BIG DATA & ANALYTICS

UNIT-1

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

What is Big Data?

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

What is an Example of Big Data?

Following are some of the Big Data examples-

The **New York Stock Exchange** is an example of Big Data that generates about **one terabyte** of new trade data per day.

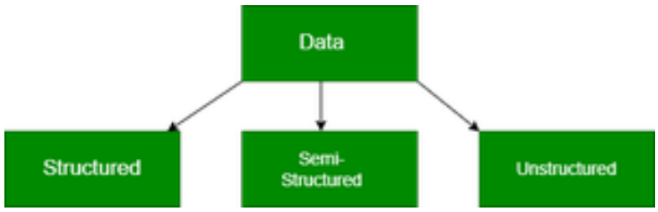
Social Media

The statistic shows that **500+terabytes** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc

Flight sites generate **10+terabytes** of data in **30 minutes** of flight time. With many thousand flights per day, generation of data reaches up to many **Petabytes.**

Types of Big Data

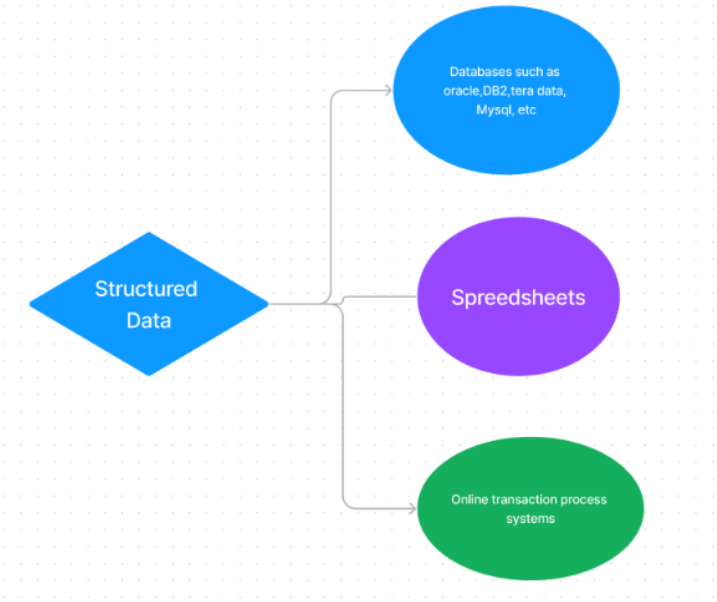
Faculty: Ankita Varshney
+91-8392877677; ankita@vision.org.in



Structured:

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

However, nowadays, size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.



Examples Of Structured Data

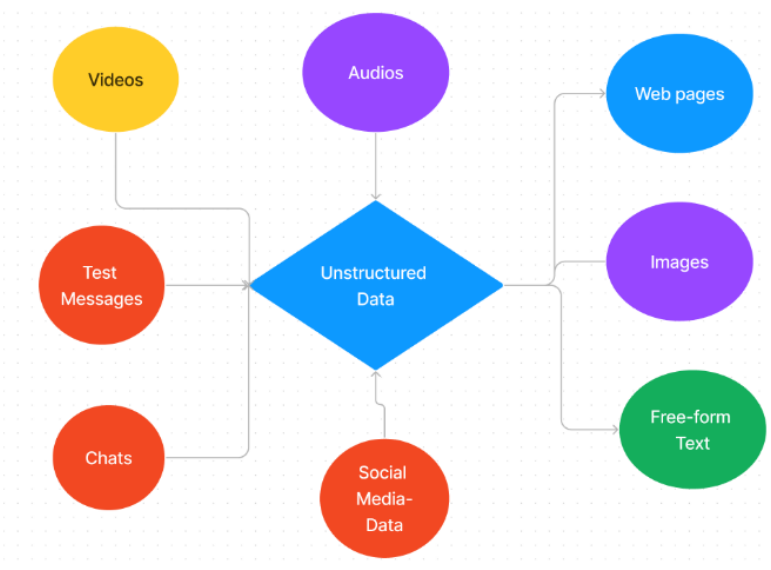
An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured:

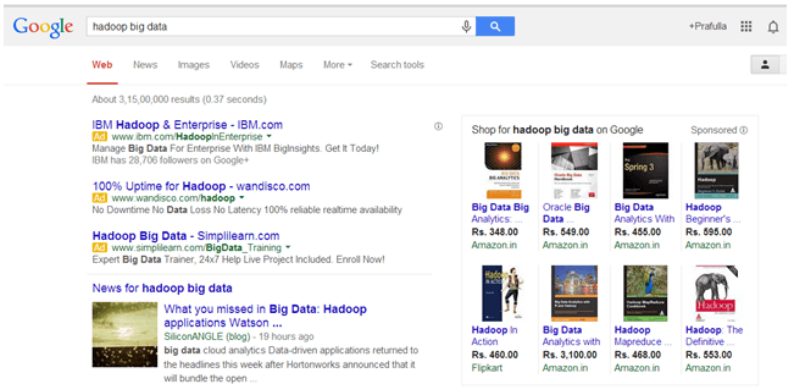
Any data with unknown form or the structure is classified as unstructured data

A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.



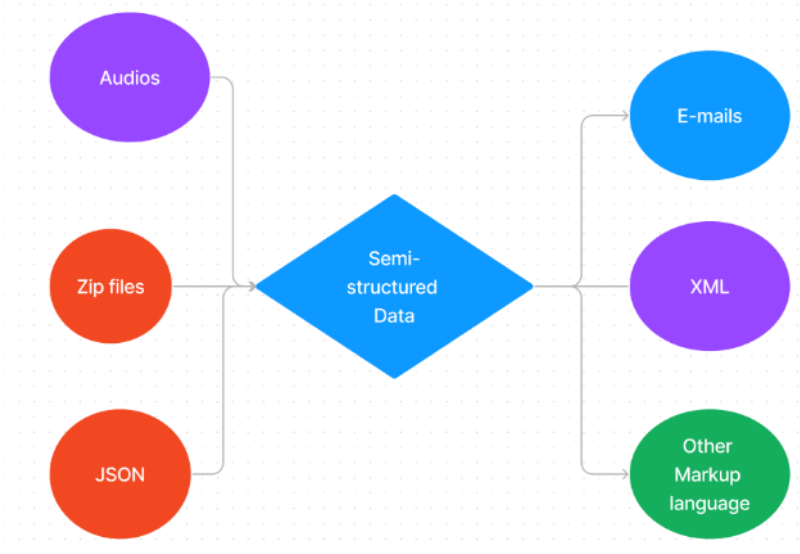
Examples Of Un-structured Data

The output returned by 'Google Search'

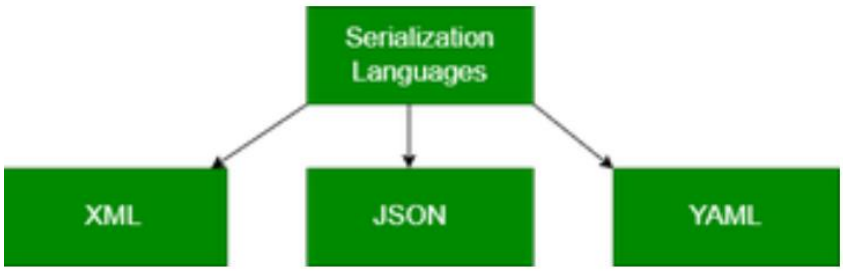


Semi-structured:

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.



Examples Of Semi-structured Data-



1. XML– XML stands for *eXtensible Markup Language*. It is a text-based markup language designed to store and transport data.

Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

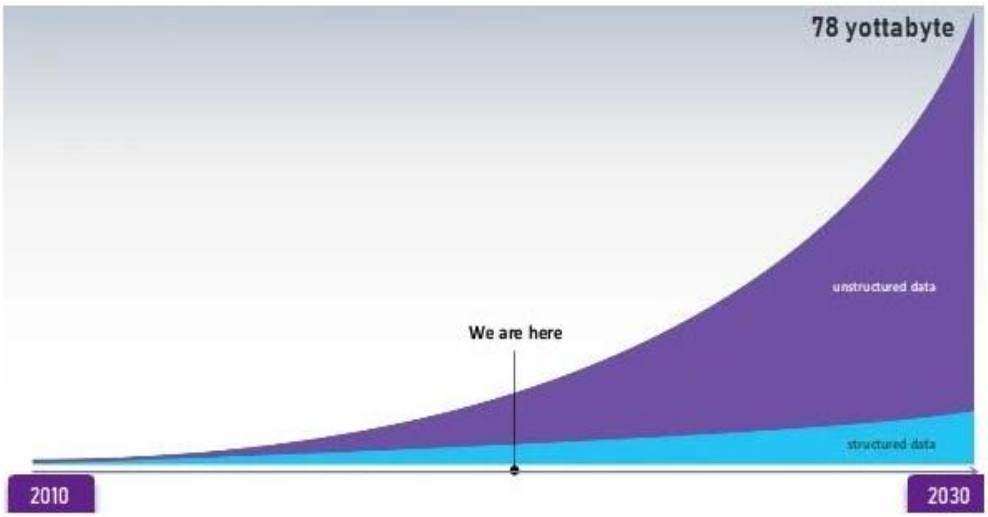
2. JSON– JSON (JavaScript Object Notation) is a lightweight open-standard file format for data interchange. JSON is easy to use and uses human/machine-readable text to store and transmit data objects.

```
{
    "firstName": "Jane",
    "lastName": "Doe",
    "codingPlatforms": [
        { "type": "Fav", "value": "Geeksforgeeks" },
        { "type": "2ndFav", "value": "Code4Eva!" },
        { "type": "3rdFav", "value": "CodeisLife" }
    ]
}
```

3. YAML(yet another markup language)– YAML is a user-friendly data serialization language. Figuratively, it stands for *YAML Ain't Markup Language.*

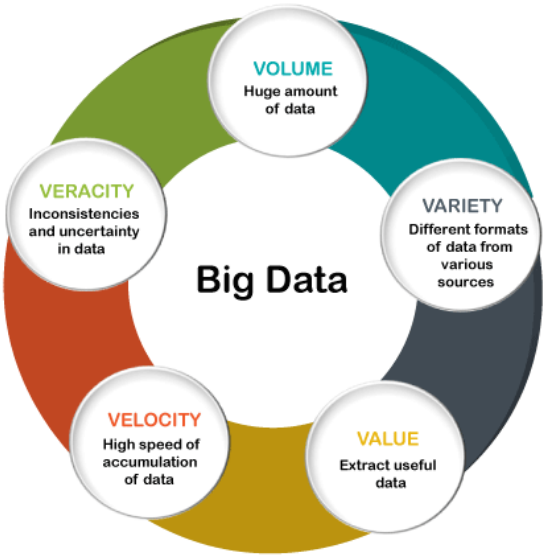
```
firstName: Jane
lastName: Doe
CodingPlatforms:
    - type: Fav
      value: Geeksforgeeks
    - type: 2ndFav
      value: Code4Eva!
    - type: 3rdFav
      value: CodeisLife
```

Data Growth over the years



Data Storage	Units
Bit	1 or 0
Byte	8 bits
Kilobyte	1,000 bytes
Megabyte	1,000 kilobytes
Gigabyte	1,000 megabytes
Terabyte	1,000 gigabytes
Petabyte	1,000 terabytes
Exabyte	1,000 petabytes
Zettabyte	1,000 exabytes
Yottabyte	1,000 zettabytes

Characteristics Of Big Data:



Big data can be described by the following characteristics:

- Volume
- Variety
- Value
- Velocity
- Variability

(i) Volume – The size and amounts of big data that companies manage and analyze. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data solutions.

(ii) Variety – The next aspect of Big Data is its **variety**.

The diversity and range of different data types, including unstructured data, semi-structured data and raw data.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

(iv) Value: Value refers to the benefits that big data can provide, and it relates directly to what organizations can do with that collected data.

Organizations can use big data tools to gather and analyze the data, but how they derive value from that data should be unique to them. Tools like Apache Hadoop can help organizations store, clean and rapidly process this massive amount of data.

(iii) Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

So,Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

(iv) Variability – Variability refers to data whose meaning is constantly changing. This refers to the inconsistency which can be shown by the data at times, the changing nature of the data companies seek to capture, manage and analyze.

Example of Variability in Big Data can be seen when investigating the amount of time spent on phones daily by diverse groups of people. The data collected from different samples (high school students, college students, and adult full-time employees) can vary, resulting in variability.

Another example could be a soda shop offering different blends of soda but having different taste every day, which is variability.

Advantages Of Big Data Processing

Ability to process Big Data in DBMS brings in multiple benefits, such as-

- Businesses can utilize outside intelligence while taking decisions

Access to social data from search engines and sites like Facebook, Twitter are enabling organizations to fine tune their business strategies.

- Improved customer service

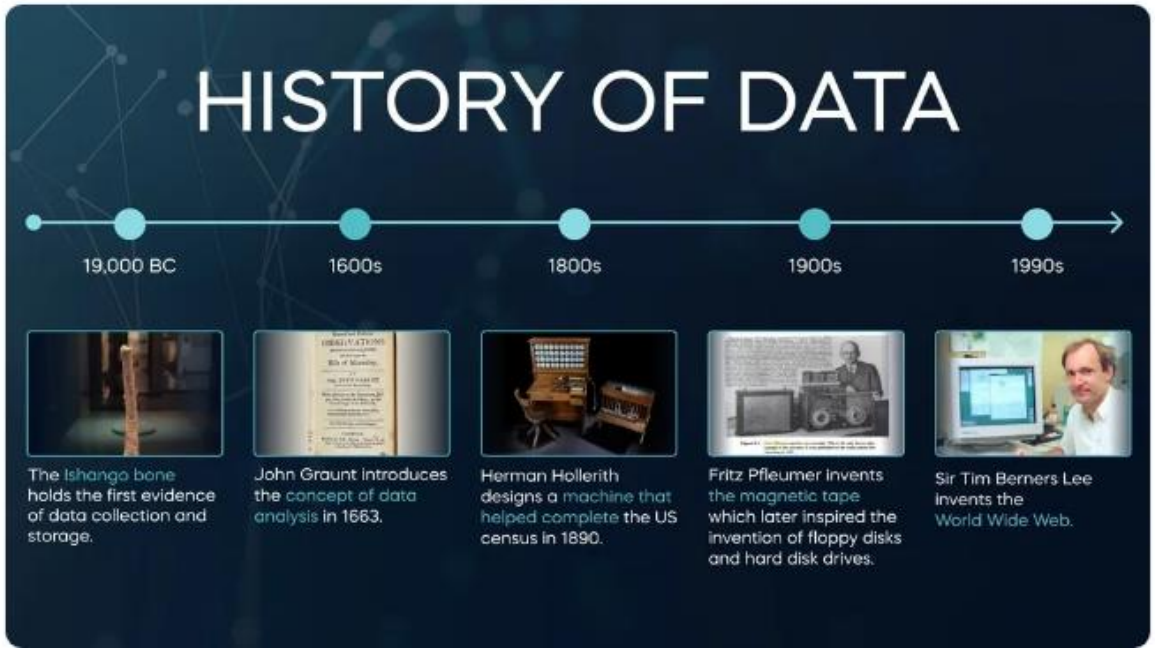
Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

- Early identification of risk to the product/services, if any
- Better operational efficiency

Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse.

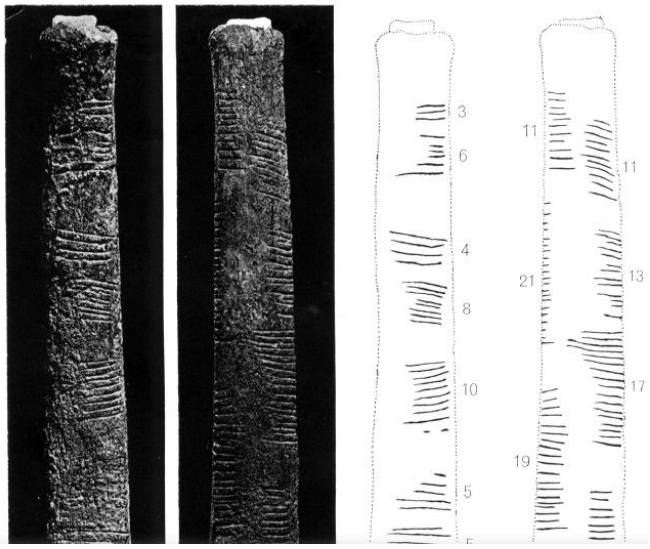
HISTORY OF BIG DATA:

The History of Data Timeline



History of Data in 19,000 BC: The Great Baboon

The first use of data goes back to 19,000 BC when our Palaeolithic(paashaan kaal) ancestors used a baboon tool called the *Ishango bone* to perform simple calculations. Back then, there were no calculators, pens, or even paper – as we know, those came much later



History of Data in 2400 BCE:

In 2400 BCE came, the abacus. The first dedicated device constructed specifically for performing calculations. The first libraries also appeared around this time, representing our first attempts at mass data storage.

History of Data in the 1640s: The Father of Public Health Statistics

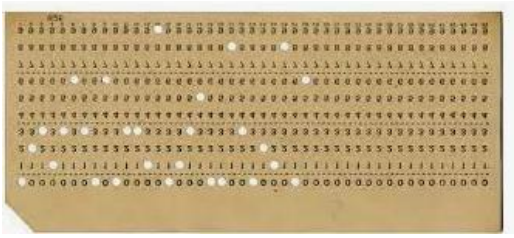
In the 1640s, John Graunt, a hat maker, started collecting information regarding deaths in London. He noted down statistics such as:

- The number of deaths
- The mortality rate among age groups
- The causes of death

In this way, he revolutionized how we use medical data to this day. In fact, Graunt was the first person to use data analysis to understand and solve a problem.

History of Data in the 1880s: The Era of Data Processing

One day back in the 1880s, the German-American statistician Herman Hollerith saw a train conductor punching train tickets for passengers. That's how the idea of using punch cards in writing and processing data was born. A punch card is normally a type of stiff paper, onto which a machine would create holes in specific locations.



History of Data in 1928: The Concept of Storage

Pfleumer's Magnetic Tape

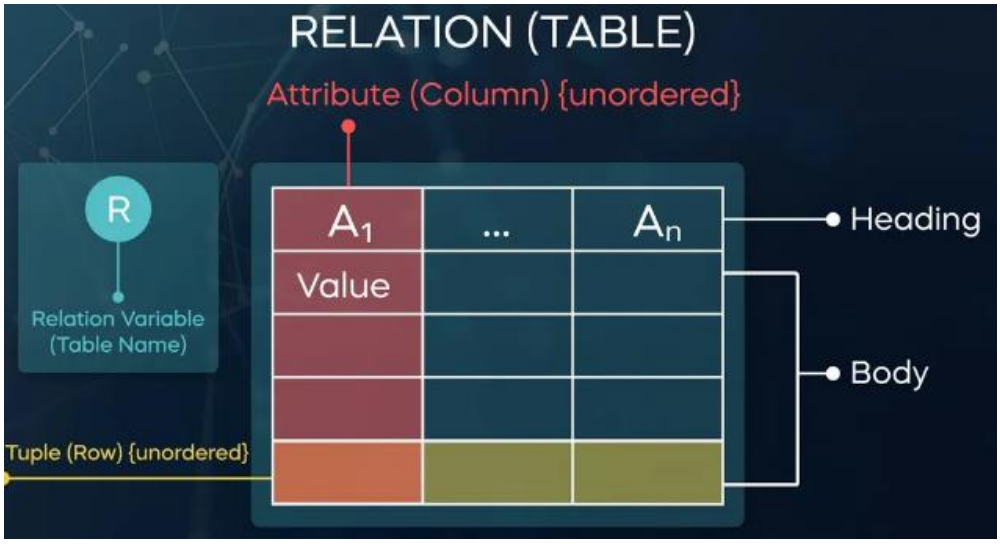
In 1928, the German engineer Fritz Pfleumer patented a magnetic tape that he used to replace wire recording for storing data.

The idea of storing information on magnetic tapes actually inspired the invention of floppy disks and hard-disk drives later on.

Codd's Relational Model 1960s:

The computer scientist Edgar Codd was the first one to introduce the idea of a relational database management system, which we know today as a "data table".

In the 1960s, he started working on a model that can describe data attributes in columns and their values in rows



Present-Day History of Data: The Internet Era

With the rise of the internet consequently comes the rise of big data, hypertext and hyperlinks. invented by Sir Tim Berners Lee made it easy to share information and connect resources.

And with the introduction of Google in 1997, data became even more widely available to everyone with access to a computer or mobile device.

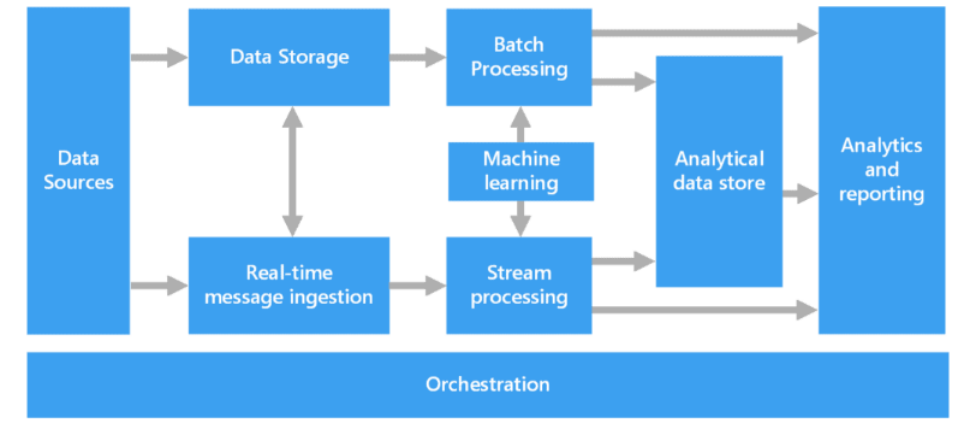
With every innovation in technology, data science, machine learning, or AI comes a new way of creating and spreading information.

Big Data Architecture:

The term "Big Data architecture" refers to the systems and software used to manage Big Data. A Big Data architecture must be able to handle the scale, complexity, and variety of Big Data.

So,Big data architecture is arranged to handle the ingestion, processing, and analysis of data that is huge or complicated for classical database systems.

Some Big Data Architecture Examples include - Azure Big Data architecture, Hadoop big data architecture, and Spark architecture in Big Data.



Data sources and integration

One of the fundamental components of big data architecture is data sources and ingestion. This component involves identifying and categorising the various data sources from which the organisation collects information. These sources can include structured data from databases, semi-structured data from sources like APIs and logs, and unstructured data from sources like social media and sensor data.

Technologies like Apache Kafka and Apache NiFi are commonly used for data ingestion, ensuring a smooth flow of data into the big data ecosystem.

Real-time message ingestion

- If the solution ingests real-time data, the architecture must consist of a way to capture and store real-time data for stream processing.
- This part of a streaming architecture is generally referred to as stream buffering. Options include Azure **IoT Hub, Azure Event Hubs, and Kafka.**

Data storage

Due to the massive size of data, traditional databases may not suffice. Big data storage solutions such as Data Lakes and Data Warehouses are employed to store raw and processed data.

Additionally, Data Lakes provide a flexible repository for storing both structured and unstructured data in its raw format, while Data Warehouses offer structured storage optimised for querying and analytics. NoSQL databases like MongoDB and Cassandra are also utilised for specific use cases, providing horizontal scalability and high performance for certain data types.

Data processing

It involves transforming and analysing the data to derive meaningful insights and patterns. Batch processing and real-time stream processing are two primary data processing approaches.

Additionally, batch processing deals with large sets of data at scheduled intervals, while stream processing handles data in real-time as it arrives. Technologies like Apache Hadoop and Apache Spark are commonly used for distributed data processing, enabling parallel computing and handling vast amounts of data efficiently.

Analytical datastore

- In this part the processed data in a structured format that can be queried using analytical tools.
- To analyze the data, the architecture contains a data modeling layer such as a tabular data model in Azure Analysis Services.
- It supports self-service BI, **Microsoft Power BI, or Microsoft Excel** for data visualization.

Orchestration

- To automate repeated data processing operations, we use an orchestration technology such as **Apache Oozie** or **Azure Data Factory and Sqoop.**

Data orchestration is an automated process for taking siloed data(A data silo is **a collection of data held by one group that is not easily or fully accessible by** other groups in the same organization.) from multiple storage locations, combining and organizing it, and making it available for analysis.

Big Data Technology

Big data technology is defined as software-utility. This technology is primarily designed to analyze, process and extract information from a large data set and a huge set of extremely complex structures.

Among the larger concepts of rage in technology, big data technologies are widely associated with many other technologies such as deep learning, machine learning, artificial intelligence (AI), and Internet of Things (IoT)

big data technologies are focused on analyzing and handling large amounts of real-time data and batch-related data.

Types of Big Data Technology:

(i) Operational Big Data Technologies:

The operational-big data includes daily basis data such as online transactions, social media platforms, and the data from any particular organization or a firm, which is usually needed for analysis using the software based on big data technologies. The data can also be referred to as raw data used as the input for several Analytical Big Data Technologies.

- Online ticket booking system, e.g., buses, trains, flights, and movies, etc.
- Online trading or shopping from e-commerce websites like Amazon, Flipkart, Walmart, etc.
- Online data on social media sites, such as Facebook, Instagram, Whatsapp, etc.

- The employees' data or executives' particulars in multinational companies.

(ii) Analytical Big Data Technologies

The actual investigation of big data that is important for business decisions falls under this type of big data technology.

- Stock marketing data
- Weather forecasting data and the time series analysis
- Medical health records where doctors can personally monitor the health status of an individual
- Carrying out the space mission databases where every information of a mission is very important

Top Big Data Technologies

We can categorize the leading big data technologies into the following four sections:

- Data Storage
- Data Mining
- Data Analytics
- Data Visualization

Data Storage & Analysis:

Big Data Technologies that come under Data Storage:

- **Hadoop**:
  Hadoop is one of the leading technologies to handle big data. This technology is based entirely on map-reduce architecture and is mainly used to process batch information. Also, it is capable enough to process tasks in batches. The Hadoop framework was mainly introduced to store and process data in a distributed data processing environment.
  Hadoop is also best suited for storing and analyzing the data from various machines with a faster speed and low cost. That is why Hadoop is known as one of the core components of big data technologies. The **Apache Software Foundation** introduced it in Dec 2011. Hadoop is written in Java programming language.

- **MongoDB**: MongoDB is another important component of big data technologies in terms of storage. No relational properties and RDBMS properties apply to MongoDb because it is a NoSQL database.
  The structure of the data storage in MongoDB is also different from traditional RDBMS databases. This enables MongoDB to hold massive amounts of data.
  The database in MongoDB uses documents similar to JSON with the schema.

MongoDB Inc. introduced MongoDB in Feb 2009. It is written with a combination of C++, Python, JavaScript, and Go language.

- **RainStor**: RainStor is a popular database management system designed to manage and analyze organizations' Big Data requirements. It uses deduplication strategies (data deduplication is a *technique for eliminating duplicate copies of repeating data*.)that help manage storing and handling vast amounts of data for reference. RainStor was designed in 2004 by a **RainStor Software Company**. It operates just like SQL. Companies such as Barclays and Credit Suisse are using RainStor for their big data needs.
- **Hunk**: Hunk allows us to report and visualize vast amounts of data from Hadoop and NoSQL data sources.
  Hunk was introduced in 2013 by **Splunk Inc**. It is based on the Java programming language.
- **Cassandra**: Cassandra is one of the leading big data technologies among the list of top NoSQL databases. It is open-source, distributed and has extensive column storage options. It is freely available and provides high availability without fail.

  Cassandra's essential features include fault-tolerant mechanisms, scalability, MapReduce support, distributed nature, , query language property, and multi-datacenter replication etc.
  Cassandra was developed in 2008 by the **Apache Software Foundation** for the Facebook inbox search feature. It is based on the Java programming language.

Data Mining

Presto: Presto is an open-source and a distributed SQL query engine developed to run interactive analytical queries against huge-sized data sources. The size of data sources can vary from gigabytes to petabytes. Presto helps in querying the data in Cassandra, Hive, relational databases and proprietary data storage systems.

RapidMiner: RapidMiner is defined as the data science software that offers us a very robust and powerful graphical user interface to create, deliver, manage, and maintain predictive analytics.

ElasticSearch: When it comes to finding information, elasticsearch is known as an essential tool.

Also, it provides a purely distributed search engine. ElasticSearch is primarily written in a Java programming language and was developed in 2010 by **Shay Banon**.

Data Analytics:

How does big data analytics work?

Big Data Analytics is a powerful tool which helps to find the potential of large and complex datasets.

- **Data Collection:** Data is the heart of Big Data Analytics. It is the process of the collection of data from various sources, which can include customer reviews, surveys, sensors, social media etc. The main goal of data collection is to gather as much relevant data as possible. The more data, the richer the insights.
- **Data Cleaning (Data Preprocessing):** Once we have the data, it often needs some cleaning. This process involves identifying and dealing with missing values, correcting errors, and removing duplicates.
- **Data Processing:** Next, we need to process the data. This involves different steps like organizing, structuring, and formatting it in a way that makes it appropriate for analysis.
- **Data Analysis:** Data analysis is performed using various statistical, mathematical, and machine learning techniques to extract valuable insights from the processed data. For instance, it can reveal customer preferences, market trends, or patterns in healthcare data
- **Data Visualization:** Data analysis results are often presented in the form of visualizations – charts, graphs, and interactive dashboards. These visual representations make complex data easy to understand and enable decision-makers to see trends and patterns at a glance.
- **Data Storage and Management:** Storing and managing the analyzed data is crucial. It's like archiving your findings. You may want to revisit the insights in the future, and well-organized storage is essential for that. Additionally, it's important to ensure data security and compliance with regulations during this critical step.

- **Continuous Learning and Improvement:** Big Data Analytics isn't a one-time process, its an ongoing process. As you collect and analyze more data, you learn more about your operations or customers.

Types of Big Data Analytics

Big Data Analytics comes in many different types, each serving a different purpose:

1. **Descriptive Analytics:** This type helps us understand past events. In social media, it shows performance metrics, like the number of likes on a post.
2. **Diagnostic Analytics:** In Diagnostic analytics analyse deeper to uncover the reasons behind past events. In healthcare, it identifies the causes of high patient re-admissions.
3. **Predictive Analytics:** Predictive analytics forecasts future events based on past data. Weather forecasting, for example, predicts tomorrow's weather by analyzing historical patterns.
4. **Prescriptive Analytics:** This type not only predicts outcomes but also suggests actions to optimize them. In e-commerce, it might recommend the best price for a product to maximize profits.
5. **Real-time Analytics:** Real-time analytics processes data instantly. In stock trading, it helps traders make quick decisions based on current market conditions.
6. **Spatial Analytics:** Spatial analytics focuses on location data. For city planning, it optimizes traffic flow using data from sensors and cameras to reduce congestion.
7. **Text Analytics:** Text analytics extracts insights from unstructured text data. In the hotel industry, it can analyze guest reviews to improve services and guest satisfaction.

Big Data Analytics Technologies and Tools

Big Data Analytics relies on various technologies and tools that might sound complex, let's simplify them:

- **Hadoop:** Imagine Hadoop as an enormous digital warehouse. It's used by companies like Amazon to store tons of data efficiently. For instance, when Amazon suggests products you might like, it's because Hadoop helps manage your shopping history.

- **Spark:** Think of Spark as the super-fast data chef. Netflix uses it to quickly analyze what you watch and recommend your next binge-worthy show.

- **NoSQL Databases:** NoSQL databases, like MongoDB, are like digital filing cabinets that Airbnb uses to store your booking details and user data. These databases are famous because of their quick and flexible, so the platform can provide you with the right information when you need it.

- **Tableau:** Tableau is like an artist that turns data into beautiful pictures. The World Bank uses it to create interactive charts and graphs that help people understand complex economic data.
- **Python and R:** Python and R are like magic tools for data scientists. They use these languages to solve tricky problems. For example, Kaggle uses them to predict things like house prices based on past data.
- **Machine Learning Frameworks (e.g., TensorFlow):** In Machine learning frameworks are the tools who make predictions. Airbnb uses TensorFlow to predict which properties are most likely to be booked in certain areas. It helps hosts make smart decisions about pricing and availability.

Big data analytics benefits

The benefits of using big data analytics include the following:

- **Real-time intelligence.** Organizations can quickly analyze large amounts of real-time data from different sources, in many different formats and types.
- **Better-informed decisions.** Effective strategizing can benefit and improve the supply chain, operations and other areas of strategic decision-making.
- **Cost savings.** This can result from new business process efficiencies and optimizations.
- **Better customer engagement.** A better understanding of customer needs, behavior and sentiment can lead to better marketing insights and provide information for product development.
- **Optimize risk management strategies.** Big data analytics improve risk management strategies by enabling organizations to address threats in real time.

Big data analytics challenges

Despite the wide-reaching benefits that come with using big data analytics, its use also comes with the following challenges:

- **Data accessibility.** With larger amounts of data, storage and processing become more complicated. Big data should be stored and maintained properly to ensure it can be used by less experienced data scientists and analysts.
- **Data quality maintenance.** With high volumes of data coming in from a variety of sources and in different formats, data quality management for big data requires significant time, effort and resources to properly maintain it.
    - **Data security.** The complexity of big data systems presents unique security challenges. Properly addressing security concerns within such a complicated big data ecosystem can be a complex undertaking.
    - **Choosing the right tools.** Selecting from the vast array of big data analytics tools and platforms available on the market can be confusing, so organizations must know how to pick the best tool that aligns with users' needs and infrastructure.
    - **Talent shortages.** With a potential lack of internal analytics skills and the high cost of hiring experienced data scientists and engineers, some organizations are finding it hard to fill the gaps.

Usage of Big Data Analytics

Big Data Analytics has a significant impact in various sectors:

- **Healthcare:** It aids in precise diagnoses and disease prediction, elevating patient care.
- **Retail:** Amazon's use of Big Data Analytics offers personalized product recommendations based on your shopping history, creating a more tailored and enjoyable shopping experience.
- **Finance:** Credit card companies such as Visa rely on Big Data Analytics to swiftly identify and prevent fraudulent transactions, ensuring the safety of your financial assets.
- **Transportation:** Companies like Uber use Big Data Analytics to optimize drivers' routes and predict demand, reducing wait times and improving overall transportation experiences.
- **Agriculture:** Farmers make informed decisions, boosting crop yields while conserving resources.
- **Manufacturing:** Companies like General Electric (GE) use Big Data Analytics to predict machinery maintenance needs, reducing downtime and enhancing operational efficiency.

Data Intelligence

Data intelligence is the use of various tools and methods to analyze and transform data into information from which valuable insight can be drawn.

Intelligent data is a core component of big data and business intelligence. Intelligent data processing provides a strong data foundation, restructuring and enhancing big datasets that AI uses; cleanses and transforms data into information that is valuable and relevant to business performance; enables businesses to identify patterns, make informed decisions, and adapt to new information; and incorporate advanced analytics techniques to enhance visualized prescriptive and predictive analytics.

Intelligent Data Analysis

Intelligent Data Analysis (IDA) is an interdisciplinary study that is concerned with the extraction of useful knowledge from data, drawing techniques from a variety of fields, such as artificial intelligence, high-performance computing, pattern recognition, and statistics.

Intelligent data analysis refers to the use of analysis, classification, conversion, extraction organization, and reasoning methods to extract useful knowledge from data.

Big Data Analytic Process:

- **Data Collection:** Data is the heart of Big Data Analytics. It is the process of the collection of data from various sources, which can include customer reviews, surveys, sensors, social media etc. The main goal of data collection is to gather as much relevant data as possible. The more data, the richer the insights.
- **Data Cleaning (Data Preprocessing):** Once we have the data, it often needs some cleaning. This process involves identifying and dealing with missing values, correcting errors, and removing duplicates. It's like sifting through a treasure trove to remove any rocks or debris, leaving only the valuable gems behind.
- **Data Processing:** Next, we need to process the data. This involves different steps like organizing, structuring, and formatting it in a way that makes it appropriate for analysis. Think of it like a chef preparing ingredients before cooking. Data processing makes the raw data more digestible for analytics tools.
- **Data Analysis:** Data analysis is performed using various statistical, mathematical, and machine learning techniques to extract valuable insights from the processed data. For

instance, it can reveal customer preferences, market trends, or patterns in healthcare data.

- **Data Visualization:** Data analysis results are often presented in the form of visualizations – charts, graphs, and interactive dashboards. These visual representations make complex data easy to understand and enable decision-makers to see trends and patterns at a glance.

- **Data Storage and Management:** Storing and managing the analyzed data is crucial. It's like archiving your findings. You may want to revisit the insights in the future, and well-organized storage is essential for that. Additionally, it's important to ensure data security and compliance with regulations during this critical step.
- **Continuous Learning and Improvement:** It's an ongoing process. As we collect and analyze more data, we learn more about your operations or customers. This insight can lead to refining data collection methods and analysis techniques for better results. **Big Data Analytics** is about collecting, cleaning, processing, and analyzing data to uncover valuable insights. It's a multi-step process that transforms raw data into fruitful insights.

Big Data Analytics Tools

- **Hadoop:** Imagine Hadoop as an enormous digital warehouse. It's used by companies like Amazon to store tons of data efficiently. For instance, when Amazon suggests products you might like, it's because Hadoop helps manage your shopping history.

- **Spark:** Think of Spark as the super-fast data chef. Netflix uses it to quickly analyze what you watch and recommend your next binge-worthy show.

- **NoSQL Databases:** NoSQL databases, like MongoDB, are like digital filing cabinets that Airbnb uses to store your booking details and user data. These databases are famous because of their quick and flexible, so the platform can provide you with the right information when you need it.

- **Tableau:** Tableau is like an artist that turns data into beautiful pictures. The World Bank uses it to create interactive charts and graphs that help people understand complex economic data.
- **Python and R:** Python and R are like magic tools for data scientists. They use these languages to solve tricky problems. For example, Kaggle uses them to predict things like house prices based on past data.
- **Machine Learning Frameworks (e.g., TensorFlow):** In Machine learning frameworks are the tools who make predictions. Airbnb uses TensorFlow to predict which properties are most likely to be booked in certain areas. It helps hosts make smart decisions about pricing and availability.

Ques: What is analysis? What is reporting? Differentiate between analysis and reporting?

- **Analysis:**

Analysis involves the systematic examination and evaluation of data or information to uncover patterns, trends, relationships, or insights. It typically involves:

- **Data Examination**: Reviewing and exploring raw data to understand its structure, quality, and content.
- **Pattern Recognition**: Identifying trends, correlations, anomalies, or other significant patterns within the data.
- **Interpretation**: Drawing conclusions, generating hypotheses based on the observed patterns.
- **Modeling**: Using statistical, mathematical, or computational techniques to analyze and make predictions based on the data.
- **Insight Generation**: Deriving actionable insights or recommendations from the analysis to inform decision-making.

Analysis often requires specialized skills in statistics, data science, domain knowledge, and critical thinking to extract meaningful insights from complex datasets.

- **Reporting**:

Reporting involves the process of presenting the findings, results, or insights derived from analysis in a structured and understandable format. It typically includes:

- **Summarization**: Condensing the key findings and insights from the analysis into a concise format.
- **Visualization**: Representing data and analysis results visually through charts, graphs, tables, or other graphical elements to aid comprehension.
- **Contextualization**: Providing background information, context, and explanations to help stakeholders understand the significance of the findings.
- **Communication**: Clearly and effectively conveying the analysis results and their implications to relevant stakeholders, such as decision-makers, clients, or team members.
- **Actionable Recommendations**: Offering actionable recommendations or next steps based on the insights to guide future actions or strategies.

Reporting often involves creating dashboards, presentations, written reports, or interactive tools to communicate the analysis findings effectively.

Differences between Analysis & Reporting:

- **Focus**: Analysis primarily focuses on exploring and understanding data to derive insights, while reporting focuses on presenting those insights in a clear and understandable manner.
- **Activities**: Analysis involves data exploration, pattern recognition, interpretation, and insight generation, whereas reporting involves summarization, visualization, contextualization, and communication.
- **Skills**: Analysis typically requires skills in statistics, data science, programming, and domain knowledge, while reporting often requires skills in data visualization, communication, and storytelling.
- **Output**: The output of analysis is insights, hypotheses, or predictive models, while the output of reporting is typically reports, presentations, dashboards, or visualizations.

Ques: what is Data? What are the properties of Data?

Ans: Data refers to raw facts, figures, observations, or information that can be collected, recorded, and analyzed. It can take various forms, including text, numbers, images, audio, video, and more. Data serves as the foundation for generating insights, making decisions, and solving problems in various domains.

Here are some key properties of data:

1. **Accuracy**: Data should be free from errors or inaccuracies and should reflect the true state of what it represents. Ensuring data accuracy is crucial for making reliable decisions and drawing meaningful conclusions.
2. **Completeness**: Data should contain all the necessary information required for its intended purpose. Incomplete data may lead to biased analysis or incorrect conclusions.
3. **Consistency**: Data should be consistent across different sources, time periods, and contexts. Inconsistencies can arise due to discrepancies in data collection methods, definitions, or standards.
4. **Relevance**: Data should be relevant to the problem or question at hand. Including irrelevant data can add noise and complexity to analysis, making it harder to derive meaningful insights.
5. **Timeliness**: Data should be collected and made available in a timely manner to ensure its relevance and usefulness. Outdated data may no longer reflect the current state of affairs and can lead to outdated or incorrect conclusions.
6. **Validity**: Data should measure what it claims to measure and be valid for its intended purpose. Validity ensures that data accurately represents the concepts or phenomena it is intended to capture.
7. **Precision**: Data should be precise and granular enough to support the desired level of analysis or decision-making. Precision refers to the level of detail or specificity contained in the data.
8. **Accessibility**: Data should be easily accessible to authorized users or stakeholders who need it for analysis, reporting, or decision-making purposes. Accessibility ensures that data can be leveraged effectively to drive insights and actions.
9. **Security**: Data should be protected from unauthorized access, manipulation, or loss to maintain its integrity and confidentiality. Implementing proper security measures helps safeguard sensitive or valuable data assets.
10. **Interpretability**: Data should be interpretable and understandable to users who need to analyze or make decisions based on it. Clear documentation, metadata, and contextual information can enhance the interpretability of data.

Ques: What is conventional system? List some of the challenges of conventional system?

Ans: A conventional system typically refers to traditional or established methods, practices, or systems that have been in use for a significant period and are widely accepted within a particular context or industry. These systems often rely on manual processes, paper-based documentation, and legacy technologies. While conventional systems have served their purposes well in the past, they may face several challenges, including:

1. **Limited Efficiency**: Conventional systems often involve manual processes and paper-based documentation, leading to inefficiencies, delays, and higher operational costs compared to automated or digital systems.

2. **Data Redundancy and Inconsistency**: With conventional systems, data may be duplicated across multiple paper documents or stored in disparate formats, leading to redundancy and inconsistency. This can make it difficult to maintain data integrity and ensure accuracy.
3. **Lack of Integration**: Conventional systems may operate in silos, with limited integration between different departments, processes, or systems. This lack of integration can hinder communication, collaboration, and data sharing across the organization.
4. **Limited Scalability**: Conventional systems may struggle to scale and accommodate growing volumes of data, transactions, or users. This can pose challenges for organizations experiencing rapid growth or expansion.
5. **Risk of Errors and Compliance Issues**: Manual data entry and paper-based processes in conventional systems are prone to human errors, which can impact data accuracy and compliance with regulations or industry standards.
6. **Difficulty in Accessing and Analyzing Data**: Retrieving and analyzing data from conventional systems can be time-consuming and labor-intensive, especially if data is stored in paper-based formats or across multiple disparate systems.
7. **Security Vulnerabilities**: Conventional systems may lack robust security measures to protect against cyber threats, data breaches, or unauthorized access. This can expose sensitive information to risks and compromise data confidentiality.
8. **Resistance to Change**: Introducing new technologies or modernizing conventional systems may face resistance from employees accustomed to existing processes and systems. Overcoming resistance to change can be a significant challenge in transitioning to more efficient or digital systems.
9. **High Maintenance Costs**: Maintaining and supporting legacy or conventional systems can be costly, especially as they age and require frequent updates, patches, or repairs to remain functional and secure.
10. **Inability to Meet Evolving Customer Expectations**: Conventional systems may struggle to adapt to changing customer preferences, expectations, or market trends. This can hinder organizations' ability to deliver seamless and personalized experiences to their customers.

Ques: What are the big data technology components what are the big data technology components?

Ans: Big data technology encompasses various components that work together to store, process, analyze, and visualize large volumes of data. Some key components include:

1. **Storage Systems**:
    - **Hadoop Distributed File System (HDFS)**: A distributed file system that provides high-throughput access to application data.

- **NoSQL Databases**: Such as Apache Cassandra, MongoDB, Couchbase, etc., which are optimized for handling large volumes of unstructured or semi-structured data.

2. **Processing Frameworks**:
   - **Apache Hadoop**: An open-source framework for distributed storage and processing of large datasets using a MapReduce programming model.
   - **Apache Spark**: A fast and general-purpose cluster computing system that provides in-memory processing capabilities for large-scale data processing and analytics.
   - **Apache Flink**: A stream processing framework for real-time data processing and analytics.
3. **Data Ingestion**:
   - **Apache Kafka**: A distributed streaming platform that is used for building real-time data pipelines and streaming applications.
   - **Apache NiFi**: A data flow automation tool that facilitates the movement of data between disparate systems.
4. **Data Warehousing**:
   - **Amazon Redshift, Google BigQuery, Snowflake**: Cloud-based data warehousing solutions that enable high-performance querying and analytics on large datasets.
5. **Data Visualization and Business Intelligence (BI)**:
   - **Tableau, Power BI, Qlik**: Tools that provide interactive data visualization and business intelligence capabilities for analyzing and presenting insights from large datasets.
6. **Machine Learning and Advanced Analytics**:
   - **Apache Mahout, TensorFlow, PyTorch**: Frameworks for building and deploying machine learning models at scale for big data analytics.

- **Spark MLlib**: Machine learning library built on top of Apache Spark for scalable machine learning algorithms.

7. **Resource Management and Orchestration**:
   - **Apache YARN**: Resource management and job scheduling in Hadoop clusters.
   - **Apache Mesos, Kubernetes**: Container orchestration platforms that provide efficient resource utilization and management for big data workloads.
8. **Data Governance and Security**:
   - **Apache Ranger, Apache Sentry**: Frameworks for managing access control and enforcing security policies in Hadoop ecosystems.
   - **Cloudera Navigator, Apache Atlas**: Metadata management and governance tools for tracking and managing data lineage, classification, and security policies.

Ques: What is Big Data Security? What are the steps for securing Big Data?

Ans: Big data security refers to the set of measures, processes, and technologies designed to protect the confidentiality, integrity, and availability of big data assets and infrastructure.

Here are the steps typically involved in securing big data:

1. **Identify and Classify Data**: Understand the types of data being collected, processed, and stored within the big data infrastructure. Classify data based on sensitivity and regulatory requirements.
2. **Access Control**: Implement strong access controls to restrict access to sensitive data based on roles and privileges. Use authentication mechanisms such as LDAP, Kerberos, or OAuth for user authentication.
3. **Encryption**: Encrypt data at rest using techniques such as disk encryption or database-level encryption. Implement encryption for data in transit using protocols like SSL/TLS for network communication.
4. **Network Security**: Secure the network infrastructure by implementing firewalls, intrusion detection/prevention systems (IDS/IPS), and network segmentation to protect against unauthorized access and network-based attacks.

5. **Monitoring and Auditing**: Implement logging and monitoring solutions to track user activities, system events, and data access. Perform regular audits to detect and investigate suspicious activities or security breaches.

6. **Data Governance and Compliance**: Establish data governance policies and procedures to ensure compliance with relevant regulations such as GDPR, HIPAA, PCI-DSS, etc.

7. **Secure Development Practices**: Follow secure coding practices and conduct regular security reviews and code audits for big data applications and infrastructure components.

8. **Patch Management**: Keep the big data infrastructure and software components up-to-date with security patches and updates to address vulnerabilities and mitigate security risks.

9. **Disaster Recovery and Business Continuity**: Develop and implement disaster recovery and business continuity plans to ensure the availability and resilience of big data infrastructure in the event of natural disasters, hardware failures, or cyber attacks.

10. **Employee Training and Awareness**: Provide regular training and awareness programs to educate employees about security best practices, data handling procedures, and potential security threats.