Big Data

Evolution of Big Data

Before exploring what is Big Data, let me begin by giving some insight into why the term Big Data has gained so much importance.

When was the last time you guys remember using a floppy or a CD to store your data? Let me guess, had to go way back in the early 21st century right? The use of manual paper records, files, floppy and discs have now become obsolete. The reason for this is the exponential growth of data. People began storing their data in relational database systems but with the hunger for new inventions, technologies, applications with quick response time and with the introduction of the internet, even that is insufficient now. This generation of continuous and massive data can be referred to as Big Data. There are a few other factors that characterize Big Data which I will be explaining later in this blog.

Forbes reports that there are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating. Internet of Things(IoT) is one such technology which plays a major role in this acceleration. 90% of all data today was generated in the last two years.

Big Data Definition

So before I explain what is Big Data, let me also tell you what it is not! The most common myth associated with Big Data is that it is just about the size or volume of data. But actually, it's not just about the "big" amounts of data being collected. *Big Data* refers to the large amounts of data which is pouring in from various data sources and has different formats. Even previously there was huge data which were being stored in databases, but because of the varied nature of this Data, the traditional relational database systems are incapable of handling this Data. Big Data is much more than a collection of datasets with different formats, it is an important asset which can be used to obtain enumerable benefits.

Types of Big Data

Big Data could be of three types:

- Structured
- Semi-Structured
- Unstructured

1. *Structured*

   The data that can be stored and processed in a fixed format is called as Structured Data. Data stored in a relational database management system (RDBMS) is one example of 'structured' data. It is easy to process structured data as it has a fixed schema. Structured Query Language (SQL) is often used to manage such kind of Data.

2. *Semi-Structured*

   Semi-Structured Data is a type of data which does not have a formal structure of a data model, i.e. a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze. XML files or JSON documents are examples of semi-structured data.

3. *Unstructured*

   The data which have unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data. Text Files and multimedia contents like images, audios, videos are example of unstructured data. The unstructured data is growing quicker than others, experts say that 80 percent of the data in an organization are unstructured.

Big Data Characteristics

The five characteristics that define Big Data are: Volume, Velocity, Variety, Veracity and Value.

1. *VOLUME*

   Volume refers to the 'amount of data', which is growing day by day at a very fast pace. The size of data generated by humans, machines and their interactions on social media itself is massive. Researchers have predicted that 40 Zettabytes (40,000 Exabytes) will be generated by 2020, which is an increase of 300 times from 2005.

2. *VELOCITY*

   Velocity is defined as the pace at which different sources generate the data every day. This flow of data is massive and continuous. There are 1.03 billion Daily Active Users (Facebook DAU) on Mobile as of now, which is an increase of 22% year-over-year. This shows how fast the number of users are growing on social media and how fast the data is getting generated daily. If you are able to handle the velocity, you will be able to generate insights and take decisions based on real-time data.

3. *VARIETY*

   As there are many sources which are contributing to Big Data, the type of data they are generating is different. It can be structured, semi-structured or unstructured. Hence, there is a variety of data which is getting generated every day. Earlier, we used to get the data from excel and databases, now the data are coming in the form of images, audios, videos, sensor data etc. as shown in below image. Hence, this variety of unstructured data creates problems in capturing, storage, mining and analyzing the data.

4. *VERACITY*

   Veracity refers to the data in doubt or uncertainty of data available due to data inconsistency and incompleteness. In the image below, you can see that few values are missing in the table. Also, a few values are hard to accept, for example – 15000 minimum value in the 3rd row, it is not possible. This inconsistency and incompleteness is Veracity.
   Data available can sometimes get messy and maybe difficult to trust. With many forms of big data, quality and accuracy are difficult to control like Twitter posts with hashtags, abbreviations, typos and colloquial speech. The volume is often the reason behind for the lack of quality and accuracy in the data.

   - Due to uncertainty of data, 1 in 3 business leaders don't trust the information they use to make decisions.
   - It was found in a survey that 27% of respondents were unsure of how much of their data was inaccurate.

- Poor data quality costs the US economy around $3.1 trillion a year.

5. *VALUE*

   After discussing Volume, Velocity, Variety and Veracity, there is another V that should be taken into account when looking at Big Data i.e. Value. It is all well and good to have access to big data but unless we can turn it into value it is useless. By turning it into value I mean, Is it adding to the benefits of the organizations who are analyzing big data? Is the organization working on Big Data achieving high ROI (Return On Investment)? Unless, it adds to their profits by working on Big Data, it is useless.

Big Data Analytics

Now that I have told you what is Big Data and how it's being generated exponentially, let me present to you a very interesting example of how *Starbucks*, one of the leading coffeehouse chain is making use of this Big Data.

I came across this article by Forbes which reported how *Starbucks* made use of Big Data to analyse the preferences of their customers to enhance and personalize their experience. They analysed their member's coffee buying habits along with their preferred drinks to what time of day they are usually ordering. So, even when people visit a "new" Starbucks location, that store's point-of-sale system is able to identify the customer through their smartphone and give the barista their preferred order. In addition, based on ordering preferences, their app will suggest new products that the customers might be interested in trying. This my friends is what we call Big Data Analytics.

Basically, Big Data Analytics is largely used by companies to facilitate their growth and development. This majorly involves applying various data mining algorithms on the given set of data, which will then aid them in better decision making.

There are multiple tools for processing Big Data such as *Hadoop*, *Pig*, *Hive*, *Cassandra*, *Spark*, *Kafka*, etc. depending upon the requirement of the organisation.

Big Data Applications

These are some of the following domains where *Big Data Applications* has been revolutionized:

- Entertainment: Netflix and Amazon use Big Data to make shows and movie recommendations to their users.
- Insurance: Uses Big data to predict illness, accidents and price their products accordingly.
- Driver-less Cars: Google's driver-less cars collect about one gigabyte of data per second. These experiments require more and more data for their successful execution.
- Education: Opting for big data powered technology as a learning tool instead of traditional lecture methods, which enhanced the learning of students as well aided the teacher to track their performance better.
- Automobile: Rolls Royce has embraced Big Data by fitting hundreds of sensors into its engines and propulsion systems, which record every tiny detail about their operation. The changes in data in real-time are reported to engineers who will decide the best course of action such as scheduling maintenance or dispatching engineering teams should the problem require it.
- Government: A very interesting use of Big Data is in the field of politics to analyse patterns and influence election results. Cambridge Analytica Ltd. is one such organisation which completely drives on data to change audience behaviour and plays a major role in the electoral process.

Scope of Big Data

- Numerous Job opportunities: The career opportunities pertaining to the field of Big data include, Big Data Analyst, Big Data Engineer, Big Data solution architect etc. According to IBM, 59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.
- Rising demand for Analytics Professional: An article by Forbes reveals that "IBM predicts demand for Data Scientists will soar by 28%". By 2020, the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000 according to IBM.
- Salary Aspects: Forbes reported that employers are willing to pay a premium of $8,736 above median bachelor's and graduate-level salaries, with successful applicants earning a starting salary of $80,265
- Adoption of Big Data analytics: Immense growth in the usage of big data analysis across the world.

Examples of Big Data

Daily we upload millions of bytes of data. 90 % of the world's data has been created in last two years.

- Walmart handles more than 1 million customer transactions every hour.
- Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.
- 230+ millions of tweets are created every day.
- More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide.
- YouTube users upload 48 hours of new video every minute of the day.
- Amazon handles 15 million customer click stream user data per day to recommend products.
- 294 billion emails are sent every day. Services analyses this data to find the spams.
- Modern cars have close to 100 sensors which monitors fuel level, tire pressure etc. , each vehicle generates a lot of sensor data.

Challenges with Big Data

Let me tell you few challenges which come along with Big Data:

1. Data Quality – The problem here is the 4th V i.e. Veracity. The data here is very messy, inconsistent and incomplete. Dirty data cost $600 billion to the companies every year in the United States.

2. Discovery – Finding insights on Big Data is like finding a needle in a haystack. Analyzing petabytes of data using extremely powerful algorithms to find patterns and insights are very difficult.

3. Storage – The more data an organization has, the more complex the problems of managing it can become. The question that arises here is "Where to store it?". We need a storage system which can easily scale up or down on-demand.

4. Analytics – In the case of Big Data, most of the time we are unaware of the kind of data we are dealing with, so analyzing that data is even more difficult.

5. Security – Since the data is huge in size, keeping it secure is another challenge. It includes user authentication, restricting access based on a user, recording data access histories, proper use of data encryption etc.

6. Lack of Talent – There are a lot of Big Data projects in major organizations, but a sophisticated team of developers, data scientists and analysts who also have sufficient amount of domain knowledge is still a challenge.