ISSN 0975-5829 (Print) ***www.delving.in***

Data Science: A Decade Survey

Yuvraj Panchal
Department of Computer Science and Engineering
Acropolis Institute of Technology and Research, India
yuvrajpanchal3176@gmail.com

Abstract — Data science is a disruptive force that enables organizations to critically evaluate widely accepted theories and models and make well-informed decisions across a wide range of industries. Inference, a critical step in deriving conclusions based only on known information, is at the core of data science. Through rigorous examination of enormous and diverse datasets, data scientists uncover patterns, trends, and correlations that provide insight into practical concepts. Inference in data science is like peering through a lens of various strategies. To extract meaning from complex data, researchers employ statistical techniques and machine learning algorithms. By using this power, businesses can improve strategy, lower risks, and confirm or refute prior beliefs. Furthermore, data science helps companies to innovate, enhance operations, and create superior products and services. Data science influences strategic decision-making and has important ramifications for operational applications. By leveraging the wealth of available data, organizations may identify customer preferences, predict market trends, and tailor offers to meet evolving needs. Additionally, data science aids firms in cutting costs, optimizing resource allocation, and maximizing returns on investment. Furthermore, data science promotes an organizational culture that places a high value on continuous learning and improvement. Data-driven businesses encourage experimentation, creativity, and fast adaptation to changing market conditions. Ultimately the focus that data science places on inference emphasizes how crucial it is as a cornerstone of modern decision-making, empowering companies to navigate an increasingly complex and interconnected business environment with clarity and confidence.

Keywords: *Data Analytics, Data Warehouse Architecture, Big Data, DBMS.*

I. INTRODUCTION

A computer's memory was limited in the past, but as technology advanced, so did the need for more memory. Because of the overwhelming demand these days, traditional methods cannot handle the data in an efficient manner. Nowadays, a huge amount of data is produced every day from social media, shopping websites, weather predictions, airline data sets, medical data sets, and other sources. We can observe the evolution of technology from the telephone to the iOS mobile phone, which is improving human intelligence; aside from that, we still use a large desktop computer with a floppy disk to process data. However, we can now store our data on the cloud. In a similar vein, cars have also evolved into smart vehicles that can sense data about their surroundings, the weather, traffic volume, and other factors. These technological advancements generate enormous amounts of data every day. In the current tech market, this is one of the fastest-growing industries. And as the future world grows more digitally advanced every day, this will continue. Furthermore, the facts undoubtedly have the power to build a

brand-new future. Data science is the study of the reasoning behind the data or the methods used to manipulate it.

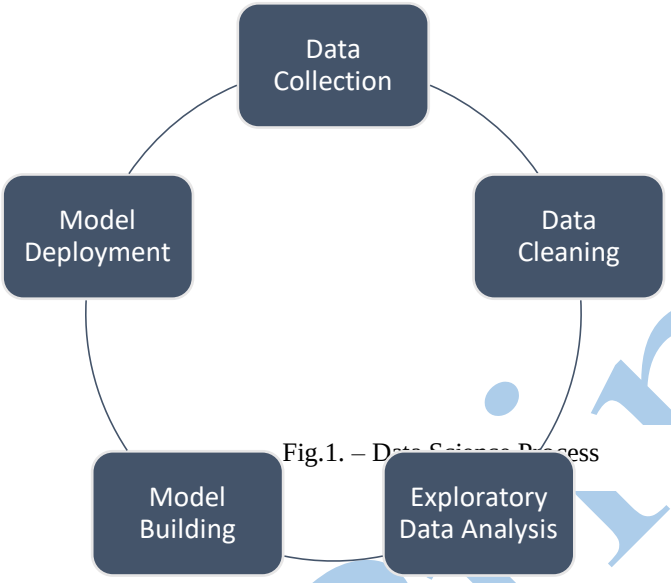

Fig.1. – Data Science Process

Fig 1 Data science process

II. DATA SCIENCE PROCEDURE

A data science framework provides a dependent approach to records challenges, beginning with a clear definition of the problem and accumulating information through strategies which includes aim putting, surveys or internet scraping for analysis, but much of the data encountered in practice is unstructured [1]. Once the facts is ready, the focus shifts to identifying hidden patterns and expertise how variables relate to goal effects Machine gaining knowledge of algorithms are specifically beneficial at this degree, as they can perceive complicated relationships without slumbering revealed without delay. After analysis, models are developed to solve specific problems and make predictions. Applying these models to the real world is critical for effective decision-making. It is important to continuously monitor the validity and relevance of the models by updating them as necessary. This comprehensive process—from defining the problem to implementing and maintaining the model ensures that insights from the data can better inform decisions and produce meaningful results [2].

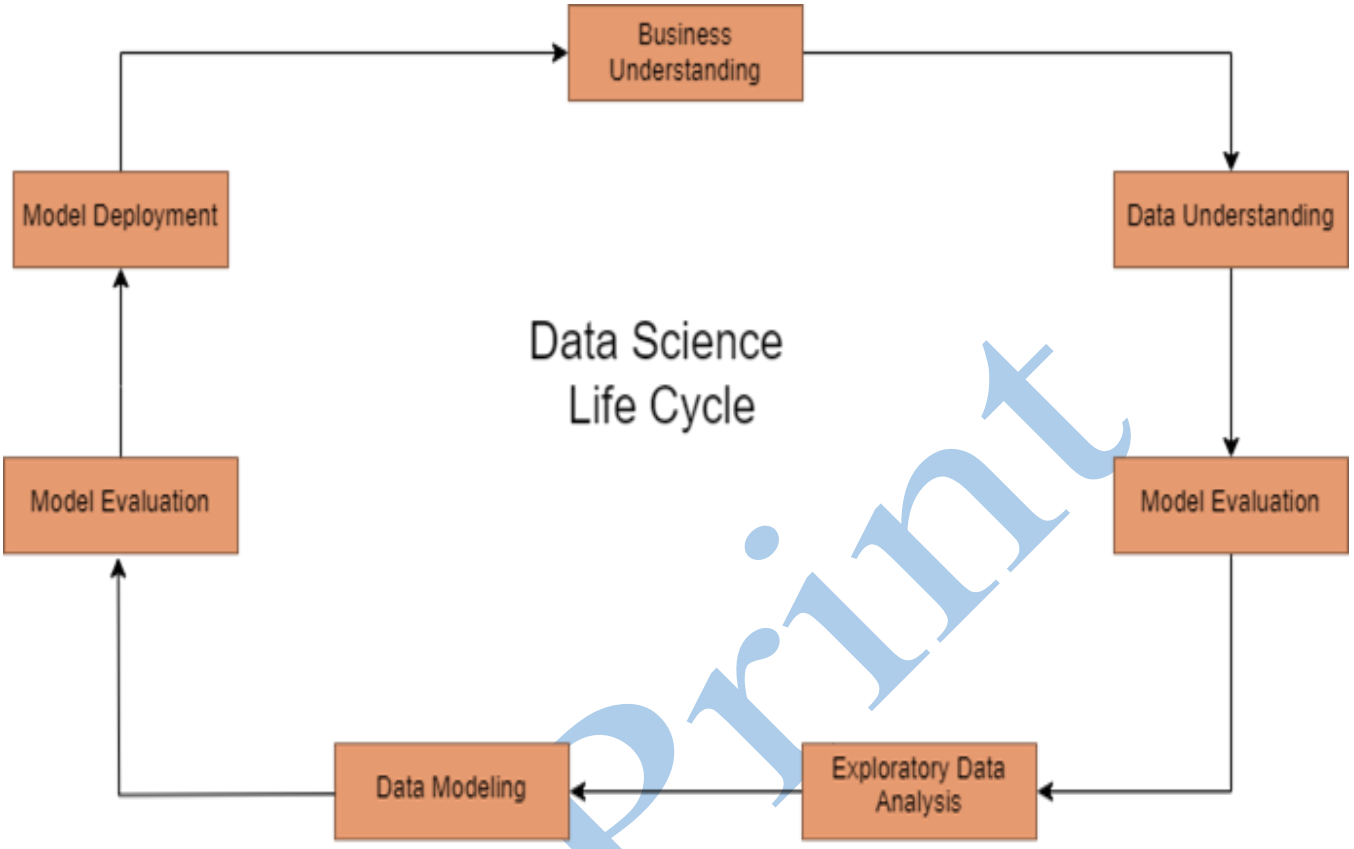

Fig. 2. – Data Science Life Cycle

III. FACTORS OF DATA SCIENCE PROCEDURE

Data analysis to get a primary concept of the statistics and patterns which are available in it. This offers us a route to paintings on if we want to use some complex evaluation methods on our data. The descriptive statistics and correlation and covariances among capabilities of the dataset assist us get a higher understanding of ways one thing is related to the opposite in our dataset. When we deal with a huge amount of records then we ought to ensure that the information is kept secure from any online threats and it is simple to retrieve and make modifications in the records. To ensure that the records is used correctly Data Engineering performs a vital role. Machine Learning has opened horizons which had helped us to build one of a kind superior programs and methodologies so, that the machines turn out to be extra efficient casting off the requirement of heavy human labour and time. Deep Learning that is a part of Artificial Intelligence and Machine Learning has high computing electricity and may manner huge corpus of data effectively. It is the examine of the collection, analysis, interpretation, presentation, and business enterprise of information. A facts scientist is higher statistician than a software program engineer and higher software engineer than a statistician [3]. Python and R are one of the maximum widely used languages by means of Data Scientists.

The number one cause is the number of applications to be had for Numeric and Scientific computing. A records scientist has to Extract records from multiple records resources like MySQL DB, MongoDB, Google Analytics after which remodel it for storing in a right format or structure for the purposes of querying and evaluation. Finally, he has to load the records in the Data Warehouse, wherein he's going to examine the facts. This Extract Transform and Load (ETL) ability is a need to for statistics technological know-how professional [4].

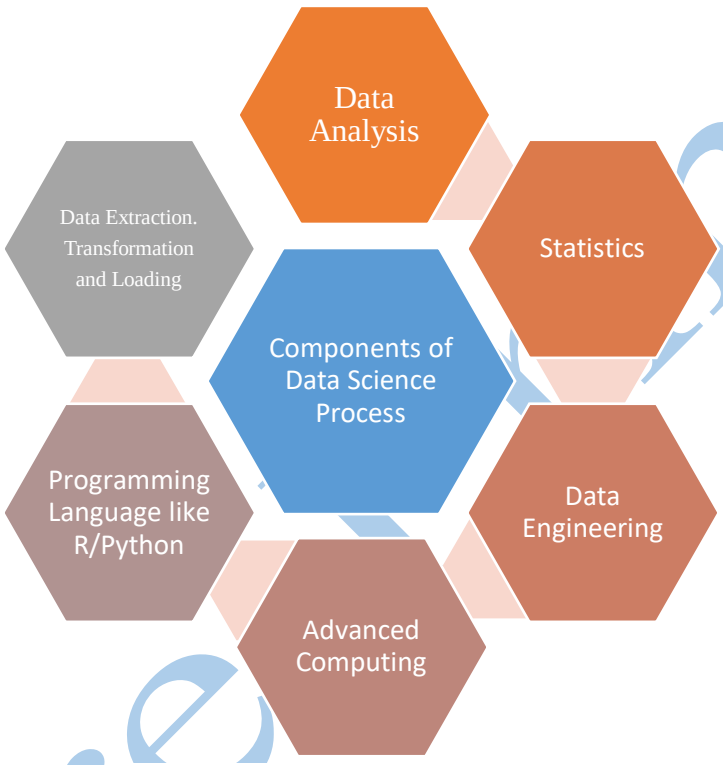

Fig. 3. - Components of Data Science Process

IV. SETS OF TOOLS

Data technology gear are utility software program or frameworks that assist data technological know-how professionals to perform diverse facts technology duties like evaluation, cleansing, visualization, mining, reporting, and filtering of data. Each of those equipment comes with a hard and fast of some of those usages [5]. Some of the gear which are very popular on this domain of Data Science like Microsoft Power BI is a effective enterprise intelligence suite and one of the maximum advocated statistics science equipment of 2023 that assist create lovely information reviews and visualization offerings for both individuals and teams. You can integrate it with different Microsoft statistics technology tools like MS. Excel, Azure Synapse Analytics, Azure Data Lake, and so on, to beautify your group's performance and as an man or woman, enhance your very own. Apache Hadoop is written in Java, has large-scale implementation over statistics technological know-how [6]. This open-supply software program is widely general for its parallel statistics processing. It can take care of storage and

processing of Big Data required for records analysis. Any large file gets disbursed or cut up into smaller chunks after which surpassed over to one-of-a-kind nodes. In other words, Hadoop works by means of dispensing large statistics units

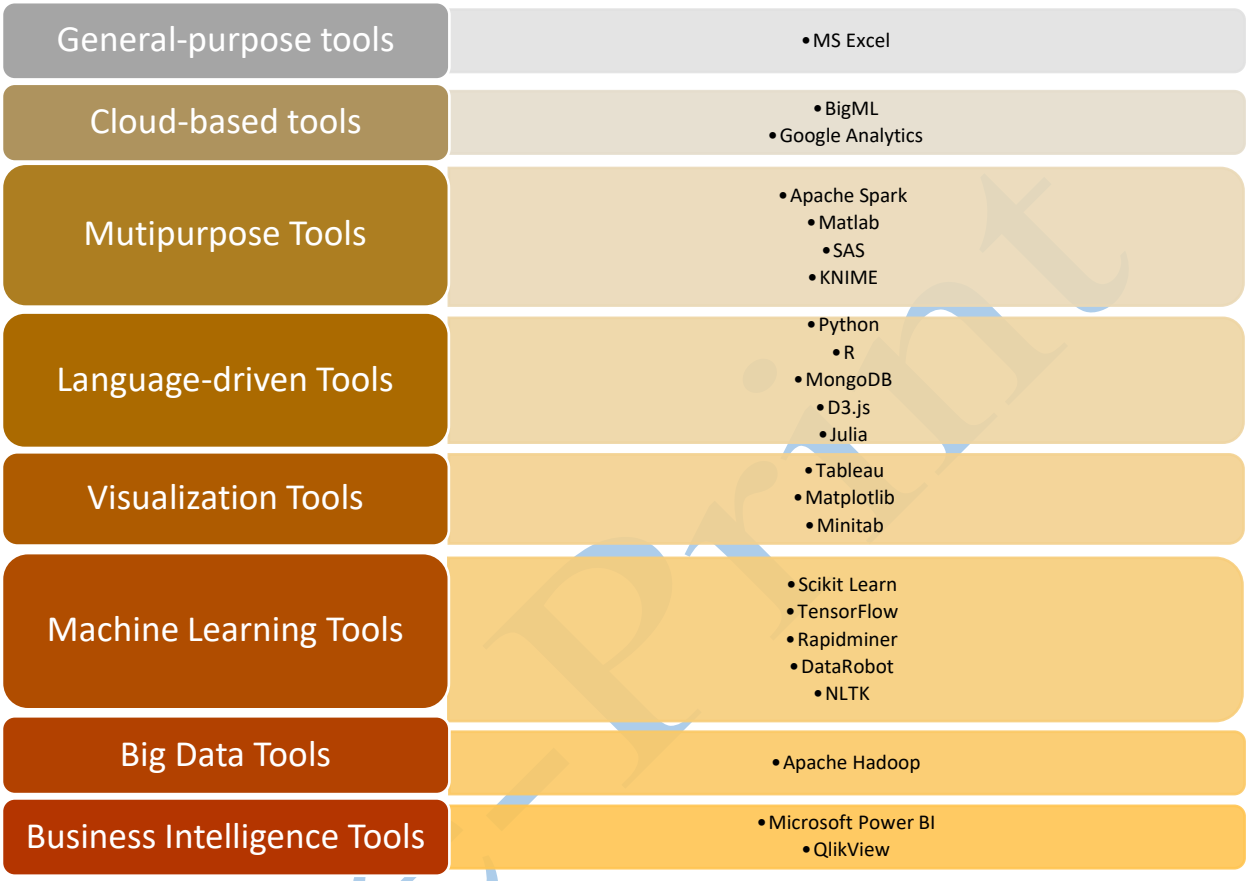

Fig. 4. – Sets of Tools

V. VARIETY OF DATA

Qualitative Data type of data can't be measured or counted in the form of numbers. These types of data are sorted by category. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data. Qualitative data talks about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly. Nominal Data is used to label variables without any order or comparison. For example, the name of language spoken by people, favourite subject of a student, gender, color, etc [7]. Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them. Ordinal data is qualitative data for which their values have some kind of relative position. The ordinal data only shows the sequences and cannot use for statistical analysis. For example, satisfaction level of a customer, educational status of a person (primary, middle, graduate, postgraduate etc.). Quantitative Data can be expressed in

numerical values, making it countable [8]. It answers the questions like "how much," "how many," and "how often." For example, the price of a phone, the computer's RAM, the height or weight of a person, etc., falls under quantitative data. Quantitative data can be used for statistical manipulation. These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, box plots, pie charts, line graphs, etc. Discrete Data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values. The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table. Continuous data are in the form of fractional numbers. It can be the version of an Android phone, the height of a person, the length of an object etc. [9].

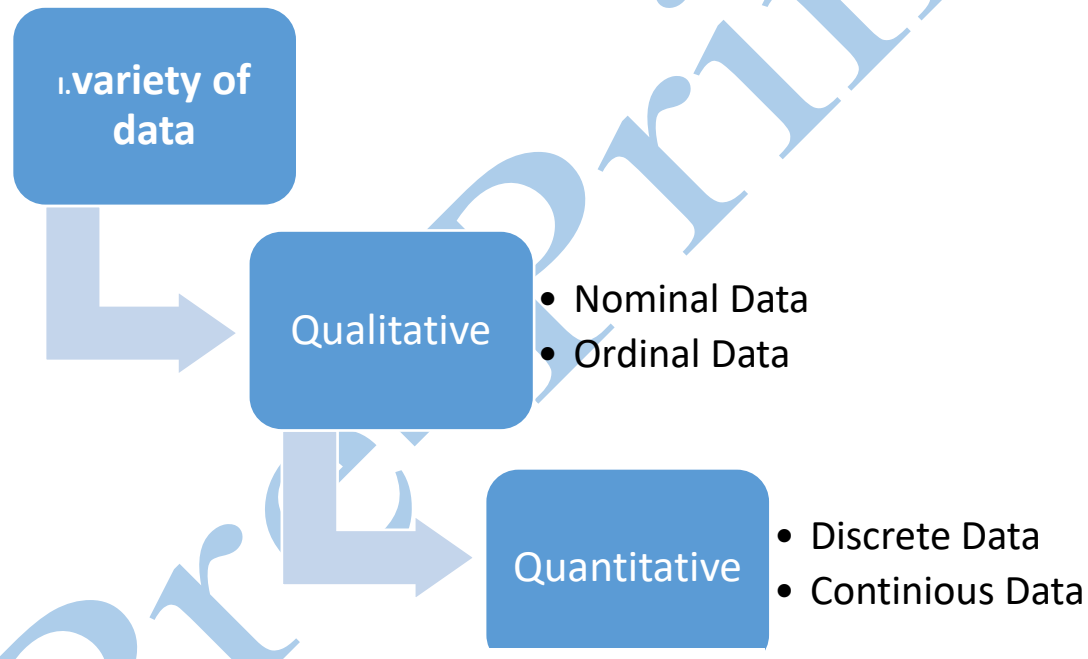

Fig. 5. – Variety of Data

VI. COLLECTION OF DATA

The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is known as Data Collection. Knowledge is power, information is knowledge, and data is information in digitized form. Hence, data is power. But before you can leverage that data into a successful strategy for your organization or business, you need to gather it and that is data collection. Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity. Primary data collection involves the collection of original data directly from the source or through direct interaction with the respondents. This method allows researchers to obtain first-hand information specifically tailored to their

research objectives. There are various techniques for primary data collection, including researchers design structured questionnaires or surveys to collect data from individuals or groups. These can be conducted through face-to-face interviews, telephone calls, mail, or online platforms. Interviews involve direct interaction between the researcher and the respondent [10]. They can be conducted in person, over the phone, or through video conferencing. Interviews can be structured, semi-structured, or unstructured. Researchers observe and record behaviors, actions, or events in their natural setting. This method is useful for gathering data on human behavior, interactions, or phenomena without direct intervention. Experimental studies involve the manipulation of variables to observe their impact on the outcome. Researchers control the conditions and collect data to draw conclusions about cause-and-effect relationships. Focus groups bring together a small group of individuals who discuss specific topics in a moderated setting [11]. This method helps in understanding opinions, perceptions, and experiences shared by the participants. Secondary data collection involves using existing data collected by someone else for a purpose different from the original intent. Researchers analyze and interpret this data to extract relevant information Researchers refer to books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data. Numerous online databases provide access to a wide range of secondary data, such as research articles, statistical information, economic data, and social surveys. Government agencies, research institutions, and organizations often maintain databases or records that can be used for research purposes. Data shared by individuals, organizations, or communities on public platforms, websites, or social media can be accessed and utilized for research. Previous research studies and their findings can serve as valuable secondary data sources. Researchers can review and analyze the data to gain insights or build upon existing knowledge. Accurate data collecting is crucial to preserving the integrity of research, regardless of the subject of study or preferred method for defining data (quantitative, qualitative). Errors are less likely to occur when the right data gathering tools are used [12].

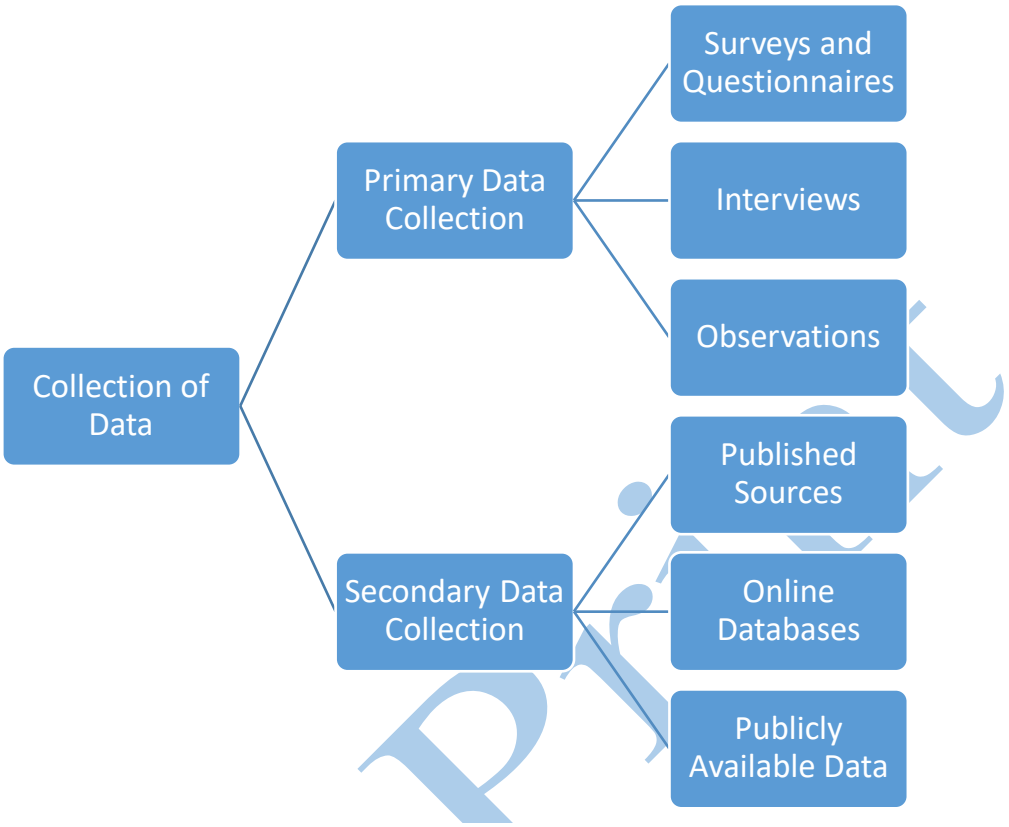

Fig. 6.- Collection of Data

VII. Dispute with Data Collection

The essential risk to the extensive and a hit utility of machine getting to know is negative information fine. Data nice must be your pinnacle precedence in case you want to make technology like gadget learning. When working with numerous data assets, it's achievable that the identical statistics may have discrepancies among sources. The differences may be in codecs, gadgets, or from time to time spellings [13]. Data is the riding pressure behind the selections and operations of records-pushed organizations. However, there may be quick durations while their records is unreliable or not prepared. A data engineer spends approximately eighty% of their time updating, retaining, and making sure the integrity of the information pipeline. Schema adjustments and migration problems are simply examples of the reasons of facts downtime. Data pipelines can be difficult because of their size and complexity [14]. Data downtime should be constantly monitored, and it must be decreased through automation. Even with thorough oversight, some errors can still occur in large databases. For statistics streaming at a quick velocity, the difficulty turns into more overwhelming. Spelling errors can move left out, formatting problems can arise, and column heads is probably deceptive. Streaming data, nearby databases, and cloud statistics lakes are only a few of the assets of information that current companies need to contend with. These assets are in all likelihood to duplicate and overlap each other pretty a piece. The likelihood

of biased analytical effects will increase while reproduction data are present. It also can result in ML models with biased education facts. While we emphasize statistics-driven analytics, a records pleasant trouble with excessive facts exists [15].

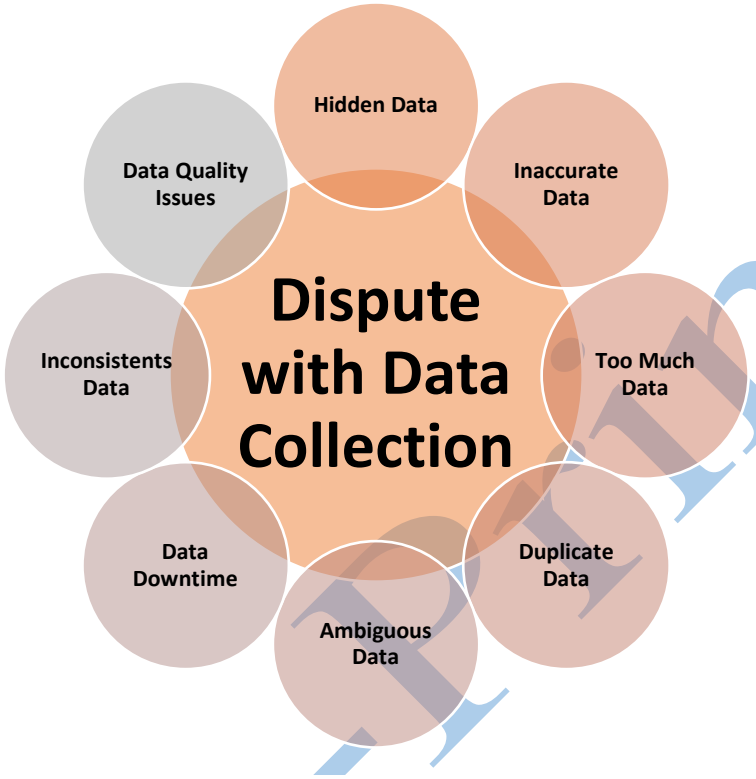

Fig. 7. – Dispute with Data Collection

VIII. ORIGIN OF DATA

The sources of data are different for primary data and secondary data. The data, which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. Data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviewsor formal interviews through telephone, face to face, email, etc. The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text,

audio, video, or any raw formats. For example, observing a group of customers and their behavior towards the products [16].

I. EXPLORATION OF DATA

Data exploration is the first step inside the statistics evaluation manner. It is also called Exploratory Data Analysis. Data analysts leverage records visualizations to discover styles and relationships within a dataset including numerical values, accuracy, quantity, length, and lots more. These characteristics allow us to get a deeper knowledge of the facts. Data analysts might use statistics visualization strategies to get a extensive view of the records and get an initial understanding of what the records says. Then manual manipulation would possibly take place, which include the usage of a drill-down approach to find anomalies or patterns within the records. Structure of a dataset, Relationships between special variables, presence of outliers or anomalies, distribution of information values techniques can be used to pick out information exploration [17]. Data streams are large volumes of continuous information that are available in special forms. For instance, a records circulate may be transactional, sensor, photo, or web visitors data, and plenty of extra. Having a big-photo of all the distinctive information streams can assist your business enterprise become aware of inefficiencies to your business, lessen threat, and streamline operations. Data exploration can assist empower teams across your organisation via supplying them with the facts they need [18]. Data democratization within an organisation ultimately enables to drive information-knowledgeable selections by using allowing anybody irrespective of their technical stage to get admission to and work with statistics without difficulty.

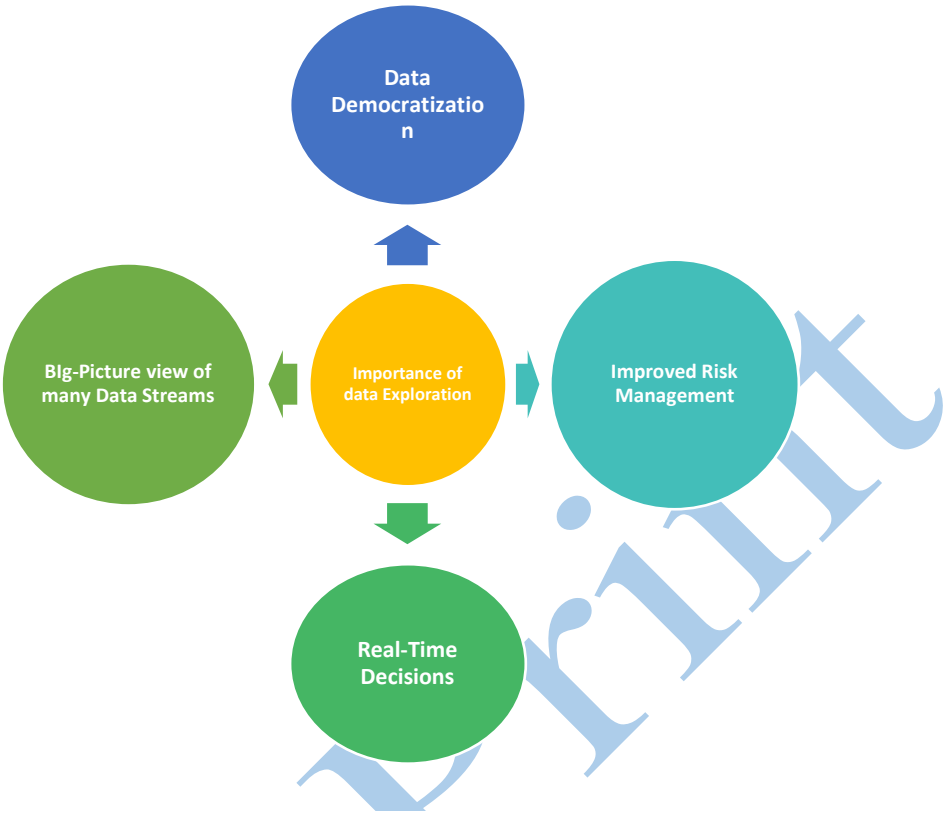

Fig. 8. – Importance of Data Exploration

II. STORAGE OF DATA

To store data, regardless of form, users need storage devices. Direct Attached Storage is as the name implies is often in the immediate area and directly connected to the computing machine accessing it [19]. Often, it's the only machine connected toit. DAS can provide decent local backup services, too, but sharing is limited. DAS devicesinclude floppy disks, optical discs—compact discs (CDs) and digital video discs (DVDs)—hard disk drives (HDD), flash drives and solid-state drives (SSD). Network-based Storage allows making it better for data sharing it better suited for backups and are allows more than one computer to access it through sharing and collaboration. Its off-site storage capability and data protection [20]. Two common network-based through a network, capability also makes based storage setups. Network Attached Storage is often a single device made up of redundant storage containers or a redundant array of independent disks (RAID). NAS storage has vast sharing capabilities due to the network connection. The device uses clustering and utilizes a file-storage system. Storage Area Network storage can be a network of multiple devices of various types, including SSD and flash storage, hybrid storage, hybrid cloud storage, backup software and appliances, and cloud storage. Flash storage is a solid-state technology that uses flash memorychips for writing and storing data. SSDs and flash offer higher throughput than HDDs, but all-flash arrays can be more expensive. Cloud storage delivers

a cost-effective, scalable alternative to storing files to on-premise hard drives or storage networks. Cloud service providers allow you to save data and files in an off-site location that you access through the public internetor a dedicated private network connection. The provider hosts, secures, manages, and maintains the servers and associated infrastructure and ensures you have access to the data whenever you need it. Hybrid cloud storage combines private and public cloud elements. With hybrid cloud storage, organizations can choose which cloud to storedata [21]. For instance, highly regulated data subject to strict archiving and replication requirements is usually more suited to a private cloud environment, whereas less sensitive data can be stored in the public cloud. Backup storage and appliances protect data loss from disaster, failure or fraud. They make periodic data and application copies to a separate, secondary device and then use those copies for disaster recovery. Backup appliances range from HDDs and SSDs to tape drives to servers, but backup storage can also be offered as a service, also known as backup-as-a-service (BaaS). File storage also called file-level or file-based storage, is a hierarchical storage methodology used to organize and store data. In other words, data is stored in files,the files are organized in folders and the folders are organized under a hierarchy of directories and subdirectories. Block storage sometimes referred to as block-level storage, is a technology used tostore data into blocks. The blocks are then stored as separate pieces, each with a unique identifier. Developers favour block storage for computing situations that require fast, efficient and reliable data transfer. Object storage often referred to as object-based storage, is data storage architecture for handling large amounts of unstructured data. This data doesn't conform to, or can'tbe organized easily into, a traditional relational database with rows and columns [22]. Examples include email, videos, photos, web pages, audio files, sensor data, and other types of media and web content (textual or non-textual).

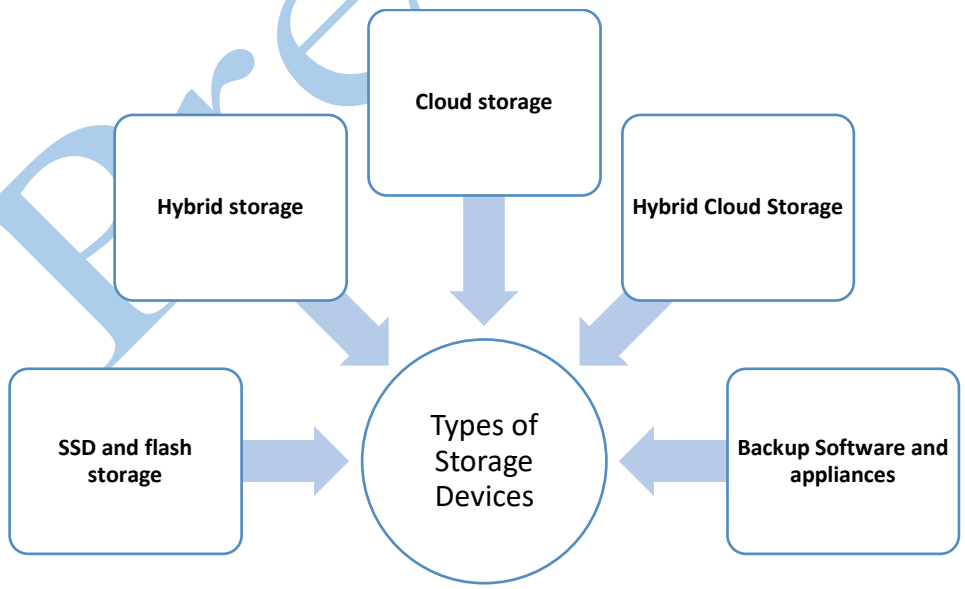

Fig. 9 - Types of Storage Devices

III. ORGANIZATION OF DATA

It is the process of collecting, storing, organizing and maintaining data to ensure that it is accurate, accessible to those who need it and reliable throughout your data science project lifecycle. Just like any management process, it requires procedures that are backed and supported by policies and technologies. Data Collection and Acquisition, Data Cleaning and Preprocessing, Data Storage, Data Security and Privacy, Data Governance and Documentation, Collaboration and Sharing are the key components of data management. Data management plays several roles in an organization's data environment, making essential functions easier and less time-intensive. Data preparation is used to clean and transform raw data into the right shape and format foranalysis, including making corrections and combining data sets. Data pipelines enable the automated transfer of data from one system to another [23]. Extract, Transform, Load are built to take the data from one system, transform it, and load it into the organization's data warehouse. Data catalogs help manage metadata to create a complete picture of the data, providing asummary of its changes, locations, and quality while also making the data easy to find Data warehouses are places to consolidate various data sources, contend with the many datatypes businesses store, and provide a clear route for data analysis. Data governance defines standards, processes, and policies to maintain data security andintegrity. Data architecture provides a formal approach for creating and managing data flow. Data security protects data from unauthorized access and corruption. Data modeling documents the flow of data through an application or organization [24].

Fig. 10. – Data Management Tool

IV. DATA COLLECTION WITH API

APIs are very flexible and may be used on web-based totally structures, operating structures, database systems and pc hardware [25]. APIs are the important constructing blocks for facts science. They provide key statistics sources and enable information integration and visualization. From the "Pay with Paypal" or "Login with Facebook" buttons to video games like Pokemon Go and tour aggregators which include Expedia, TripAvisor, and Booking.Com that let you compare costs of flights and lodges, APIs are all around us. They help connect our world and convey precious statistics from one website or software to every other. Amazon Machine Learning API constructed at the AWS cloud platform with a user-pleasant interface, Amazon facilitates with prediction models, generates useful visualizations, and enables statistical evaluation. IBM Watson permits you to sift thru online seek content material and locate patterns in company data [26]. It is all approximately making use of applications. Google API has a gaggle of APIs like Google Maps is an essential device for any mapping software and for calculating the distance between places. Google Maps includes 17

exceptional APIs underneath Maps, Places, and Routes and has emerge as one of the most popular net application development APIs, serving over a million websites and apps and 1000000000 customers [27]. Census.Gov API presents essential demographic and economic statistics from the U.S. Authorities; this API enables you query that statistics and prepare interesting programs and statistics projects built on one of the maximum authentic facts series businesses. Spotify API: Get the metadata related to the most popular songs (or maybe the maximum obscure).

V. EVOLUTION OF BIG DATA

In the previous era a computer have a small amount of memory and as the technology upgraded this demand is increasing in nature. Now a days the demand is so huge that it is not easy to manage that data in an efficient manner with the traditional approaches. In current era on daily basis enormous amount of data is generated in terms of social media, purchasing sites, weather forecasting, airline, hospital data set etc. To process this data without thinking of Big Data is beyond the imagination so manage this data and process this data Big Data comes into picture. The term Big Data is misleading as an impression that after certain size the data is big and upto that certain size that data is small [28]. The Big Data could start from any point there is no fix definition however it is mostly defined this way the Big Data is a data that become difficult to process because of its size using traditional system. If we want to send a 100 MB file via Gmail it is not possible with the traditional system because in Gmail maximum size of an attachment should be less than 25 MB. If we want to view a 100 GB image on our normal computer, it is not possible to view due to capability of system. If we want to edit due to limitation of software. Big Data is related to the capabilities of the system and at higher level the term is related to organization [29].

Table 1 Different symbol and memory size

Name	Data	Symbol
Kilobyte	$10^{\wedge 3}$	KB
Megabyte	$10^{\wedge 6}$	MB
Gigabyte	$10^{\wedge 9}$	GB
Terabyte	$10^{\wedge 12}$	TB
Petabyte	$10^{\wedge 15}$	PB
Exabyte	$10^{\wedge 18}$	EB

Zettabyte	$10^{\wedge 21}$	ZB
Yottabyte	$10^{\wedge 24}$	YB

We can see that how technology is evolved, from telephone to mobile phone with IOS that are making human being life smarter; apart form that we are using bulky desktop where we uses floppy to process the data & now we can store our data to cloud and in the similar way now the car is also converted into smart car that sense the data in the form of environment, weather condition, traffic volume etc. and these advancement in technology numerous amount of data on daily basis. IOT connects your physical device with internet and makes it smarter [30]. For example, Smart Air Conditioner which reads room temperature, outside temperature & body temperature and accordingly decide the temperature of room. In order to do if AC collect data from internet and process the same to function. Social media is actually one of the important factor in the evolution of data. Almost every single user uses different social media platform and they generate huge amount of data and most challenging task is that it that it generated unstructured data that is difficult to manage. So here it is not only generating data but also generates different forms of data. So these are few major examples for evolution of data, there are so many other reason for evolution of data. Big data is the term for collection of data sets so large and complex that it become difficult to process using on-hand database system tools or traditional data processing application. 'Big Data' is a term used to describe collection of data that is huge in size and yet growing exponentially with time. "Big Data" is the data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analysed with traditional computing techniques. A big data management architecture is designed to grip the absorption, processing and investigation of data that is too large or complex for traditional database systems [31].

Big data solutions typically involve one or more of the following types of workloads:

- Batch processing of big data sources at rest.
- Real-time processing of big data in motion.
- Interactive exploration of big data.
- Predictive analytics and machine learning.

Big Data architecture consists of components like all big data solutions start with one or more data sources like Application data stores, such as relational databases; Static files produced by applications, such as web server log files; Real-time data sources, such as IOT devices. Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a data lake [32]. Options for implementing this storage include Azure Data Lake Store or blob containers

in Azure Storage. Batch Processing data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to new files. Real-time Message Ingestion if the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing. However, many solutions need a message ingestion store to act a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics. Stream Processing after capturing real-time message, the solution must process them by filtering, aggregating and otherwise preparing the data for analysis. The processed stream data is then written to an output sink. Azure Stream Analytics provides a managed stream processing service based on perpetually running SQL queries that operate on unbounded streams [33]. Analytical Data store many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools. Azure SQL Data Warehouse provides a managed service for large-scale, cloud-based data warehousing. Analysis and reporting the goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower users to analyse the data, the architecture may include a data modelling layer, such as a multidimentional OLAP cube or tabular data model in Azure Analysis Services. Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts. Most big data solutions consist of repeated data processing operations, encapsulated in workflows that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report or dashboard [34]. There are several sources of Big Data which generate a lot of new data. The unstructured data, now mostly generated through the internet and social media. Text messages, tweets(blog) posts are increasingly becoming an important source of data relevant for any organization. Capturing unstructured data, traditionally documents and email messages, has been the territory of Enterprise Content Management. Document capture software has been along for some decennia. But can those systems cope with the increasing amounts and number of sources in a Big Data environment. When we focus on the first of the Big Data challenges, the capture of data, the level of challenge is mixed. It's no real problem to suck in large amounts of data. Well, that been said, this can still pose some major technical issues. What data do we keep and what data do we discard as redundant, obsolete, trivial and irrelevant. It's all about keeping your collection of Big Data fit for purpose, now and in the future [35]. This also implies that you know what you store. So one of the tasks of capturing is filtering the data to only keep the relevant information. A big data research platform needs to process massive quantities of data – filtering, transforming and sorting it before loading it into a data warehouse. Oracle offers a choice of products for organizing data. In addition, Oracle enables end-to-end control of structured and unstructured content, allowing you to manage all your data from application-to archive efficiently, securely, and cost effectively with the Oracle content management and tiered storage solution designed specifically for research organizations. Big data integration is discovering information, profiling them, understanding

the data, the value of that data, tracking through metadata, improving the quality of data and then transforming it into the form that is required for big data. Every big data use case needs integration. There are several challenges one can face during this integration such as analysis, data curation, capture, sharing, search, visualization, information privacy and storage. The infrastructure required for analysing big data must be able to support deeper analytics such as statistical analysis and data mining on a wider variety of data types stored in diverse systems; scale to extreme data volumes; deliver faster response times; and automate decisions based on analytical models. Oracle offers a portfolio of tools for statistical and advanced analysis [36].

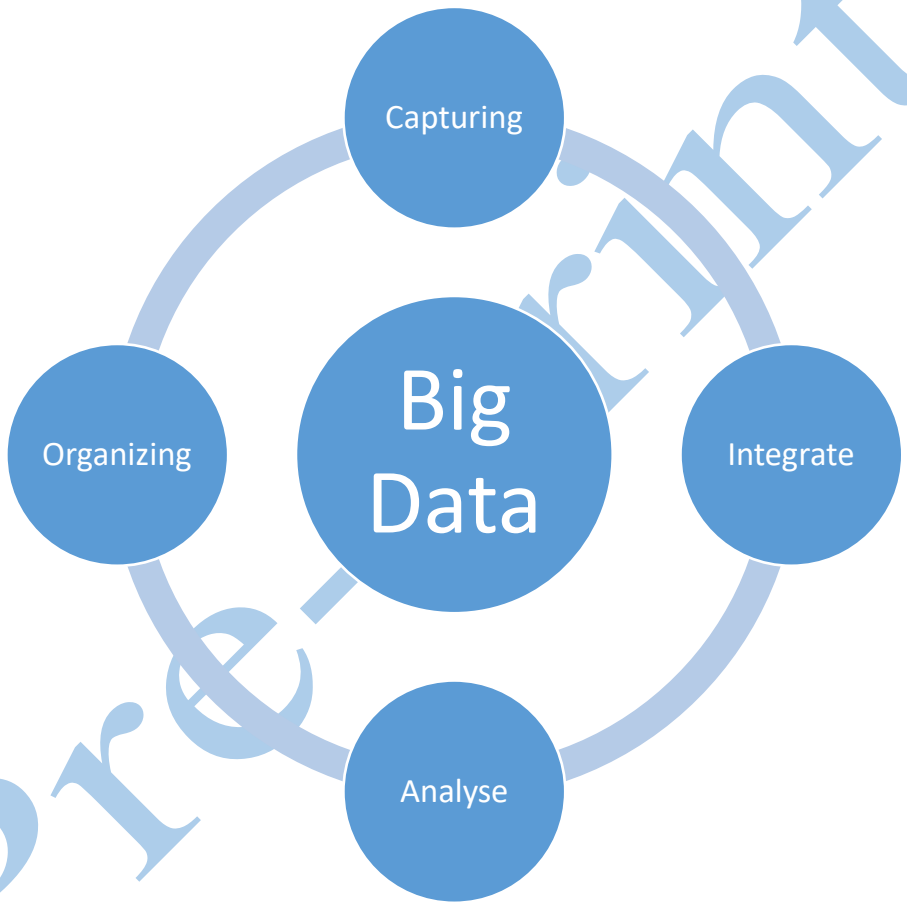

Fig. 11. – Capturing, Organizing, Integrate, Analyse of Big Data

VI. ARCHITECTURE FOUNDATION OF BIG DATA

Big fact's structure is the muse for big statistics analytics. Architects begin by way of know-how the goals and targets of the mission, and the insights of different processes. Thus, the architecture foundation desires the proper planning and effective tools. System structure basis desires the right making plans and effective equipment. System architects undergo a same sample to plot big fact's structure. Meeting with stakeholders to recognize business enterprise targets for its massive information and plan the framework with suitable hardware and software program, information assets and formats, analytics gear, information garage choices, and outcomes intake. The need of big information architecture depends on the pattern and size of facts. Single computing tasks rarely pinnacle extra than a hundred GB of records, does now

not require a huge statistics architecture. Unless you're analysing terabytes and petabytes of facts on a consistent to a scalable server in place of a massively scale-out structure [37]. A character probably do need huge records architecture to extract statistics from sizeable networking or internet logs. Process massive datasets over 100GB in length. Invest in a massive facts venture, such as third-birthday party products to optimize your environment. Store huge amounts of unstructured records and to summarize or remodel into a dependent format. Have a couple of big information assets to examine, which includes structured and unstructured. Want to proactively examine huge records for enterprise wishes. Performance testing for big information utility entails checking out of large volumes of statistics, and it requires a selected checking out approach. Big statistics structure also wishes to perform in live performance with the employer's supporting infrastructure [38].

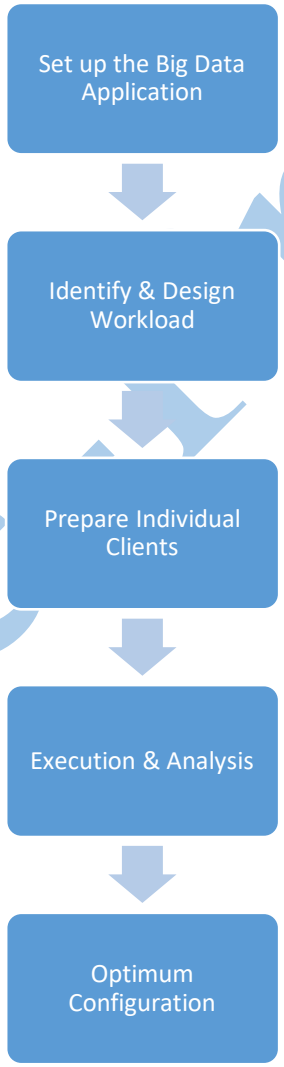

Fig. 12. - Distributed Computing Model

VII. TYPES OF BIG DATA

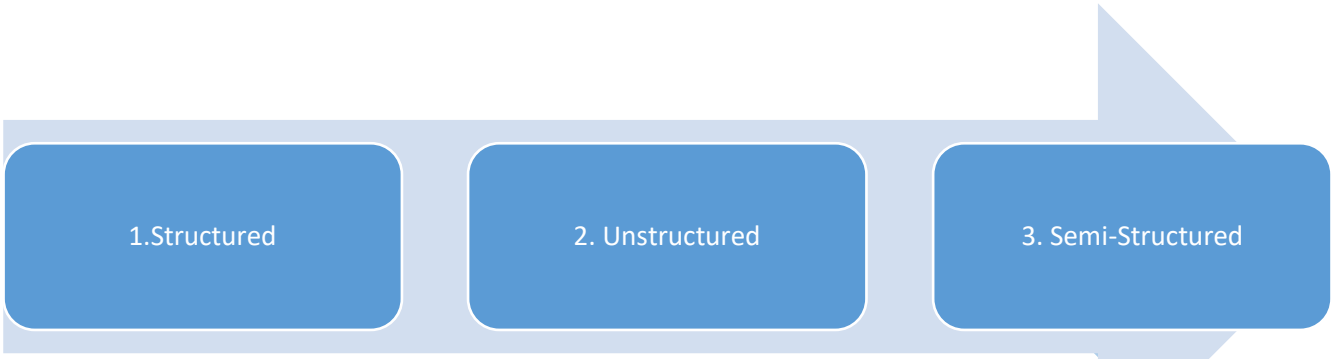

Fig. 13. – Big Data Types

Structured data, that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the student table in a database will be structured as the student details, their achievements, their semester grades, etc., in an organized manner [39].

Unstructured data refers to the data that lacks any specific form or structure. This makes it very difficult and time-consuming to process and analyse unstructured data.

Example: Output returned by 'Google Search'

Semi-Structured: Semi-Structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database) yet contains vital information or tags that segregate individual elements within the data.

Example: Personal data stored in a XML file [40].

VIII. CHARACTERISTICS OF BIG DATA

The facts that consequences into massive files. For instance, social media that is an increasing number of day by day. Processing special kind of information that may be of structured, semi-based or unstructured in nature. Data is being generated at an alarming fee. Finding correct meaning of the facts. Uncertainty and inconsistencies in the facts. The term dependent statistics generally refers to facts that has a defined duration and format for massive data. Examples of established information encompass numbers, dates, and corporations of words

and numbers called strings. Most experts agree that this kind of facts money owed for accounts for about 20 percent of the information this is out there. Although this could appear like business as typical, in truth, established statistics is taking over a brand-new function in the world of big statistics. The evolution of technology offers newer assets of established facts being produced often in actual time and in big volumes. The sources of information are divided into categories first one is Computer or Machine generated: Machine-generated facts generally refer to statistics this is created via a gadget without human intervention. Example: Sensor statistics, Web log records, Point-of-sale records, monetary information etc. Second one is Human Generated: This is information that people, in interplay with computer systems, supply. Example: Input facts, Click-movement facts, gaming associated statistics and many others. Big records are becoming an vital element within the way businesses are leveraging high-volume statistics at the proper pace to resolve unique statistics troubles. Relational Database Management Systems are essential for this excessive extent. Big facts do no longer stay in isolation. To be powerful, agencies often need in an effort to combine the results of big records evaluation with the data that exists in the commercial enterprise [41].

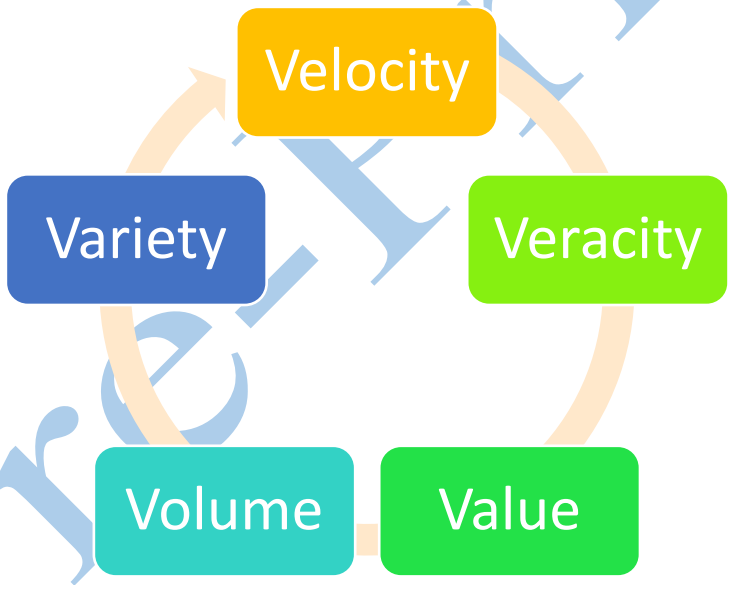

Fig. 14. - Big Data Characteristics

IX. DATA INTEGRATION TYPES

Data Integration is the process of transferring the data from source to destination format. Many data warehousing and data management approaches has been supported by integration tools for data migration and transportation by using Extract-Transform-Load (ETL) approach. These tools are widely fit for handing large volumes of data and not flexible to handle semi or unstructured data. To overcome these challenges in big data world, programmatically driven parallel techniques such as map-reduce models were introduced. Data Integration as a process is highly cumbersome and iterative especially to add new data sources. Traditionally

waterfall approach is used in EDW (Enterprise Data Warehouse) , where one cannot move to the next phase before completing the earlier one. This approach has its developed to sustain the usefulness of EDW. In big data environment, the situation is completely different. Therefore the traditional approaches of integration are inefficient in handling the current situation. So people are expected to do something regarding this issue. Extract, Transform and Load (ETL) comes from Data Warehousing and stands for Extract - Transform-Load. ETL performs the process of loading data from the source system to target system. It is converting data of same type or different types to centralized Data Warehouse which is having standard format for all data. **Extract:** The "Extract" task involves gathering data from external sources that needs to be brought to the required systems and databases. The goal of this task is to understand the format of data, assess the overall quality of the data and to extract the data from its source so that it can be manipulated in next task. **Transform:** In the "transform" step a variety of software tools and even custom programming are used to manipulate the data so that it integrates with data that is already present. **Load:** After the successfully transformation of the source data, it is required to physically load it into the target system or database. Before loading the data, it is required to make sure that there is a backup of the current system so that roll back or undo can be initiated in case of failure of the Load process. After loading the data, it is common to run audit reports so that there can be review of the results of the merged databases and systems to make sure the new data has not caused any errors [42].

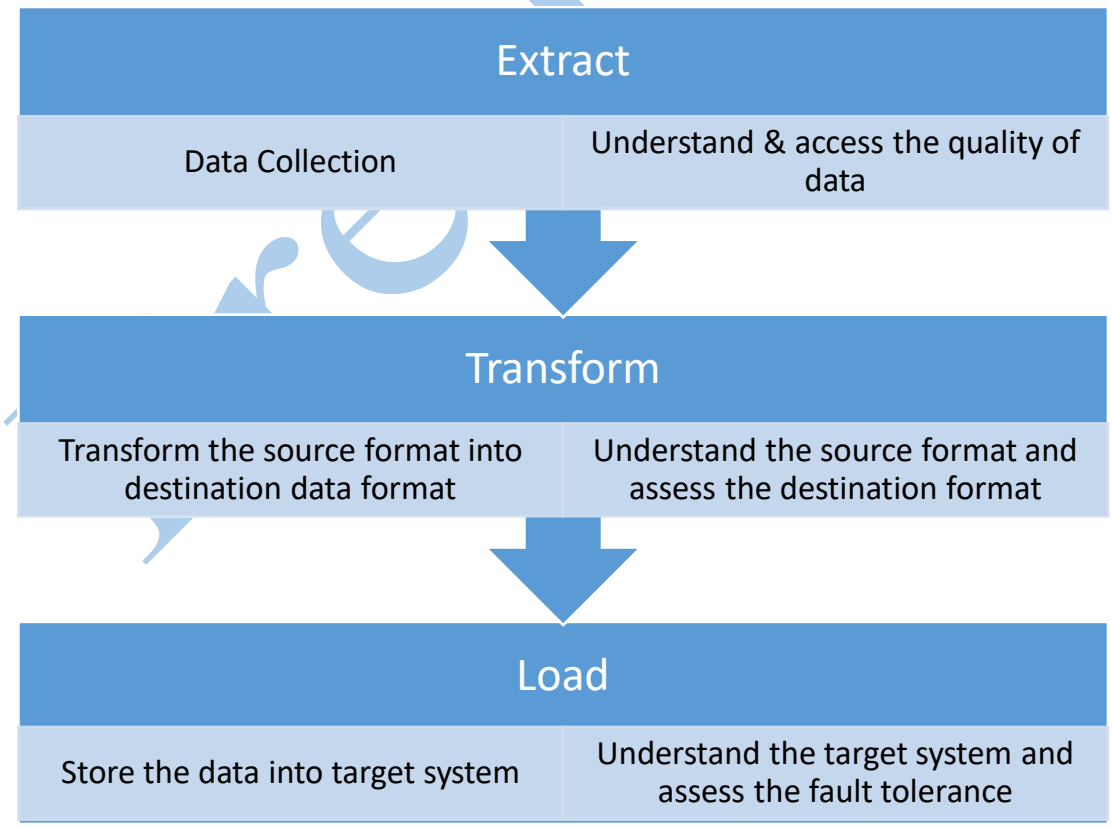

Fig. 15. – ETL Process

X. INTRODUCTION OF ANALYTICS

Data Analytics the science of examining raw data with the purpose of drawing conclusions about that information. Data Analytics involves applying an algorithmic or mechanical process of derive insights. For example, running applying an algorithmic or mechanical process to derive insights. For example, running through a number of data sets to look for meaningful correlations between each other. It is used in a number of industries to allow the organizations and companies to make better decisions as well as verify and disprove existing theories or models. The focus of Data Analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows. Data Analysis in healthcare the main challenge for hospitals with cost pressures tightens is to treat as many patients as they can efficiently, keeping in mind the improvement of the quality of care. Data analytics is able to optimize the buying experience through the mobile/weblog and the social media data analysis. Travel sights can gain insights into the customer's desires and preferences. Data Analytics helps in collecting data to optimize and spend within as well as across games. Game companies gain insight into the dislikes, the relationships, and the likes of the users. Most firms are using data analytics for energy management, including smart-grid management, energy optimization, energy distribution, and building automation in utility companies. Core architecture data model (CADM) in enterprise architecture is a logical data model of information used to describe and build architectures [43].

XI. DATA WAREHOUSE ARCHITECTURE

The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice. Two-layer architecture separates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations. Three-tier architecture is the most widely used architecture. It consists of the Top, Middle and Bottom Tier. The database of the Datawarehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed, and loaded into this layer using back-end tools. The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database. The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be Query tools, reporting tools, managed query tools, Analysis tools and Data mining tools. The data warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key components to make the entire environment functional, manageable and accessible. A column-oriented database stores each column continuously. i.e. on disk or in-memory each column on the left will be stored in sequential blocks. For analytical queries that perform

aggregate operations over a small number of columns retrieving data in this format is extremely fast. As PC storage is optimized for block access, by storing the data beside each other we exploit locality of reference. On hard disk drives this is particularly important which due to their performance characteristics provide optimal performance for sequential access. The goal of a columnar database is to efficiently write and read data to and from hard disk storage in order to speed up the time it takes to return a query. One of the main benefits of a columnar database is the data can be highly compressed. The compression permits columnar operations – like MIN, MAX, SUM, COUNT and AVG – to be performed very rapidly. Another benefit is that because a column-based DBMSs is self-indexing, it uses less disk space than a relational database management system (RDBMS) containing the same data. The best example of a Column – Oriented data store is HBase Database, which is basically designed from the ground up to provide scalability and partitioning to enable efficient data structure serialization, storage, and retrieval.

A computer performs tasks according to the instructions provided by the human. Parallel computing and distributed computing are two computation types. Parallel computing is used in high – performance computing such as supercomputer development. Distributed computing provides data scalability and consistency. Google and Facebook use distributed computing for data storing. The key difference between parallel and distributed computing is that parallel computing is to execute multiple tasks using multiple processors simultaneously while in distributed computing, multiple computers are interconnected via a network to communicate and collaborate in order to achieve a common goal. Each computer in the distributed system has their own users and helps to share resources. Shared-nothing architecture (SNA) is a patten used in distributing computing in which a system based on multiple self-sufficient nodes that have their own memory, HDD storage and independent input/output interfaces. Each node shares no resources with other nodes, and there is a synchronization mechanism that ensure that all information is available on at least two nodes. Shared-nothing architecture is very popular in web applications, because it provides almost infinite horizontal scaling that can be made with very inexpensive hardware. It is widely used by Google, Microsoft and many other companies that need to collect and process massive sets of data. One of the good examples of using SNA architecture is a MySQL cluster. It features a Network Data Base (NDB) storage engine that automatically distributes MySQL data across multiple storage nodes and provides great performance in write-heavy application. Shared-nothing is popular for web development because of its scalability. As Google has demonstrated, a pure SN system can scale simply by adding nodes in the form of inexpensive computers, since there is no single bottleneck to slow the system down. Google calls this sharding. A SN system typically partitions its data among many nodes on different databases (assigning different computers to deal with different users or queries),or many require every node to maintain its own copy of the application's data, using some kind of coordination protocol. This is often referred to as database sharding. There is some doubt about whether a web application with many independent web nodes but a single, shared database should be counted as SN. One of the approaches to achieve SN architecture

for stateful applications is the use of a data grid, also known as distributed caching. This still leaves the centralized database as a single point of failure

XII. CONCLUSION

Over the past decade, data science has changed dramatically due to rapid advances in artificial intelligence, big data and cloud computing. These developments have expanded the range of applications, enabling sectors such as healthcare, finance and urban planning to use data to make more informed decisions but challenges remain, especially in terms of data privacy, ethics problems and the need to address biases in algorithms. Managing and scaling unstructured data solutions is complex. Despite these issues, the sector continues to offer great opportunities, particularly in emerging areas such as personalized healthcare and smart cities, where data science can drive innovation and social improvement.

REFERENCE

[1] Kou, G., Peng, Y., Shi, Y., & Chen, Z. (2007). Privacy-preserving data mining of medical data using data separation-based techniques. *Data science journal*, *6*, S429-S434.

[2] Boulemtafes, A., Derhab, A., & Challal, Y. (2020). A review of privacy-preserving techniques for deep learning. *Neurocomputing*, *384*, 21-45.

[3] Carvalho, T., Moniz, N., Faria, P., & Antunes, L. (2022). Survey on privacy-preserving techniques for data publishing. *arXiv preprint arXiv:2201.08120*.

[4] Singh, A. P., & Parihar, M. D. (2013). A review of privacy preserving data publishing technique. *International Journal of Emerging Research in Management &Technology*, *2*(6), 32-38.

[5] Taric, G. J., & Poovammal, E. (2017). A survey on privacy preserving data mining techniques. *Indian Journal of Science and Technology*.

[6] Nayak, Gayatri, and Swagatika Devi. "A survey on privacy preserving data mining: approaches and techniques." *International Journal of Engineering Science and Technology* 3.3 (2011): 2127-2133.

[7] Rashid, Asmaa Hatem, and Norizan Binti Mohd Yasin. "Privacy preserving data publishing." *International Journal of Physical Sciences* 10.7 (2015): 239-247.

[8] Iezzi, Michela. "Practical privacy-preserving data science with homomorphic encryption: an overview." *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020.

[9] Rajesh, N., K. Sujatha, and A. Arul Lawrence. "Survey on privacy preserving data mining techniques using recent algorithms." *International Journal of Computer Applications* 133.7 (2016): 30-33.

[10] Vaghashia, Hina, and Amit Ganatra. "A survey: privacy preservation techniques in data mining." *International Journal of Computer Applications* 119.4 (2015).

[11] Bertino, Elisa, Dan Lin, and Wei Jiang. "A survey of quantification of privacy preserving data mining algorithms." *Privacy-preserving data mining: Models and Algorithms* (2008): 183-205.

[12] Vaghashia, Hina, and Amit Ganatra. "A survey: privacy preservation techniques in data mining." *International Journal of Computer Applications* 119.4 (2015).

[13] Kiran, P., and N. P. Kavya. "A survey on methods, attacks and metric for privacy preserving data publishing." *International Journal of Computer Applications* 53.18 (2012).

[14] https://www.researchgate.net/publication/220670223_Survey_on_Privacy_Preserving_Data_Mining

[15] Aggarwal, Charu C., and Philip S. Yu. "A condensation approach to privacy preserving data mining." *International Conference on Extending Database Technology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[16] Churi, Prathamesh P., and Ambika V. Pawar. "A systematic review on privacy preserving data publishing techniques." *Journal of Engineering Science & Technology Review* 12.6 (2019).

[17] Aggarwal, Charu C., and Philip S. Yu. *A general survey of privacy-preserving data mining models and algorithms*. Springer US, 2008.

[18] Kiran, Ajmeera, and D. Vasumathi. "A comprehensive survey on privacy preservation algorithms in data mining." *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2017.

[19] Sreedhar, K. C., M. N. Faruk, and B. Venkateswarlu. "A genetic TDS and BUG with pseudo-identifier for privacy preservation over incremental data sets." *Journal of intelligent & fuzzy systems* 32.4 (2017): 2863-2873.

[20] Murugaboopathi, G., and V. Gowthami. "Slicing based efficient privacy preservation technique with multiple sensitive attributes for safe data distribution." *Journal of Intelligent & Fuzzy Systems* 40.2 (2021): 2661-2668.

[21] Agrawal, Dakshi, and Charu C. Aggarwal. "On the design and quantification of privacy preserving data mining algorithms." *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2001.

[22] Cabrero-Holgueras, José, and Sergio Pastrana. "Sok: Privacy-preserving computation techniques for deep learning." *Proceedings on Privacy Enhancing Technologies* (2021).

[23] Keshk, Marwa, et al. "Privacy-preserving big data analytics for cyber-physical systems." *Wireless Networks* 28.3 (2022): 1241-1249.

[24] Shah, Alpa, and Ravi Gulati. "Privacy preserving data mining: techniques, classification and implications-a survey." *Int. J. Comput. Appl* 137.12 (2016): 40-46.

[25] Ratra, Ritu, and Preeti Gulia. "Privacy preserving data mining: techniques and algorithms." *International Journal of Engineering Trends and Technology* 68.11 (2020): 56-62.

[26] Torkzadehmahani, Reihaneh, et al. "Privacy-preserving artificial intelligence techniques in biomedicine." *Methods of information in medicine* 61.S 01 (2022): e12-e27.

[27] Ram Mohan Rao, P., S. Murali Krishna, and A. P. Siva Kumar. "Privacy preservation techniques in big data analytics: a survey." *Journal of Big Data* 5.1 (2018): 33.

[28] Dargan, Shaveta, et al. "A survey of deep learning and its applications: a new paradigm to machine learning." *Archives of Computational Methods in Engineering* 27 (2020): 1071-1092.

[29] Devi, I., G. R. Karpagam, and B. Vinoth Kumar. "A survey of machine learning techniques." *International Journal of Computational Systems Engineering* 3.4 (2017): 203-212.

[30] Al-Sahaf, Harith, et al. "A survey on evolutionary machine learning." *Journal of the Royal Society of New Zealand* 49.2 (2019): 205-228.

[31] Guliyev, Hasraddin, Natiq Huseynov, and Nasimi Nuriyev. "The relationship between artificial intelligence, big data, and unemployment in G7 countries: New insights from dynamic panel data model." *World Development Sustainability* 3 (2023): 100107.

[32] Teo, Zhen Ling, et al. "Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture." *Cell Reports Medicine* (2024).

[33] Krichen, Moez. "Convolutional neural networks: A survey." *Computers* 12.8 (2023): 151.

[34] Bhawantha, Praveen, and Maneesha Attanayake. "Data Science: What is it and the Importance of it." *Authorea Preprints* (2024).

[35] Ali, Zulqarnian. "Data Science: Its Role and Importance."

[36] Shaikh, Mustaq, and Farjana Birajdar. "Artificial intelligence in groundwater management: Innovations, challenges, and future prospects." *International Journal of Science and Research Archive* 11.1 (2024): 502-512.

[37] Gajawada, Satish. *Data Science Plus Plus (DS++): The Definition*. No. 12477. EasyChair, 2024.

[38] LASTNAME, A., U. SELBERG, and BO MÖBIUS. "DATA SCIENCE TECHNIQUES FOR COMPUTER SCIENCE CHALLENGES."

[39] Zamani, Efpraxia D., et al. "Artificial intelligence and big data analytics for supply chain resilience: a systematic literature review." *Annals of Operations Research* 327.2 (2023): 605-632.

[40] Wen, Jie, et al. "A survey on federated learning: challenges and applications." *International Journal of Machine Learning and Cybernetics* 14.2 (2023): 513-535.

[41] Qiu, Junfei, et al. "A survey of machine learning for big data processing." *EURASIP Journal on Advances in Signal Processing* 2016 (2016): 1-16.

[42] Lee, Jea Woog, et al. "Soccer's AI transformation: deep learning's analysis of soccer's pandemic research evolution." *Frontiers in Psychology* 14 (2023): 1244404.

[43] Pramanik, M. Ileas, et al. "Privacy preserving big data analytics: A critical analysis of state-of-the-art." *Wiley*

Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11.1 (2021): e1387.