(DA) Data Analytics

05/01/2023

Data:-

↳ Data → Information → knowlage → Actionable insights

↳ since the mid-19 cars, people have used the word data to mean computer information that is transmitted or stored

↳ Data means stored in a computer and the data are various types data is important because the

↳ it is interpreted, by a human or machine to derive meaning

↳ prefer homogeneous source as well as heterogenous source

↳ need of the hour is to understand, manipulate and take data, for analysis to derive valuable insights

Importance of Data:

↳ searching data and analise data and get concluded to user preference and create and send notification (ex:- chatbots)

↳ Boost Company Revenue

↳ customer decision

Characteristics of Data :-

↳ Data → Composition, condition, context

↳ Deal with the structure of Data analysis use

Deal with wheres has been generated

09/01/2025

Human vs. Machine Readable Data

↳ Human-readable refers to the information that can be read and interpreted by a human, it required a person to interpret it, that information is human-readable.

↳ Machine readable information is a set of information for manipulating Data.
ex: CSV, JSON, XML.

Classification of Digital Data

1) structured Data (10-15%)
2) semi-structured Data (15-20%)
3) unstructured Data (80-85%)

} Data (World)

1) Structured Data

↳ stored in table format (pattern makes it easier for any person to sort, read & process Data.

↳ Any type of operation (using SQL)

↳ ordered predefined format, stored table format, reside in a fixed filed within a record field, attributed mapped, used to query and operations against predetermined data types.

Structured Data

↳ Relational Database

↳

↳

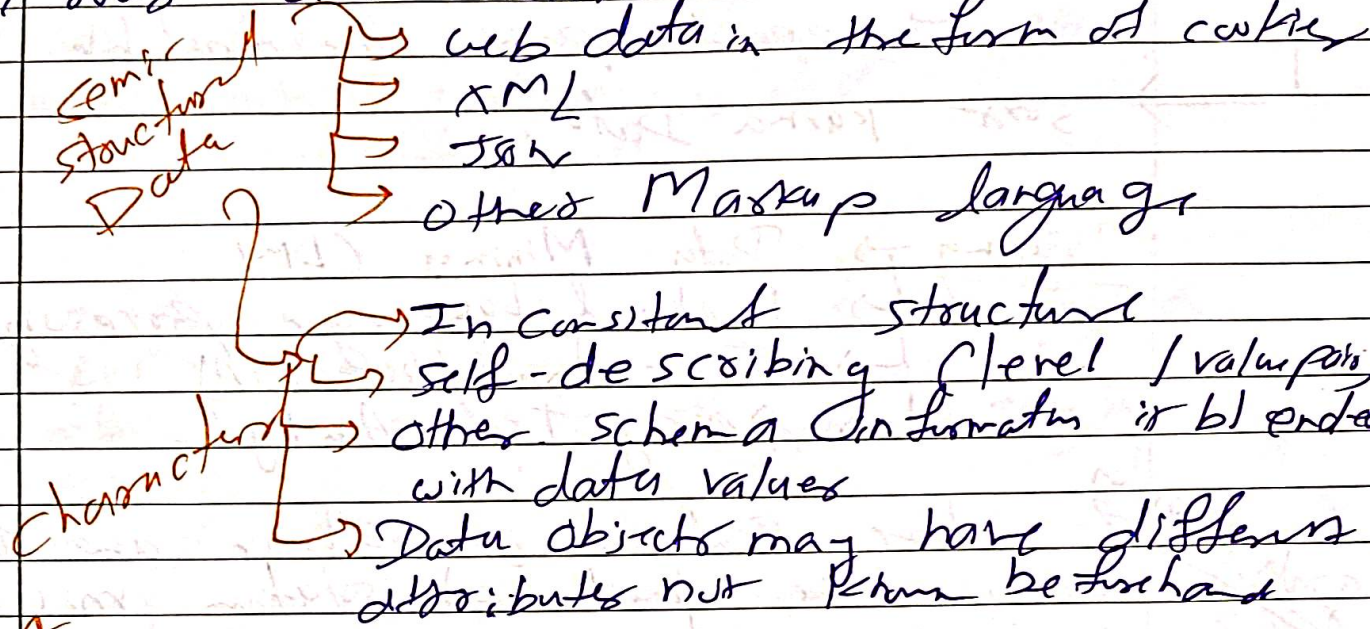
↳ Insert / update / delete, security, index, scalability, transaction processing.

Initial to final state

09/01/2024

Semi-structured Data:-

↳ schema-less or self-describing structure
structure
↳ ex:- email, XML, markup language like HTML
↳ rows and columns



12/01/2024

Unstructured Data:-

↳ Data may have not logic, logical organizing
↳ so :- Unstructured Data
↳ Typically consists of metadata
i.e. additional information related to data

unstructured data → text both internal and external to org.
↳ Social media data
↳ Body of email
↳ chats, Text messages
↳ media data
↳ Images, audios, videos.

02/01/2024

Challenges associated with Unstructured Data

- Identify Unstructured Data
- sorting
- storage
- cost
  - cost management
  - depend upon the available Data

Sort karna Data.

Dealing → Data Mining (DM)
- Natural Language processing (NLP)
- Text Analytics (TA)
- Noisy Text Analytics

Data Mining → store Data in one storage and analysis karne ke bad ek solution mil jaye hai us domain mai.

Natural language processing → e.g. Google Assistance

Text Analytics → Unstructured Data

Big Data → 3V → High Volume → velocity → veracity

why → structured Data + Semi-structured Data + Unstructured Data = Big Data

12/01/2024

Elements of Big Data :-

Volume, variety, velocity, veracity

More data → More accurate analysis → Greater confidence in decision making

Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offering etc.

challenges of Traditional system

↳ outstructure data --→ (constructure Data)
↳ RDMS using Relation (No Deferate tabhi associate Data → unstructure Data)
↳ use ETL process (extract, transfer and load)
↳ Active parallelism with the help hardware MPP
↳ independent support of aggregated summer Data.

conditions:-

↳ Data challenges → Volume, velocity, variety
↳ process challenges → Data Discovery
↳ Management challenges → Scalability

→ capturing Data, Aligning Data, Transform Data suitable form, Modeling data
→ security, piracy, Governance, Ethical use

*) Web Data:-

↳ Available public Domain
↳ documents pdf, doc, docx, Plain text

15/01/2024

- Best - known source
- Data unstructured and is appropriate.

* parallel computing Vs Distributed computing

parallel computing	Distributed computing
1) shared memory sys.	Distributed Memory sys.
2) Multiple processor share a single bus and memory unit	Autonomous computing nodes connected via network
3) processors in order of Tbps	GPPs
4) Limited Scalability	Better scalability and cheap

Distributed computing in local network (called cluster computing), Distributed computing in wide-area network (grid computing)

* EDW, OLTP, MPP

- Enterprise Data Warehouse
- Online transcational processing
- Messively parallel processing

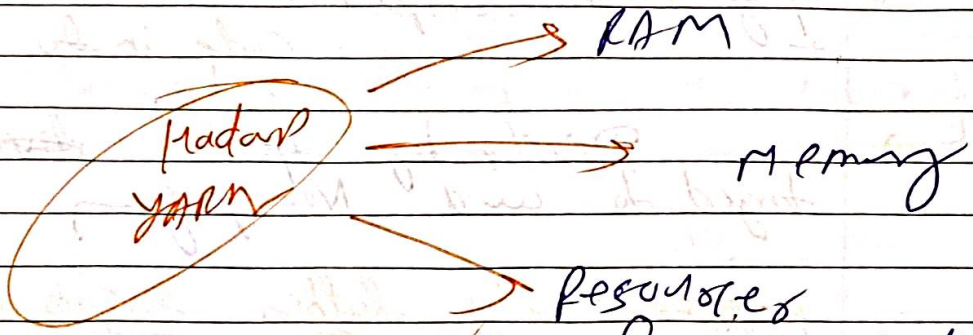
* Hadoop

Store Big Data using Distributed architecture

- store Huge Amount of Data and process
- Computing power (Distributed power) / Provider work
- Topic: computing power, Fault tolerance, Flexibility, Low cost, scalibility

18/01/2024

3) YARN: (Resource Management Unit)
Jiska use karke Data ko process kiya jata hai



↳ Resource Negotiator. Jo handle the cluster of Node and acts as Hadoop's resource managers unit. YARN allocate RAM, memory, & other resources of yarn

↳ Resource Manager (Master)
↳ Node Manager (Slaves)

MAP Reduce: (Data processing)

1) The Mapreduce perform the process of job dataset in a distributed parallel Manner.

2) The MapReducer consist of 2 steps
↳ Map → to divide the data in a determine format
↳ Reduce → and combines them into key value pair

↳ To collect intermediate Results

23/1/2024

Apache Pig :-

It consist of
1) Pig latin :- This is the language of script
2) Pig latin compiler :- this convert pig latin code into executable code

Develop by ~~apache~~ ~~Pig~~ yahoo research
target to used Non programmer

~~Pig Latin~~ Pig Python Scripts :-

Grunt Shell

Pig server

Parser
Optimizer
Compiler
Execution Engine
↓
MAP Reduce
↓
HDFS.

Apache Pig

Hadoop

Hive :-

↳ Hive Command line
↳ JDBC/ ODBC driver

Sparkly open source

↳ All Data Process
↳ Processing

Mahout :-

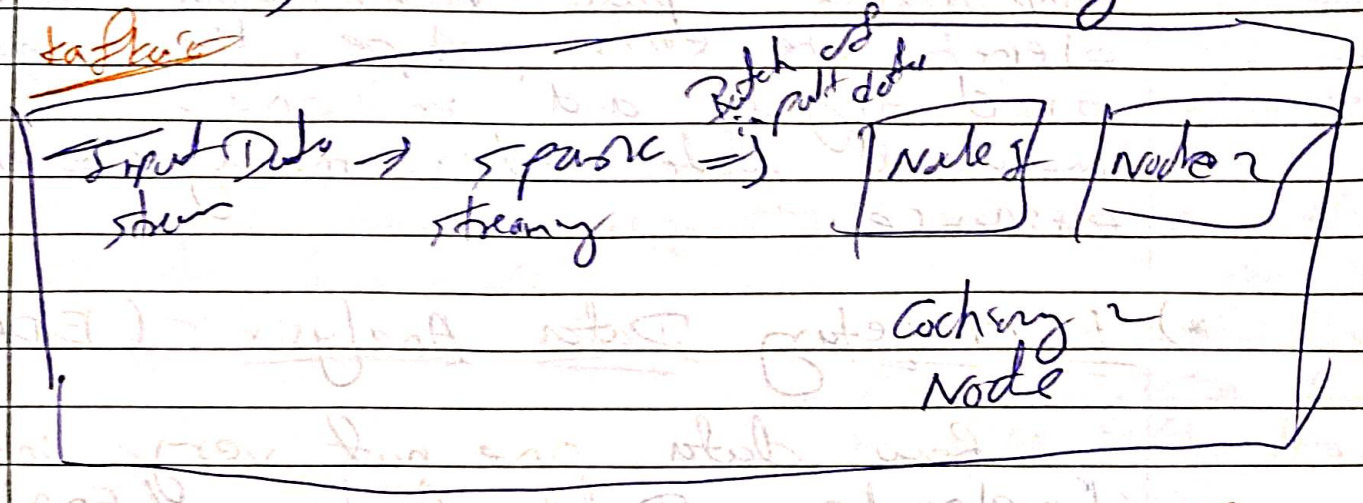
Mahout builds machine learning algorithm environment

↳ provide platform

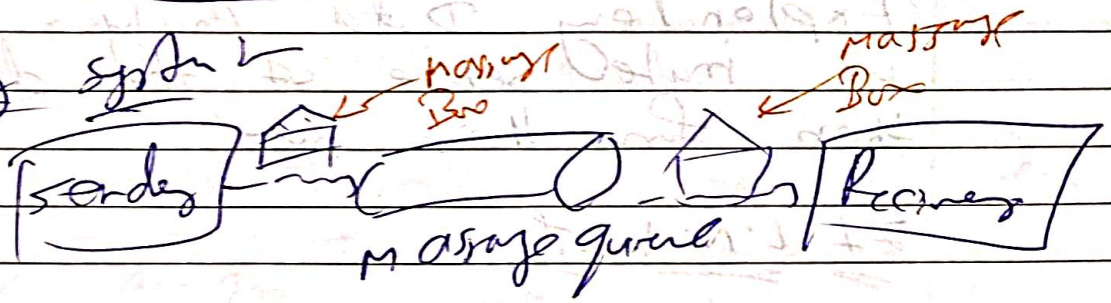
23/10/2024

Ambari :- ↳ Distributed Hadoop cluster
↳ How many applications running

Kafka :-



∴ Messaging system :-



» Storm :- Real time streaming of Data processing

✓ Ranger :- Security Features implemented in Ranger

✓ Hbase Knox :- proxy services
Authentication services
Client services

» Oozie :- 1) workflow engine
2) coordinator engine

» Data Explecation :-
Data explanation is the first step in Data analysis involving the use of data visulization tool and statistical to uncover data set characteristics

and initial pattern.

Importance:- Data visualization tools and elements like color, shape, line, graph and image aid in challenging & recognizing flow of data under many circumstances existence.

*) Exploring Data Analysis:- (EDA):-

Raw data are not very informative Exploratory Data Analysis (EDA) in how we make sense of the data by converting them from their raw.

It Limits of:-

i) Analyzing and summarizing the new data
ii) Discovering important features and patterns in the data and any striking deviations from those patterns
iv) Interpreting our findings in the context of the problem

*) Sampling:-

→ Sampling is a technique of selecting individual members or a subset of the population to make a statistical conclusion in the base of evidence from them and estimate characteristics of the whole population

Sampling Method:-

probability sampling	Non-probability sampling
* Any elements can be chosen randomly from the population It deals with choosing the sample randomly	Every element is not the subjective judgement of the person on basis of their past experience & knowledge rather than random selection

ii) Non-sampling Methods:-

Every elements will be chosen in the subjective judgements from the the person on the basis of their past experience & knowledge with the random selection

ob| ont

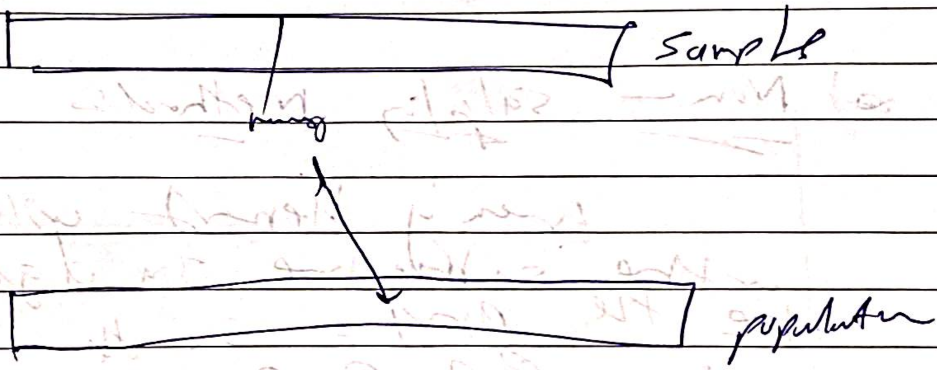
After checking

12/02/2024

Estimation:-

- it is often of interest to learn about the characteristics of a large group of elements such as individuals, households, building, product.
- Data from the sample are then used for developing estimates characteristic of the larger population.

Point Estimation



Point Estimation Conditions

- Most often, the methods of finding the parameters of large population are through e.g. using finding the average age of humans taking into length, it is impossible to collect the exact age of every human being in the world.

Interval Estimation

- Consider Interval Estimation cont — suppose a student measuring the boiling temperature of a certain liquid

12/04/2024

Observed the reading (in degree, Celsius) 102.8, 102.7, 102.1, 103, 104, and 102.4 on a different sample of the liquid. What is the interval estimation for the population mean at a 95% confidence level

mean = 102.82

μ = 102.82 ± 1.96 × (0.984/√5) = 102.92

standard error = (SE) σ/√n = 0.442

Margin of error = Z (σ/√n) = 1.96 × (0.984/2.108) = 0.286