INTRODUCTION TO BIG DATA

What is Big Data

Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time. These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.

The amount and availability of data is growing rapidly, spurred on by digital technology advancements, such as connectivity, mobility, the Internet of Things (IoT), and artificial intelligence (AI). As data continues to expand and proliferate, new big data tools are emerging to help companies collect, process, and analyze data at the speed needed to gain the most value from it.

Big data describes large and diverse datasets that are huge in volume and also rapidly grow in size over time. Big data is used in machine learning, predictive modeling, and other advanced analytics to solve business problems and make informed decisions

The Vs of big data

Big data definitions may vary slightly, but it will always be described in terms of volume, velocity, and variety. These big data characteristics are often referred to as the "3 Vs of

- **Volume**
  As its name suggests, the most common characteristic associated with big data is its high volume. This describes the enormous amount of data that is available for collection and produced from a variety of sources and devices on a continuous basis.
- **Velocity**
  Big data velocity refers to the speed at which data is generated. Today, data is often produced in real time or near real time, and therefore, it must also be

  processed, accessed, and analyzed at the same rate to have any meaningful impact.

- **Variety**
  Data is heterogeneous, meaning it can come from many different sources and can be structured, unstructured, or semi-structured. More traditional structured data (such as data in spreadsheets or relational databases) is now supplemented by unstructured text, images, audio, video files, or semi- structured formats like sensor data that can't be organized in a fixed data schema. big data" and were first defined by Gartner in 2001.

In addition to these three original Vs, three others that are often mentioned in relation to harnessing the power of big data: **veracity**, **variability**, and **value**.

- **Veracity**:

  Big data can be messy, noisy, and error-prone, which makes it difficult to control the quality and accuracy of the data. Large datasets can be unwieldy and confusing, while smaller datasets could present an incomplete picture. The higher the veracity of the data, the more trustworthy it is.

- **Variability:**

  The meaning of collected data is constantly changing, which can lead to inconsistency over time. These shifts include not only changes in context and interpretation but also data collection methods based on the information that companies want to capture and analyze.

- **Value:**

  It's essential to determine the business value of the data you collect. Big data must contain the right data and then be effectively analyzed in order to yield insights that can help drive decision-making.

Sources of Big Data

These data come from many sources like

- **Social networking sites:** Facebook, Google, LinkedIn all these sites generate huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge number of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

How does big data work?

The central concept of big data is that the more visibility you have into anything, the more effectively you can gain insights to make better decisions, uncover growth opportunities, and improve your business model.

Making big data work requires three main actions:

1. **Integration:**

Big data collects terabytes, and sometimes even petabytes, of raw data from many sources that must be received, processed, and transformed into the format that business users and analysts need to start analyzing it.

2. **Management:**

Big data needs big storage, whether in the cloud, on-premises, or both. Data must also be stored in whatever form required. It also needs to be processed and made available in real time. Increasingly, companies are turning to cloud solutions to take advantage of the unlimited compute and scalability.

3. **Analysis:**

The final step is analyzing and acting on big data—otherwise, the investment won't be worth it. Beyond exploring the data itself, it's also

critical to communicate and share insights across the business in a way that everyone can understand. This includes using tools to create data visualizations like charts, graphs, and dashboards.

What is big data analytics?

Big data analytics is the process of collecting, examining, and analysing large amounts of data to discover market trends, insights, and patterns that can help companies make better business decisions. This information is available quickly and efficiently so that companies can be agile in crafting plans to maintain their competitive advantage.

Big data analytics is important because it helps companies leverage their data to identify opportunities for improvement and optimisation. Across different business segments, increasing efficiency leads to overall more intelligent operations, higher profits, and satisfied customers. Big data analytics helps companies reduce costs and develop better, customer-centric products and services.

Technologies such as business intelligence (BI) tools and systems help organisations take unstructured and structured data from multiple sources. Users (typically employees) input queries into these tools to understand business operations and performance. Big data analytics uses the four data analysis methods to uncover meaningful insights and derive solutions.

Types of big data analytics

Four main types of big data analytics support and inform different business decisions.

1. Descriptive analytics

Descriptive analytics refers to data that can be easily read and interpreted. This data helps create reports and visualise information that can detail company profits and sales.

Example: During the pandemic, a leading pharmaceutical company conducted data analysis on its offices and research labs. Descriptive analytics helped them identify consolidated unutilised spaces and departments, saving the company millions of pounds.

2. Diagnostics analytics

Diagnostics analytics helps companies understand why a problem occurred. Big data technologies and tools allow users to mine and recover data that helps dissect an issue and prevent it from happening in the future.

Example: An online retailer's sales have decreased even though customers continue to add items to their shopping carts. Diagnostics analytics helped to understand that the payment page was not working correctly for a few weeks.

3. Predictive analytics

Predictive analytics looks at past and present data to make predictions. With artificial intelligence (AI), machine learning, and data mining, users can analyse the data to predict market trends.

Example: In the manufacturing sector, companies can use algorithms based on historical data to predict if or when a piece of equipment will malfunction or break down.

4. Prescriptive analytics

Prescriptive analytics solves a problem, relying on AI and machine learning to gather and use data for risk management.

Example: Within the energy sector, utility companies, gas producers, and pipeline owners identify factors that affect the price of oil and gas to hedge risks.

Benefits of big data analytics

Incorporating big data analytics into a business or organisation has several advantages. These include:

Cost reduction: Big data can reduce costs in storing all business data in one place. Tracking analytics also helps companies find ways to work more efficiently to cut costs wherever possible.

Product development: Developing and marketing new products, services, or brands is much easier when based on data collected from customers' needs and wants. Big data analytics also helps businesses understand product viability and to keep up with trends.

Strategic business decisions: The ability to constantly analyse data helps businesses make better and faster decisions, such as cost and supply chain optimisation.

Customer experience: Data-driven algorithms help marketing efforts (targeted ads, for example) and increase customer satisfaction by delivering an enhanced customer experience.

Risk management: Businesses can identify risks by analysing data patterns and developing solutions for managing those risks.

UNSTRUCTURED DATA

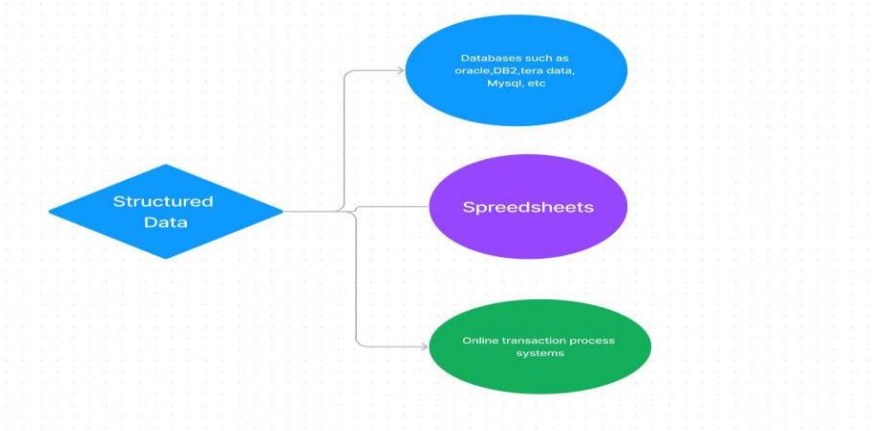
Types of Big Data

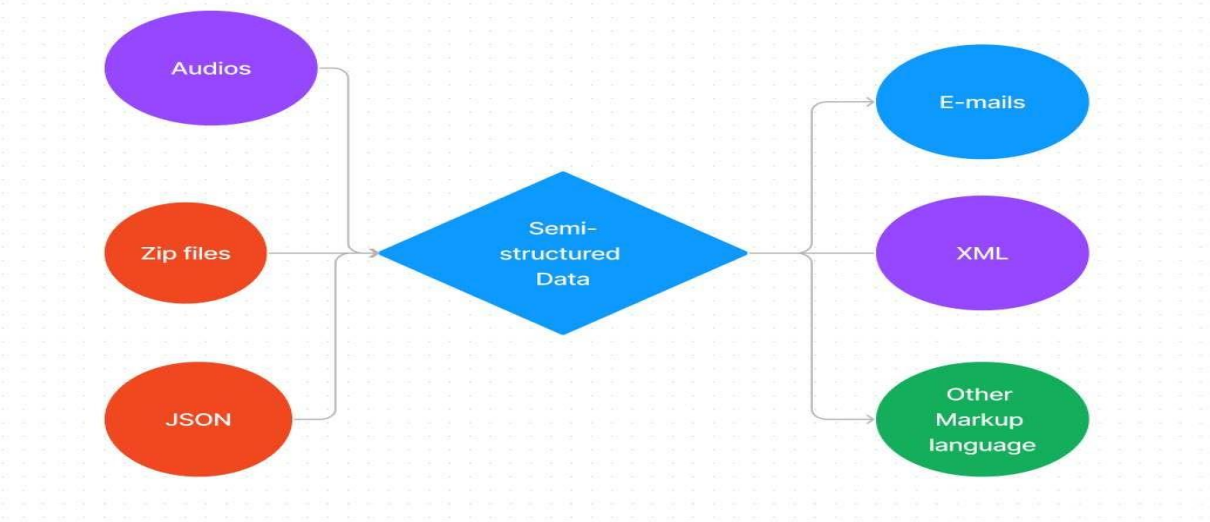
All data cannot be stored in the same way. The methods for data storage can be accurately evaluated after the type of data has been identified

1. Structured data

Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns

all data which can be stored in database in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. Example: Relational data.





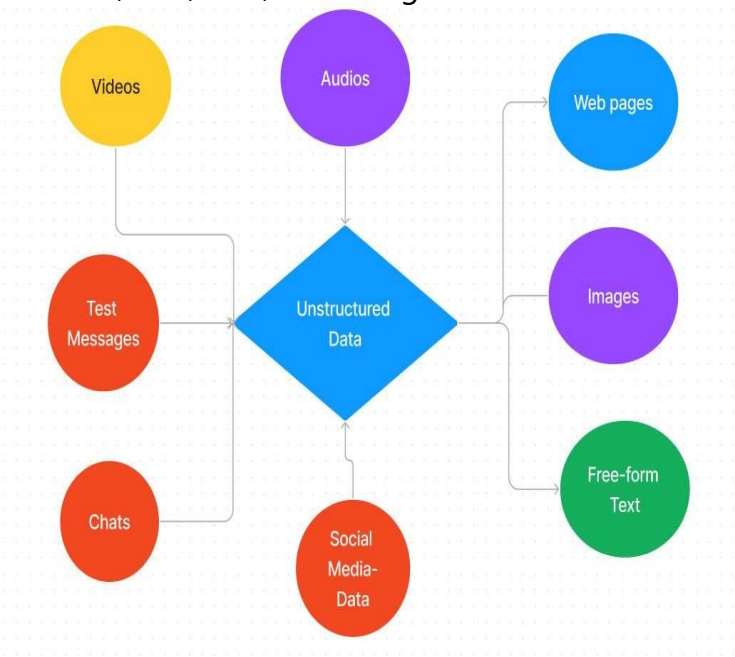
2. Semi-Structured data

Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to analyze. With some processes, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. Example: XML data.

3. Unstructured data

Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream

relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs.



Unstructured data is the data which does not conforms to a data model and has no easily identifiable structure such that it can not be used by a computer program easily. Unstructured data is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database.

From 80% to 90% of data generated and collected by organizations is unstructured, and its volumes are growing rapidly — many times faster than the rate of growth for structured databases.

Unstructured data stores contain a wealth of information that can be used to guide business decisions. However, unstructured data has historically been very difficult to analyze. With the help of AI and machine learning, new software tools are emerging that can search through vast quantities of it to uncover beneficial and actionable business intelligence.

Unstructured data vs. structured data

Let's take structured data first: it's usually stored in a relational database or RDBMS, and is sometimes referred to as relational data. It can be easily mapped into designated fields — for example, fields for zip codes, phone numbers, and credit cards. Data that conforms to RDBMS structure is easy to search, both with human-defined queries and with software.

Unstructured data, in contrast, doesn't fit into these sorts of pre-defined data models. It can't be stored in an RDBMS. And because it comes in so many formats, it's a real challenge for conventional software to ingest, process, and analyze. Simple content searches can be undertaken across textual unstructured data with the right tools.

Beyond that, the lack of consistent internal structure doesn't conform to what typical data mining systems can work with. As a result, companies have largely been unable to tap into value-laden data like customer interactions, rich media, and social network conversations. Robust tools for doing so are only now being developed and commercialized.

What are some examples of unstructured data?

Unstructured data can be created by people or generated by

machines. Here are some examples of the *human-generated*

variety:

- Email: Email message fields are unstructured and cannot be parsed by traditional analytics tools. That said, email metadata affords it some structure, and explains why email is sometimes considered semi-structured data.
- Text files: This category includes word processing documents, spreadsheets, presentations, email, and log files.
- Social media and websites: data from social networks like Twitter, LinkedIn, and Facebook, and websites such as Instagram, photo-sharing sites, and YouTube.
- Mobile and communications data: For this category, look no further than text messages, phone recordings, collaboration software, chat, and instant messaging.
- Media: This data includes digital photos, audio, and

video files. Here are some examples of *unstructured data*

generated by machines:

- Scientific data: This includes oil and gas surveys, space exploration, seismic imagery, and atmospheric data.
- Digital surveillance: This category features data like reconnaissance photos and videos.
- Satellite imagery: This data includes weather data, land forms, and military movements.

  le business intelligence.

Characteristics of Unstructured Data:

- Data neither conforms to a data model nor has any structure.
- Data cannot be stored in the form of rows and columns as in Databases
- Data does not follow any semantic or rules
- Data lacks any particular format or sequence
- Data has no easily identifiable structure
- Due to lack of identifiable structure, it cannot used by computer programs easily

Sources of Unstructured Data:

- Web pages
- Images (JPEG, GIF, PNG, etc.)
- Videos
- Memos
- Reports
- Word documents and PowerPoint presentations
- Surveys

Advantages of Unstructured Data:

- Its supports the data which lacks a proper format or sequence

- The data is not constrained by a fixed schema
- Very Flexible due to absence of schema.
- Data is portable
- It is very scalable
- It can deal easily with the heterogeneity of sources.
- These types of data have a variety of business intelligence and analytics applications.

Disadvantages of Unstructured data:

- It is difficult to store and manage unstructured data due to lack of schema and structure
- Indexing the data is difficult and error prone due to unclear structure and not having pre-defined attributes. Due to which search results are not very accurate.
- Ensuring security to data is difficult task.

Problems faced in storing unstructured data:

- It requires a lot of storage space to store unstructured data.
- It is difficult to store videos, images, audios, etc.
- Due to unclear structure, operations like update, delete and search is very difficult.
- Storage cost is high as compared to structured data
- Indexing the unstructured data is difficult

Possible solution for storing Unstructured data:

- Unstructured data can be converted to easily manageable formats
- using Content addressable storage system (CAS) to store unstructured data. It stores data based on their metadata and a unique name is assigned to every object stored in it. The object is retrieved based on content not its location.
- Unstructured data can be stored in XML format.

- Unstructured data can be stored in RDBMS which supports BLOBs

Extracting information from unstructured Data:

unstructured data do not have any structure. So it cannot easily interpreted by conventional algorithms. It is also difficult to tag and index unstructured data. So extracting information from them is tough job. Here are possible solutions:

- Taxonomies or classification of data helps in organising data in hierarchical structure. Which will make search process easy.

Data can be stored in virtual repository and be automatically tagged. For example Documentum.

- Use of application platforms like XOLAP. XOLAP helps in extracting information from e-mails and XML based documents
- Use of various data mining tools