

Emerging Technologies

chapter 2

More Quick Notes, Telegram: @campus_handout / https://t.me/campus_handout

Chapter 2: Data Science

2.1. An Overview of Data Science

- Data science
 - multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured, semi-structured and unstructured data.
 - is much more than simply analyzing data.
 - one of the most promising and in-demand career paths for skilled professionals
- Data scientists need to be
 - curious and result-oriented
 - have exceptional industry-specific knowledge
 - communication skills
 - strong quantitative background in statistics and linear algebra
 - programming knowledge focus on Data:
 - ✓ warehousing
 - ✓ mining
 - ✓ modeling

2.1.1. What are data and information?

❖ Data

- unprocessed facts and figures
- represented with the help of characters
- ❖ information
 - processed/ interpreted data
 - base for decisions and actions

2.1.2. Data Processing Cycle

- Data processing: re-structuring or re-ordering of data
- Data processing consists three steps
 1. Input
 2. Processing
 3. Output

2.3 Data types and their representation

2.3.1. Data types from Computer programming perspective

- data type is simply an attribute of data that tells the compiler or interpreter how the programmer intends to use the data.
- Common data types include:
 - 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
 - Alphanumeric strings(string)- used to store a combination of characters and numbers
- data type defines:
 - ✓ operations that can be done on the data
 - ✓ meaning of the data
 - ✓ the way values of that type can be stored

2.3.2. Data types from Data Analytics perspective

- there are **three common types of data types/ structures**:
 - 1. Structured Data**
 - adheres to a pre-defined data model
 - **straightforward to analyze**
 - conforms to a tabular format with rows and columns
 - **Example: Excel** files or **SQL** databases
 - 2. Semi-structured Data**
 - also known as a **self-describing structure**
 - form of structured data that does not conform with the formal structure of data models: relational databases, tables
 - contains tags or other markers
 - **Example: JSON** and **XML**
 - 3. Unstructured Data**
 - information that **either does not have a predefined data model or is not organized in a pre-defined manner**
 - typically text-heavy
 - **Examples: audio, video files**, No-SQL databases

Metadata – Data about Data

- this is not a separate data structure
- provides additional information about a specific set of data
- Example metadata provides fields for dates and locations for a photograph taken

2.4. Data value Chain

- describe the information flow within a big data system
 - I. Data Acquisition**
 - ✓ process of **gathering, filtering, and cleaning data** before it is put in a data warehouse or any other storage

- ✓ major big data challenges in terms of infrastructure requirements

II. Data Analysis

- ✓ making the raw data acquired **amenable to use** in decision-making as well as domain-specific usage.
- ✓ Involves: **exploring, transforming, and modeling data**

III. Data Curation

- ✓ the **active management** of data over its life cycle to **ensure** it meets the necessary **data quality** requirements for its effective usage
- ✓ ensuring that data are **trustworthy, discoverable, accessible, reusable** and **fit their purpose**

IV. 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**
- ✓ NoSQL technologies present a wide range of solutions based on alternative data models.

V. 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

2.5. Basic concepts of big data

- a blanket term for the **non-traditional strategies and technologies** needed to gather, organize, process, and gather insights from **large datasets**
- data that **exceeds the computing power** or storage of a **single computer**

2.5.1. What Is Big Data?

- large and complex data that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- Large dataset
- Big data is characterized by 3V and more:
 1. Volume - large
 2. Velocity - live streaming or in motion
 3. Variety - In many different forms from diverse sources
 4. Veracity - trust the data? How accurate is it?

2.5.2. Clustered Computing and Hadoop Ecosystem

2.5.2.1. Clustered Computing

- Big data clustering software combines the resources of many smaller machines
- provide a number of benefits:
 - ✓ Resource Pooling - Combining the available storage, CPU, and memory
 - ✓ High Availability - provide varying levels of fault tolerance and availability guarantees
 - ✓ Easy Scalability - easy to scale horizontally by adding additional machines

2.5.2.2. Hadoop and its Ecosystem

- It is an open-source framework allows for the distributed processing of large datasets across clusters of computers
- four key characteristics of Hadoop are:
 1. Economical
 2. Reliable - stores copies of the data on different machines and is resistant to hardware failure.
 3. Scalable - easily scalable both, horizontally and vertically
 4. Flexible - you can store as much structured and unstructured data as you need

- Hadoop has an ecosystem that has evolved from its four core components:
 - ✓ data management – e.g Zookeeper
 - ✓ access – e.g PIG, HIVE
 - ✓ processing – e.g YARN
 - ✓ storage – e.g HDFS
 - It is continuously growing to meet the needs of Big Data
- 2.5.3. Big Data Life Cycle with Hadoop
- 2.5.3.1. Ingesting data into the system
- data is ingested or transferred to Hadoop from various sources
- 2.5.3.2. Processing the data in storage
- data is stored and processed
- 2.5.3.3. Computing and analyzing data
- data is analyzed by processing frameworks
- 2.5.3.4. Visualizing the results
- analyzed data can be accessed by users

