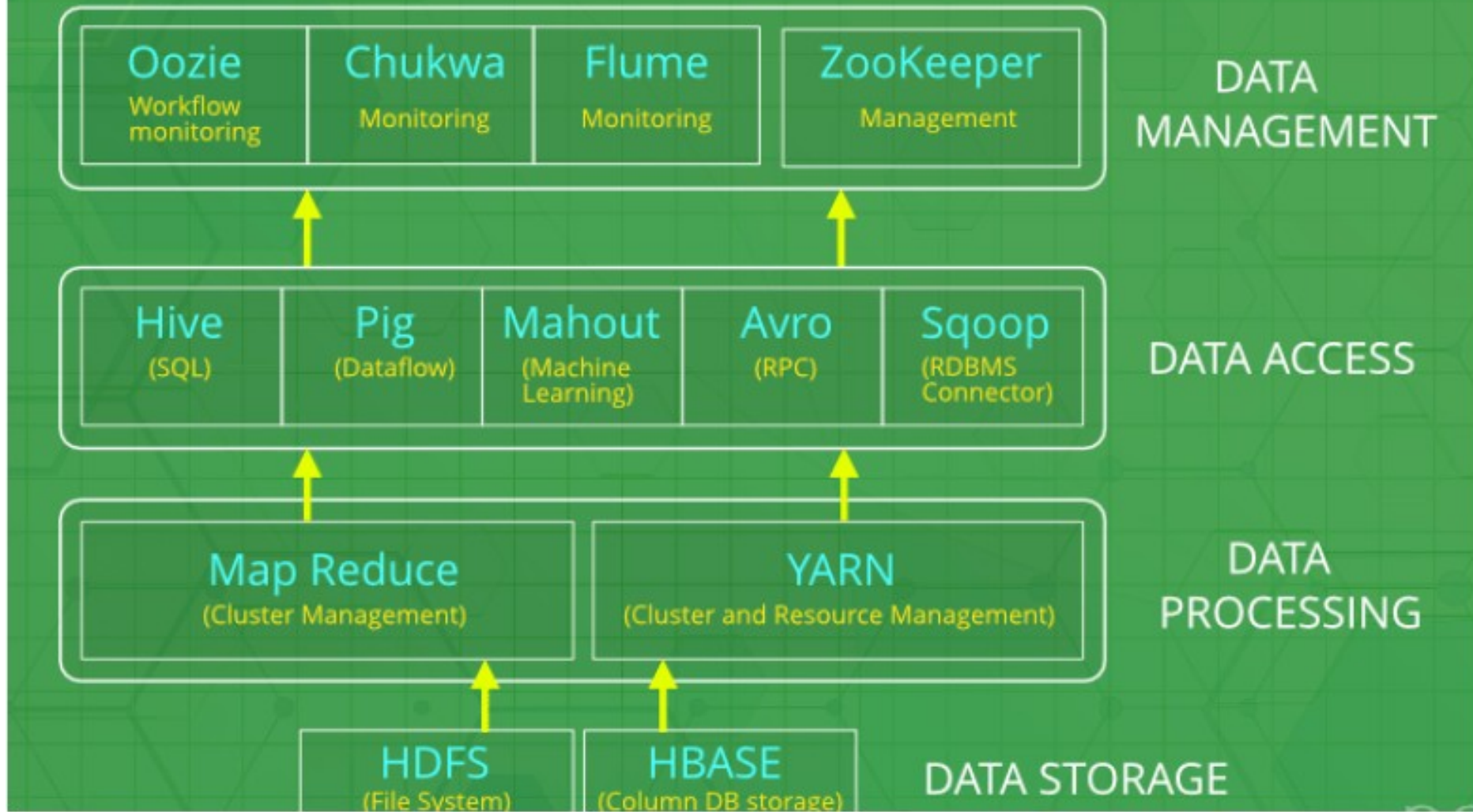
...Hadoop and its Ecosystem

Hadoop Ecosystem has evolved from its four core components:

1. Data management,
2. Data access,
3. Data processing, and
4. Data storage.

It is continuously growing to meet the needs of Big Data.

Hadoop Ecosystem

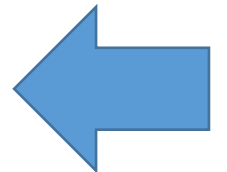


Source: Introduction to emerging technology module page 31

Big Data Life Cycle with Hadoop

Has 4 stages:

1. **Ingesting:** transferring data into to Hadoop from various sources such as relational databases, systems, or local files. Sqoop transfers data from RDBMS to HDFS
2. **Processing:** 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.
3. **Computing and analyzing:** data analyzation using processing frameworks such as Pig, Hive, and Impala. Pig converts the data using a map and reduce and then analyzes it.
4. **Visualizing:** accessing the result, performed by tools such as Hue and Cloudera Search.



Review Questions

- Briefly explain data Vs information?
- Discuss data and its types from computer programming and data analytics perspectives?
- Briefly explain each steps of data value chain?
- List out and discuss the characteristics of Big Data?
- What is Hadoop system? What is it used for?

END!!