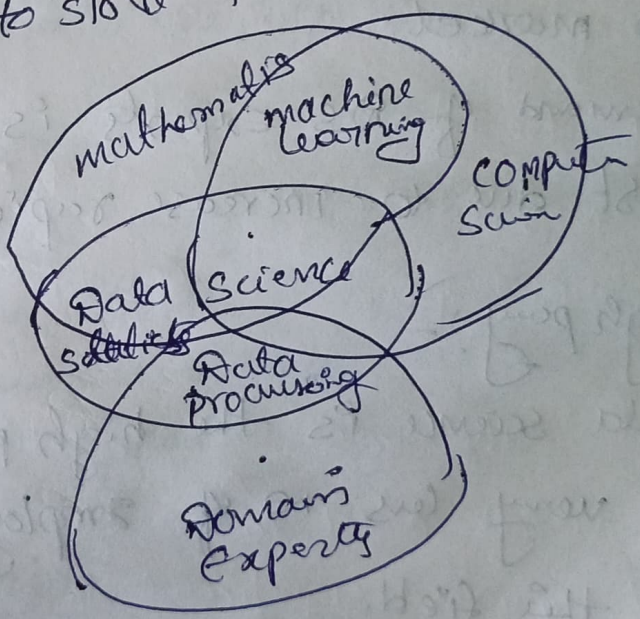
Data Science:-

* Data science is a field that involves extracting insights & knowledge from data using various techniques such as mathematics, statistics, Artificial intelligence & computer science.
* Data science is the domain of study that deal with vast volumes of data using modern tools.
* Data science uses the most powerfull hardware, programming systems & most efficient algorithms to slove the data related problems



Advantages of Data Science:-

* Data Science has become an essential part of any industry today. Every second lots of data is generated from users of fb or any other social N/w sites.

* Because of this Data Science has a no of advantages.
* Some of the advantages are mentioned below.

1) IN Demand
2) High pay
3) Versatile
4) Customize
5) multiple Job options
6. Buseiness Benefits.
7. Decision making

1) In Demand:-

* Data Science is most Demanding field in Job market.
* Demand of DS experts is increasing very fast due to increase rapid data generation.

2) High pay:-

* Data science is the high pay job because of very less no of employees are available in this field.
* The average salary of DS start from 6 to 8 Lakhs per annum.
* There are a lot of opportunities with high pay for DS professionals in india as well as over the world.

3. Versatile:-

* Data science is a versatile field which has many opportunities in different areas.
* It can be used in Health care, Banking & E-commerce industries.

4. Customize products:-

* Data science helps organizations to understand customer requirement & develop products as per customer requirements.
* companies can increase their sale & revenue with the help of ML.

5. Multiple Job options:-

→ It gives large no. of career opportunities in its various fields.

* Some of them are Data scientist, Data Analyst, ML Engineer, Big data Engineer,

6. Business Benfits:-

* Data science helps organizations knowing how & when their products sell best,
* how the products are delivered always to the right place & right time.
* helps to take faster & better decisions

by the organization to improve efficiency & profits.

Disadvantages of Data science:-

1) Data privacy
2) cost
3) Domain knowledge
4) complexity
5) mastery in challenging

i) Data privacy:-

* Data is the core component that can increase the productivity & the revenue of industry by making game-changing business decisions.
* But the Information obtained from the data can be misused against any organization or group of people or committee.

2) cost: The tools used for data science & Analytics can cost a lot to an organization.
* It requires significant investments in tools, talent & infrastructure.

3) Domain knowledge:-

A personal requires a particular domain knowledge to applying Data science.

* you must require the knowledge of the domains where you are applying Data Science.

4.) Complexity:-

* Data science is a complex field. because it includes - mathematics, programming & visualization etc.

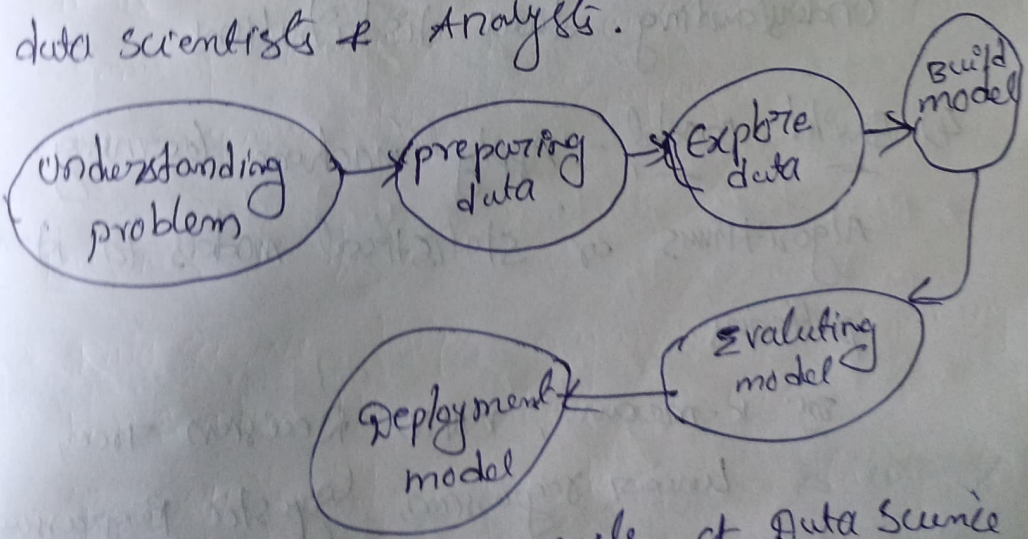
- some times it is a complex task to choose right tools for analysis in Data science.

5) Data Quality:- poor data quality can lead to inaccurate insights.

* Data science process:-

* The data science process is structured / life cycle to extracting insights & knowledge from data.

* It involves a series of steps that help to data scientists & Analysts.



life cycle of Data Science

1. Understanding problem, It is used to understanding problem & Identify a business problem or opportunity.
   * It Grather data from different sources such as social medias, websites etc.

2) preparing data:- → eliminate duplicate data & null values.
   * It is used to clean the data. It ensure data quality, handle missing values & transform data.

3) Explore data:-
   * we have a large amount of organized, high quality data you can conducting an exploratory data Analysis (EDA).
   * use visualization & statistical techniques to understanding data distributions.

4) Build model:- Apply machine learning Algorithms or statistical models to identify patterns.
   ex: K-mean, KNN, Decision Tree, Linear regression, Logistic Regression

5) Evaluation of model:-
   It is used to validate & test the

data to improve performance & accuracy of model.

5. Deployment:-

=> put the model into production & monitor its performance.

* Gather feedback from end-users or stakeholders
* maintain a versioning system to keep track of model iterations/updates.

Facets of data:-

* very large amount of data will generated in data science.
* These data is main categories of ~~data of~~ different types, These types as follows.

1. structured
2. Natural language
3. Graph-based
4. Streaming
5. unstructured
6. machine-generated.
7. Audio, video, & images.

* Structured

-> structured data is a data which is arranged in rows & column format.
-> It helps for application to retrieve & process data easily.

* Structured data is identifiable because it is organized in structure.
* example of structured data is Excel & tables.
* It has predefined format.

Unstructured data:-

* Unstructured data is a data which is doesn't follow a specific format.
* It is very difficult to retrieve required information.
* Unstructured data is not identifiable.
example of unstructured data is text, email, images, videos etc.
* It does not have any predefine format

Natural language:-

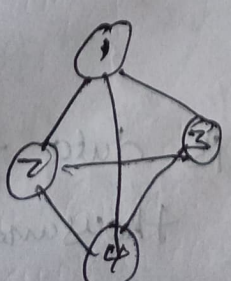
* Natural language is special type of data.
* NLP enables machine to recognize characters, words & sentences. then apply meaning & understand the information.
* NLP helps machine to understand languages as human do.
* NLP is success in entity recognition & Sentiment Analysis.
* It helps machine to understand human languages ex: speech recognization & text analysis.

Machine Generated data:-

* machine generated data is a information that created without human interaction as a Result of a computer process.
* machine data is generated continuously by every processor based data systems & consumer-oriented systems.
* It can be either structure or unstructured.
* machine data is increase due to mobile devices, virtual based servers, desktops web servers etc.

Graph based / network based:-

* graphs are data structures to describes r/nship b/w entities / nodes.
* graph contains collections of vertices is called nodes & Edges is called arcs.
* Graph database is used to store graphical data & required specialized query language such SPARQL.
* It is good choice for recommendation Applications



ex:
→ fraud prevention.

Audio, Image & video:-

* Audio, image & video are data types, that pose specific challenges to a data scientist.
* The audio & video is time based media storage format for sound / moving pictures information.
* The multimedia data is one of the most important source of information & knowledge. The data integration & transformation, Indexing is challenge in data management & Analysis.
* Data science playing an important role to address these challenges. in multimedia data.
* multimedia comes with various forms such as text, image, video, geographic coordinates & pulse waveforms. ~~come~~ with different sources.
* DS can use many instruments covering big data, m2, & data mining solutions to store, handle & analyze such heterogeneous data.

Streaming Data:-

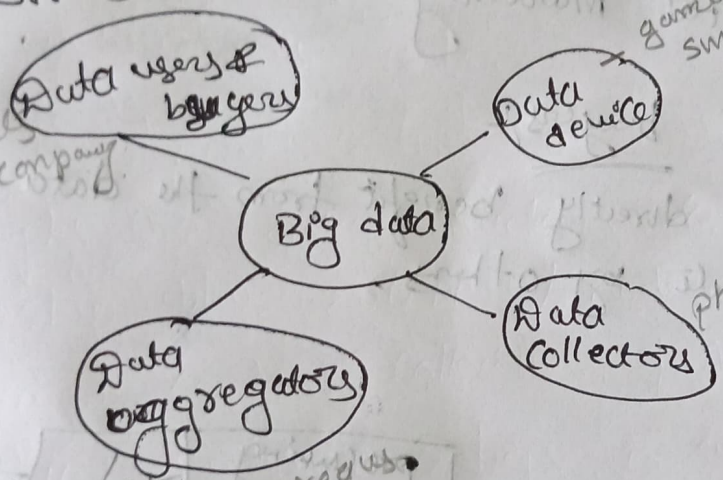
* Streaming data is data that is generated continuously by thousands of data sources
* Size of data is small (KB).

* Streaming data includes a wide variety of data such as log files generated by customers using mobile, web applications, e-commerce, game player & social N/w.

Big data ecosystem:-

Big data is a collection of data that is larger more complex than traditional data.

Big data ecosystem:- Big data ecosystem is a massive volumes of both structured & unstructured data.



Data Device:-

→ It is also known as data source. Gather data from multiple locations & continuously generated new data about this data.

* Each gigabyte of new data created & petabyte of data is created.

Ex: online video games, smart phones data. Retail shopping. etc.

Data collectors:-

* The data collectors used collect data from the devices & users.

ex: Retail stores tracking the path a customer.

Data Aggregators:-

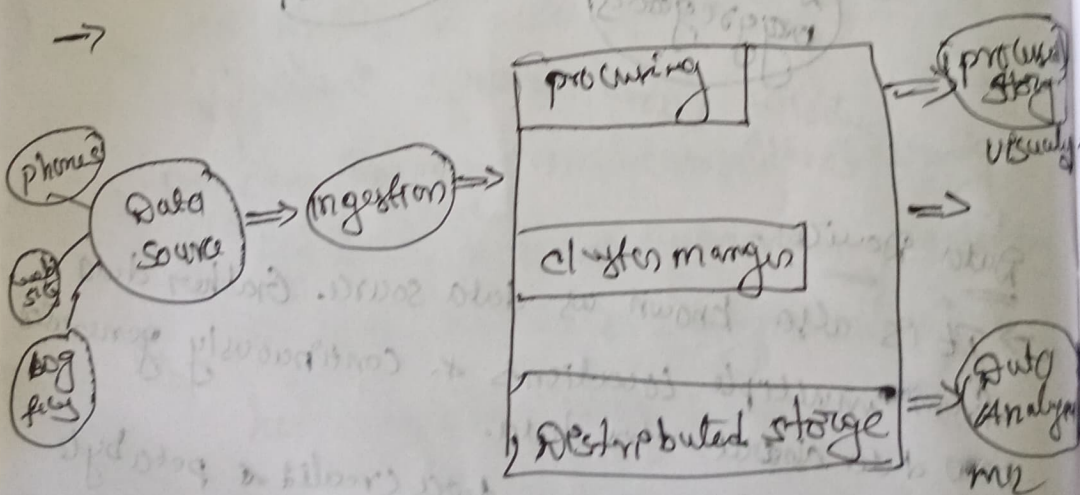
* It is used to make sense of data, i.e, they transform & package the data as products to sell to list brokers for specific ad campaigns.

ex: Digital marketing

Data users & buyers:-

These group directly benefit from the data collections & aggregate by others.

→



Ingestion:-

It means Tools like Apache, kafka, Apache NiFi for collecting & transporting data.

processing:- It use Frameworks like Apache Hadoop, Apach Sparks, these are used to processing

and Analyzing data.

Cluster manager:- It is used to manage data.

Data storage:- It is distributed storage systems like Hadoop Distributed Filesystem (HDFS), Amazon S3.

Data visualization:- Tools like Apache mahout & mllib for ml & DM.

Data Analystics:- Tools like apache Hive, Apache Impala & Apache Drill for SQL based.

Data procuring:-
To procure the data using different tools such as RDBMS.