**What is Big Data?**

Think of following as per 2020 numbers

* Every second, there are around 822 tweets on Twitter
* Every minute, nearly 510 comments are posted, 2,93,000 statuses are updated, and 1,36,000 photos are uploaded on Facebook
* Every hour, Walmart, a global discount departmental store chain, handles more than 1 million customer transactions.
* Every day, consumers make around 11.5 million payments by using PayPal

We live in a digital world where data is increasing rapidly because of the ever-increasing use of the Internet, sensors, and heavy machines at a very high rate. The sheer volume, variety, velocity, and veracity of such data is signified by the term “Big Data”.

Big data is structured, unstructured and semi-structured or heterogeneous in nature. It becomes difficult for computing system to manage “Big Data” because of the immense speed and volume at which it is generated.

Traditional data management, warehousing, and analysis systems fizzle to analyze this type of data. Due to its complexity, big data is stored in distributed architecture file system.

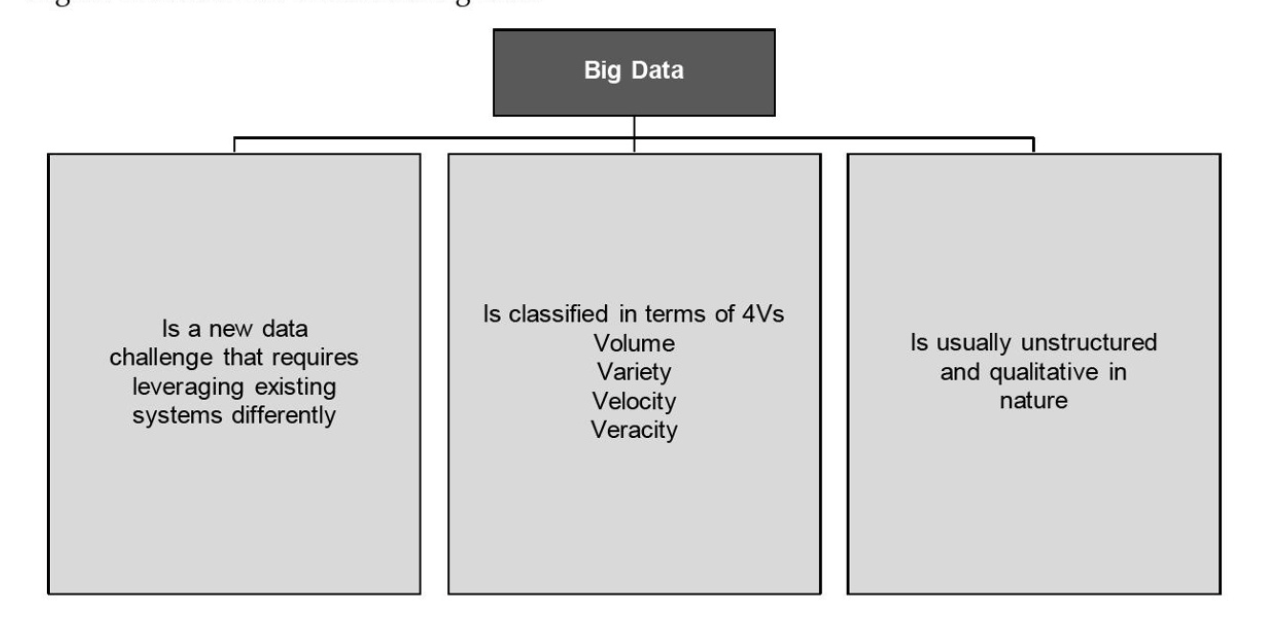
Hadoop by Apache is widely used for storing and managing Big Data. Analyzing Big Data is a challenging task as it involves large, distributed file systems, which should be fault tolerant, flexible, and scalable.

According to IBM, “Every day, we create 2.5 quintillion bytes of data – so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is Big Data.

Data is everywhere, in every industry, in the form of numbers, images, videos and text. As data continue to grow, so does the need to organize it. Collecting such huge amount of data would just be a waste of time, effort, and storage space if it cannot be put to any logical use. The need to sort, organize, analyze and offers this critical data in a systematic manner leads to the rise of the much discussed term, Big Data.

The process of capturing or collecting Big Data is known as “datafication”. Big Data is “datafied” so that it can be used productively. Big Data cannot be made useful by simply organizing it, rather the data’s usefulness lie in determining what we can do with it.

According to IBM, Big data is being generated by nearly everything around us at all times at an alarming velocity, volume and variety. To extract meaningful value from Big Data, you need optimal processing power, analytical capabilities, and skills.



**Real-World Examples of Big Data:**

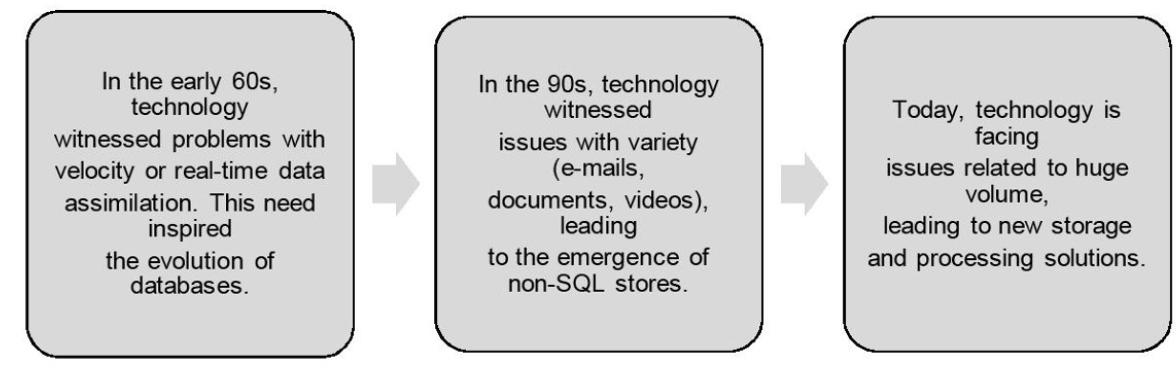
* Consumer product companies and retail organizations are observing data on social media website such as Facebook and Twitter. These sites help them to analyze customer behaviours, preferences, and product perception. Accordingly, the companies can line up their upcoming products to gain profits. This phenomenon is also known as social media analytics.
* Manufacturers are also monitoring minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to replace or maintain. Replacing it too soon wastes money and replacing it too late triggers an expensive work stoppage.
* Manufacturers are also monitoring social networks but with a different goal than marketers. They are using it to detect aftermarket support issue before a warranty failure becomes publicly detrimental.
* Financial service organization are using the data mined from customer interactions to slice and dice their users into finely tuned segments. This enables these financial institutions to create increasingly relevant and sophisticated offers.
* Advertising and marketing agencies are tracking social media to understand responsiveness to campaigns, promotions, and other advertising mediums.
* Insurance companies are using Big Data analysis to see which home insurance applications can be immediately processed and which one need a validating in-person visit from an agent.
* By embracing social media, retail organizations are engaging brand advocates, changing the perception of brand antagonists, and even enabling enthusiastic customers to sell their products.
* Hospitable are analyzing medical data and patient records to predict those patients that are likely to seek readmission within a few months of discharge. The hospital can then intervene in hopes of preventing another costly hospital stay. The hospitals also analyze patients data to prepare themselves to handle diseases.
* Web-based businesses are developing information products that combine data gathered from customers to offer more appealing recommendations and more successful coupon programs.
* The government is making data public at the national, state, and city level for users to develop new applications that can generate public good. For example weather data that is helpful for various industries.
* Sports team are using data for tracking ticket sales and even for tracking team strategies. This is known as sports analytics.

**Types and Source of Data**

|  |  |  |
| --- | --- | --- |
| **Type** | **Description** | **Source** |
| Social Data | Refers to the information collected from various social networking sites and online portals. | Facebook, Twitter and LinkedIn |
| Machine Data | Refers to the information generated from RFID chips, bar code scanners, and sensors. | RFID chip reading, Global positioning system(GPS) result |
| Transactional Data | Refers to the information generated from online shopping sites, retailers, and business to business (B2B) transactions. | Retail websites like eBay, and Amazon |

**History of Data Management – Evolution of Big Data**

Big Data is the new term of data evolution directed by the enormous velocity, variety and volume of data. Velocity implies the speed with which the data flows in an organization; Variety refers to the varied forms of data, such as structured, semi-structured or unstructured; and Volume defines the amount or quantity of data an organizations has to deal with.



The advent of IT, the Internet, and globalization has facilitated increased volume of data and information generation at an exponential rate, which has led to “information explosion”. This, in turn, fueled the evolution of Big Data that started in 1940s and continues till date.

Information explosion is described as a continuous increase in the volume of the published information or data and the effects of this abundant information.

|  |  |
| --- | --- |
| **Year** | Milestone |
| 1940s | An American librarian speculated the potential shortfall of shelves and cataloguing staff, realizing the rapid increase in information and limited storage. |
| 1960s | Automatic Data Compression was published in the communications of the ACM. It states that the explosion of information in the past few years makes it necessary that requirements for storing information should be minimized.  The paper described “Automatic Data Compression” as a complete automatic and fast three-part compression that can be used for any kind of information in order to reduce the slow external storage requirements and increase the rate of transmission from a computer system. |
| 1970s | In Japan, the Ministry of Posts and Telecommunications initiated a project to study information flow in order to track the volume of information circulating in the country. |
| 1980s | A research project was started by the Hungarian Central Statistics Office to account for the country’s information industry. It measured the volume of information in bits. |
| 1990s | Digital storage system become more economical than paper storage. Challenges related to the amount of data and the presence of obsolete data become apparent.  Some papers that discussed his concert are as follows:   * Michel Lesk published How much information is there in the world ? * John R. Masey presented a paper title Big Data…and the Next Wave of InfraStress. * K.G. Coffman and Andrew Odlyzko published The Size and Growth Rate of the Internet. * Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes published Visually Exploring Gigabyte Datasets in Real Time. |
| 2000 onwards | Many researcher and scientist published papers raising similar concerns and discussing way to solve them.  Various methods were introduced to streamline information.  Techniques for controlling the Volume, Velocity and Variety of data emerged, thus introducing 3D data management.  A study was carried out in order to estimate the new and original information created and stored worldwide in four types of physical media : paper, film, optical media, and magnetic media. |

**Structuring Big Data**

Structuring of data, in simple terms, is arranging the available data in a manner such that it becomes easy to study, analyze, and derive conclusion from it. But, why is structuring required?

In daily life, you may have come across questions like:

* How do I use to my advantage the vast amount of data and information I come across?
* Which news articles should I read of the thousands I come across?
* How to I chose a book of the millions available on my favourite sites or stores?
* How do I keep myself updated about new events, sports, inventions, and discoveries taking place across the globe?

Today, solutions to such question can be found by information processing systems. These systems can analyze and structure a large amount of data specially for you on the basis of what you searched, what you looked at, and for how long you remained at a particular page or website, thus scanning and presenting you with the customized information as per your behaviours and habits. In other words, structuring data helps in understanding user behaviours, requirements, and preferences to make personalized recommendations for every individual.

When a use regularly visit or purchase from online shopping sites, say eBay, each time he/she logs in, the system can present as recommended list of products that my interest the user on the basis of his/her earlier purchases or searches, thus presenting a specially customized recommendation set for every user. This is the power of Big Data analytics.

Today, various sources generated a variety of data, such as image, text, audios, etc. All such different types of data can be structured only if it is sorted and organized in some logical pattern. Thus, the process of structuring data requires one to first understand the various types of data available today.

**Types of Data**

Data that comes from multiple sources, such as database, Enterprise Resource Planning(ERP) system, weblogs, chat history, and GPS maps, varies in its format. However, different formats of data need to be made consistent and clear to be used for analysis. Data is obtained primarily from the following types of source:

* Internal sources, such as organizational or enterprise data.
* External sources, such as social data

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Definition** | **Examples of Sources** | **Application** |
| Internal | Provides structured or organized data that originates from within the enterprise and helps to run business. | * Customer Relationship Management(CRM) * Enterprise Resource Planning(ERP), system * Customer, details * Products and sales data * Generally OLTP and operation data | This data (current data in the operation system) is used to support daily business operation of an organization |
| External | Provides unstructured or unorganized data that originates from the external environment of an organization | * Business partners * Syndicate data suppliers * Internet * Government * Market Research organization | This data is often analyzed to understand the entities mostly external to the organization, such as customers, competitors, market , and environment |

On the basis of the data received from the sources , Big Data comprises:

* Structured data
* Unstructured data
* Semi-structured data

In a real-world scenario, typically the unstructured data is larger in volume than the structured and semi-structured data, approximately 70% to 80% of data in unstructured form.

Diagram

Description automatically generated

**Structured Data**

Structured data can be defined as the data that has a defined repeating pattern. This pattern makes it easier for any program to sort, read, and process the data. Processing structured data is much easier and faster than processing data without any specific repeating patterns.

* Is organized data in a predefined format
* Is stored in tabular form
* Is the data that resided in fixed fields within a record or file
* Is formatted data that has entities and their attributes mapped
* Is used to query and report against predetermined data types.

Some sources of structured data include:

* Relational databases (in the form of tables)
* Flat files in the form of records (like comma separated values(csv) and tab separated files)
* Multidimensional databases(majorly used in data warehouse technology)
* Legacy databases.

**Sample of Structured data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CustomerID** | **Name** | **ProductID** | **City** | **State** |
| 12365 | Smith | 241 | Graz | Styria |
| 23658 | Jack | 365 | Wolfsberg | Carinthia |
| 32456 | Kady | 421 | Enns | Upper Austria |

**Unstructured Data**

Unstructured data is a set of data that might or might not have any logical or repeating patterns. About 80% of enterprise data consist of unstructured content.

* Consists typically of metadata ie. The additional information related to data
* Comprises inconsistent data, such as data obtained from file, social media websites, satellites, etc.
* Consists of data in different formats such as e-mails, text, audio, video or images.

Some sources of unstructured data includes:

* Text both internal and external to an organization – Documents, logs, survey results, feedbacks, and e-mails from both within and across the organization.
* Social media – Data obtained from social networking platforms, including YouTube, Facebook, Twitter, LinkedIn and Flickr
* Mobile data – Data such as text messages and location information.

**Challenges Associated with Unstructured data**

Working with unstructured data poses certain challenges, which are as follows:

* Identifying the unstructured data that can be processed.
* Sorting, Organizing, and arranging unstructured data in different sets and formats
* Combining and linking unstructured data in a more structured format to derive any logical conclusions out of the available information
* Costing in terms of storage space and human resource (data analysis and scientists) needed to deal with the exponential growth of unstructured data.

**Semi-Structured Data**

Semi-structured data, also known as having a schema-less or self describing structure, refers to a form of structured data that contains tags or markup elements in order to separate elements and generate hierarchies of records and fields in the given data. Such type of data does not follow the proper structure of data models as in relational databases. In other words, data is stored inconsistently in rows and columns of database.

Some source for semi-structured data include:

* File systems such as Web data in the form of cookies.
* Data exchange formats such as JavaScript Object Notation (JSON) data

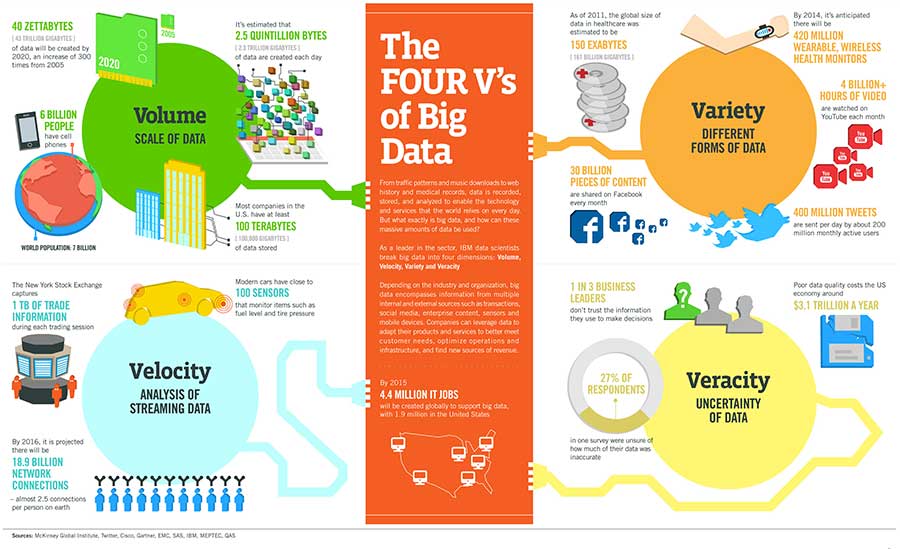
**Sample Semi-Structured Data**

|  |  |  |
| --- | --- | --- |
| **Sl.No** | **Name** | **Email** |
| **1** | **Sam Jocabs** | [**smj@zyz.com**](mailto:smj@zyz.com) |
| **2** | **First Name: David**  **Last Name: Brown** | [**davidb@xyz.com**](mailto:davidb@xyz.com) |
| **3** | **Name : Nick Sager** | Email-1:[nick.sager@xzy.com](mailto:nick.sager@xzy.com) Email-2:[nicksager@gmail.com](mailto:nicksager@gmail.com) |

**Elements of Big Data**

According to Gartner, data is growing at the rate of 59% every year. This growth can be depicted in terms of the following four Vs.

* Volume
* Velocity
* Variety
* Veracity



**Volume**

Volume is the amount of data generated by organizations or individuals. Today, the volume of data in most organizations is approaching exabytes. Some experts predict the volume of data to reach zettabytes in the coming year. Organizations are doing their best to handle this ever-increasing volume of data. For example, according to IBM, over 2.7 zetabytes of data is present in the digital universe today. Every minutes, over 571 new websites are being created. IDC estimates that by 2025, online business transaction will reach up to 850 billion per day.

The Internet alone generates a huge amount of data. The following figures help us to get an idea of the internet traffic:

* Internet has around 14.3 trillion live web pages, and 48 billion web pages are indexed by Google Inc; 14 billion Web pages are indexed by Microsoft Bing.
* Internet has around 672 exabytes of accessible data.
* Total world-wide internet traffic in the year 2013, was 43,639 petabytes.
* Over 9,00,000 servers are owned by Google Inc, which is the largest in the world.
* Total data stored on the Internet is over 1 yottabyte.

Even by underestimation, the total data stored on the Internet, including images, videos, audio, etc has crossed 1 yottabyte. The exact size of the internet will never be known.

**Velocity**

Velocity describes the rate at which data is generated, captured, and shared. Enterprises can capitalize on data only if it is captured and shared in real time. Information processing systems such as CRM and ERP face problems associated with data, which keeps adding up but cannot be processed quickly.

These systems are able to attend data in batches every few hours; however, even this time lag causes the data to lose its importance as new data is constantly being generated. For example eBay analyses around 5 million transactions per day in real time to detect and prevent frauds arising from the use of PayPal.

The sources of high velocity data include the following:

* IT devices, including routers, switches, firewalls, etc. constantly generate valuable data.
* Social media, including Facebook posts, tweets, and other social media activities, create huge amount of data, which is to be analyzed instantly at a fast speed because the value degrades quickly with time
* Portable device, including mobile PDA, etc, also generated data at a high speed.

**Variety**

We all know that data is being generated at a very fast pace. Now, this data is generated from different types of sources, such as internal , external, social and behavioural, and comes in different formats, such as image, text, videos, etc. Even a single source can generate data in varied formats, for example, GPS and social networking sites, such as Facebook, produce data of all types, including text, images, videos, etc.

Diagram, sunburst chart

Description automatically generated

**Veracity**

Veracity generally refers to the uncertainty of data, ie. Whether the obtained data is correct or consistent. Out of the huge amount of data that is generated in almost every process, only the data that is correct and consistent can be used for further analysis. Data when processed becomes information; however, a lot of effort goes in processing the data. Big Data especially in the unstructured and semi-structured forms, is messy in nature, and it takes good amount of time and expertise to clean the data and make it suitable for analysis.

**Advantages of Big Data Analytics**

According to Atul Buttle, Stanford, “Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world” So, the real power of Big Data lies in its analysis.

Processing, studying, and implementing the conclusions derived from the analysis of Big Data help you to collect accurate data, take timely and more informed strategic decisions, target the right set of audience and customers, increase benefits and reduce wastage of costs.

The right analysis of the available data can improve major business processes in various ways. For example, in a manufacturing unit, data analytics can improve the functioning of the following processes:

* **Procurement** – To find out which suppliers are more efficient and cost-effective in delivering products on time.
* **Product Development** – To draw insights on innovative product and service formats and designs for enhancing the development process and coming up with demanded products.
* **Manufacturing** – To identify which marketing campaigns will be the most effective in driving and engaging customers and understanding customer behaviour and channel behaviours.
* **Marketing** – To identify which marketing campaigns will be the most effective in driving and engaging customers and understanding customer behaviours and channel behaviours.
* **Price Managements** –To optimize price based on the analysis of external factors.
* **Merchandising** – To improve merchandise breakdown on the basis of current buying patterns ana increasing inventory levels and product interest insights on the basis of the analysis of various customer behaviours.
* **Sales -**  To optimize assignment of sales resources and accounts, product mix, and other operations.
* **Store Operations –** To adjust inventory levels on the basis of predicted buying patterns, study of demographics, weather, key events and other factors.
* **Human Resources –** To find out the characteristics and behaviours of successful and effective employees, as well as other employee insights for managing talent better.

Every business and industry today is affected by and benefitted from Big Data analytics in multiple ways.

**Future of Big Data**

In today competitive world, the need of Big Data is evident. If leaders and economies want exemplary growth and wish to generate value for all their stakeholders, Big Data has to be embraced and used extensively to :

* Allow the storage and use of transactional data in digital form
* Provide more specific information
* Refine analytics that can improve decision making
* Classify customers for providing customized products and services based on buying patterns

Most organizations today consider data and information to be their most valuable and differentiated asset. By analyzing this data effectively, organizations worldwide are now finding new ways to compete and emerge as leaders in their fields to improve decision making and enhance their productivity and performance. At the same time, the volume and variety of data is also increasing at the immense rate every day. The global phenomena of using Big Data to gain business value and competitive advantage will only continue to grow as will the opportunities associated with it.

Sample MCQ:

1. Which of the following is not a characteristics for Big Data?
   1. Volume
   2. **Variability**
   3. Variety
   4. Velocity
2. Who among the following do you think would be able to deal with the growing number of data sources efficiently?
   1. Business developer
   2. **Data Scientist**
   3. Sales executive
   4. Web designer
3. Which one of the following is not an example of external data sources?
   1. **Data from CRM**
   2. Data from Web logs
   3. Data from government sources
   4. Data from market surveys.
4. Which of the following does not belong to the traditional database technology?
   1. RDBMS
   2. DBMS
   3. Flat files
   4. **NoSQL**
5. If a Big Data analyst were to analyze data from a database of call logs provided by a telecom service provider, which element of Big Data would be dealing with?
   1. **Volume**
   2. Variable
   3. Variety
   4. Velocity
6. Some people call this data as “structured but not relations.” Which data are we talking about?
   1. Structured data
   2. Unstructured data
   3. **Semi-structured data**
   4. Mixed data
7. The data generated from a GPS satellite and web logs is classified as\_\_\_\_\_\_\_\_\_
   1. Structured data
   2. Unstructured data
   3. Both structured and unstructured data
   4. **Semi-structured data**
8. The data being captured can be in any form or structure. Which characteristics of Big Data are we talking about?
   1. Volume
   2. Velocity
   3. **Variety**
   4. Value