SNS COLLEGE OF TECHNOLOGY

Coimbatore-35.
An Autonomous Institution

COURSE NAME : 19CST203 - DATA ANALYTICS

II YEAR /IV SEMESTER

UNIT 1 - INTRODUCTION

Brain Storming

1. What is Data?
2. What is Big data ?
3. What is Data science ?

What is a Data?

- **Data** is the piece of information or facts or figures that can be stored and retrieved accordingly when needed.

Introduction

- **Analytics** The science that analyze crude data to extract useful knowledge (patterns) from them.
- This process includes - data collection, organization, pre-processing, transformation, modeling and interpretation.
- Data mining - to extract the knowledge or data.

Big data

- ❖ **Big Data** refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods.
- ❖ Top 5 Big Data Tools [Most Used in 2022]
- Apache Storm.
- MongoDB.
- Cassandra.
- Cloudera.
- OpenRefine.

Big data

❖Big data is classified in three ways:

- Structured Data.
- Unstructured Data.
- Semi-Structured Data

❖ A natural taxonomy that exists in data analytics is:

- • Descriptive analytics: extract patterns
- • Predictive analytics: extract models

❖ Method or technique - systematic procedure

❖ Algorithm - step-by-step set of instructions

Big data

❖**Example:** The method to obtain the average age of my contacts uses the ages of each (we could use other methods, such as using the number of contacts for each different age). A possible algorithm for this very simple example is

❖**Algorithm :** An algorithm to calculate the average age of our contacts

```
INPUT: A: a vector of size N with the ages of all contacts.
$S \leftarrow 0$ ⊳ Initialize the sum S to zero
for $i = 1$ to $N$ do ⊳ Iterate through all the elements of A.
    $S \leftarrow S + A_i$ ⊳ Add the current (ith) element of A to S.
$A \leftarrow S/N$ ⊳ Divide the sum by the number N of contacts.
return(A) ⊳ Return the result, i.e. the average age of the N contacts.
```

Big Data Architectures

- **F**irst techniques using clusters - MapReduce.
- MapReduce is a programming model that has two steps: map and reduce.
- Famous implementation of MapReduce is called Hadoop.
- chunks