Big Data

What is Big Data

As the name suggests, Big Data is a massive collection of data that continues to increase exponentially over time. It is a data set that is so huge and complicated that cannot be handled, managed or processed by the traditional system. Big data is similar to regular data, except it is much larger in its volume.

Evolution of Big Data

The term Big data was coined by Roger Mougalas from O'Reilly media in the year 2005.
Earlier we were having Data warehouses and most of the companies were doing data related work with data warehouses. Later on when Hadoop map reduce came into existence in 2006 then, a lot of companies started to shift their data processing workload into hadoop map reduce because that was open source, fault tolerant on commodity hardware, was beyond SQL workloads and also it was fault tolerant, but it was not providing SQL and optimization that's why there was a need of something which could provide it. That's where Hive came into picture in the year 2007 which was providing the facility of SQL and optimization but it was not beyond SQL workloads. After this Kafka came into picture in the year 2008 which provided the facility of streaming ingestion, transactional API support and vast library of connectors to connect external systems which was not there in Hadoop Map Reduce and Hive.

Why to learn Big Data Technologies

Nowadays a large volume of data is getting generated all around the world by every single organisation. Whether it be texting to someone on social media or posting something, whether it be searching something on YouTube or shopping for the specific thing, each and every activity of ours is generating some kind of data. It is the application and the way big data is serving the human needs is making it so important and valuable. And there are also many future scope in the field of big data as a career perspective because in today's world everything is data.

Examples of Big Data

In today's world, everyone is aware about Google. Google is a company that famously uses big data. Apart from having access to user's information through its chrome browser and products offered by Gmail, Google is responsible for receiving billions or trillions of search requests every day through its search engine. Afterwards Google uses that data to train its algorithms and get better at fundamental search tasks such as parsing of sentences, correcting misspelt words and understanding what a user is trying to search for. Google also analyses data on previous and current search keywords to propose search phrases to users before they finish typing, providing helpful autocomplete features.

Let's take one more example to understand it better. All of us know that Uber is a ride-sharing service. Uber keeps track of its data in order to forecast demand spikes and changes in driver availability. This information enables the firm to determine suitable trip pricing and provide incentives to drivers, ensuring that the required number of cars are available to meet demand. Uber's anticipated time of arrival estimates are also based on data analysis, which goes a long way toward ensuring client happiness.

Who is using Big Data?

Nowadays, Big data is critical in every business, whether it be large or small, such as Healthcare, sports, insurance, e-commerce, IT, etc. Here are some of the examples:-

Amazon

Amazon is a well-known online retailer. They keep track of every piece of information on their clients in order to figure out how they spend their money on certain products. All of this data is being gathered in order to feed into social media advertising algorithms that may be used to improve customer interactions, propose items, and improve consumer experience and services, among other things.

Facebook

Facebook uses big data to enhance the user's experience such as you have noticed how Facebook reminds us of birthdays, friendship and anniversaries, but have you thought about how it does that? It uses big data to generate a short film that includes our old photos as a remembrance. Also it relies heavily on data to construct flashbacks depending on the information gathered.

Instagram

When we visit instagram then while scrolling we see the same things which we scrolled for or searched for in the previous search.This is the job of big data in instagram, which gathers our data and makes recommendations based on it. Also Big data aids in the monitoring of likes, followers, and new users, which aids in the development of instagram.

Why is Data so important?

We hear so much about data in today's world, but do we really understand the importance of data? Data is something that is effectively changing the world we live in and the way that we work. The more data you have, the more equipped you will be to make good choices or decisions and seize new chances. Good data will also provide you with the ability and proof you need to justify your judgments, allowing you to explain your thinking with confidence in the future.

Putting in place a sensible data gathering system will save you time in the long run. Going back and forth to acquire the same information wastes a lot of time and resources. A smart system will collect and present data in an easy-to-understand and navigate format, saving time for everyone in your business.

Characteristics of Big Data

Big Data is made up of a lot of data that isn't handled by typical data storage or processing units.Big Data is created on a massive scale, and it is being processed and analysed by many global corporations in order to unearth insights and enhance their businesses. Let's have a look at the characteristics of Big Data:

1. Variety
   The variety of Big Data can be either structured, unstructured or semi-structured. These data are gathered from multiple sources. In earlier days, the source of data collection was mainly spreadsheets and databases but in today's world the source of data collection has reached beyond our thoughts such as photos, videos, audios, emails, PDFs etc. Generally the data collected is categorised into:-
   a) Structured data
      Structured data is arranged in a tabular format i.e, in a spreadsheet. Relational database management systems are used to store such kinds of data.
   b) Unstructured data
      Unstructured data is information that does not have a pre-defined structure or a pre-defined data model. These are generally in the form of videos, audios, pictures etc.
   c) Semi-structured data
      Semi-structured data is a kind of data that is neither structured nor unstructured. It is not exactly structured but still has some structure in it. These are generally in the form

of JSON i.e, Javascript Object Notation having key-value pairs.

2. Velocity

As you know velocity generally means speed. So, here velocity refers to the speed or rate at which data is being created or generated.

3. Veracity

The term "veracity" refers to how trustworthy the data is. It allows you to filter or transform data in a variety of ways. The ability to handle and manage data effectively is referred to as veracity. Big Data is also critical for corporate growth.

4. Volume

The term volume here indicates unimaginable amounts of data or information generated every second. On a daily basis, massive 'volumes' of data are created from many sources such as social media platforms, corporate processes, machines, networks, human interactions, and so on.

5. Value

The term value represents here, how valuable and reliable our data is. A large volume of important, dependable, and trustworthy data that must be saved, processed, and evaluated in order to uncover insights

Challenges of Big Data

The challenges in Big Data refers to the obstacles in real-world implementations. Before attempting to deploy or utilise big data, an organisation must have a clear business justification that is tied to the organisation's strategy (as with any change). This will secure buy-in from top management and a clear focus on what has to be done. It's also a good idea to do some type of cost/benefit analysis to see whether the advantages exceed the expenses, stress, and obstacles of implementation.

Integration:

Data in a business originates from a number of places, including employee-created reports, social networking sites, ERP programmes, customer logs, financial reports, e-mails, and presentations. Putting all of this information together to create reports is a difficult undertaking.

Security:

Security is another challenge with Big Data. Due to the vast volume of data generated, most firms are unable to maintain frequent inspections. However, real-time security inspections and surveillance should be required since it is the most useful.

Tools Selection:

Also while selecting the tools , it creates a confusion for an organisation to choose the best tools for data analysis and storage as well.

Exponential increase in volume:

Big Data in itself is huge and it becomes even huge when generated exponentially and that's where it becomes difficult to handle them which grow as a challenge for the organisations.

Common Problems of Big Data

While dealing with Big Data, There are several problems we deal with:

1. Size of Files
   One of the most aggravating aspects of working with massive data is the file size. We often work with tens of gigabytes of data. In fact, if you look at the screenshot at the top of this page, you'll see that it's a 90GB file! Try opening a 90-gigabyte file in Access or Excel; your machine will almost certainly freeze or crash.

2. Missing Data
   Our lives are made a living hell by missing data. However, various data sources might have variable degrees of completeness. You can't uncover patterns in data if it doesn't exist in the first place, but

by doing some detective work and thinking outside the box, you can reduce the quantity of missing data.

3. Messy Data

   The longer it takes to process a bigger file, the longer it will take to process it.

   A file containing millions of entries and hundreds of fields will always be large, even before it's combined with additional data - that's simply the nature of big data. However, a lot of massive data suffers from bloat, which means that files are larger than they need to be.

Management of Big Data

The organisation, administration, and control of vast amounts of both organised and unstructured data is known as big data management. The purpose of big data management is to guarantee that business intelligence and big data analytics programmes have access to high-quality data. Also managing large volumes of structured and unstructured data is critical to the success of any company that has to guarantee that their data is of good quality and acceptable for analytics and business intelligence applications.

Storage of Big Data

When it comes to Huge Data, it's not only the big companies that have to deal with it; even tiny firms acquire a lot of data from various sources such as emails, social media interactions, sales, and a number of other things. Data must be kept someplace before it can be sorted and processed for analysis, regardless of the size of the firm or the sector in which it operates.

Currently, there are two well-known data storage methods:

1. Warehouse Storage: The capacity to convert raw data into knowledge and insight is the most significant advantage of data warehouses. Data warehouses are a powerful tool for

supporting queries, analytics, reporting, and forecasting and trending based on gathered data.

2. Cloud Storage: The cloud not only offers easily accessible infrastructure, but also the capacity to swiftly scale that infrastructure to handle big surges in traffic or demand. Accessibility and usefulness are also provided by the cloud

Processing of Big Data

Big data processing is a collection of methodologies or programming models for accessing massive amounts of data and extracting meaningful information for decision-making. Here are some tools:-

a) One of the most common programming methods for massive data processing on large-scale commodity clusters is MapReduce

b) Apache Pig is a structured query language (SQL)-like environment created by Yahoo and utilised by a variety of companies including Yahoo, Twitter, AOL, LinkedIn, and others.

c) Facebook has created another MapReduce wrapper, Hive. These two wrappers improve the code development environment and make it easier.

d) Hadoop is an open-source MapReduce solution that is frequently used for large-scale data processing. Some Cloud providers, such as Amazon EMR, provide this software to establish Hadoop clusters to handle massive data utilising Amazon EC2 resources.

OLAP AND OLTP

OLAP

OLAP stands for Online analytical processing. The term "online analytical processing" refers to a set of software tools that are used to analyse data in order to make business choices. Basically OLAP helps in decision support. In OLAP, the present data cannot be modified or changed. OLAP is mainly used in Netflix, spotify etc.

OLTP

OLTP stands for Online transaction processing. Basically OLTP is used in our day to day operations or day to day transactions of an organisation. In OLTP, the present data can be modified or changed. OLTP is mainly used for online banking, online air ticket booking etc

Operational vs Analytical Big Data

Operational data is exactly what it sounds like: data generated by the day-to-day operations of your company. Customer, inventory, and purchase data are examples of this kind of information. This kind of data is rather easy and will appear in most businesses in a similar manner.