Source of material: https://www.edureka.co/blog/big-data-tutorial

ig Data Tutorial: All You Need To Know About Big Data!

Last updated on May 22,2019 *74.1K Views*

Awanish
Awanish is a Sr. Research Analyst at Edureka. He has rich expertise...

ig Data Tutorial

Big Data, haven't you heard this term before? I am sure you have. In the last 4 to 5 years, everyone is talking about Big Data. But d you really know what exactly is this Big Data, how is it making an impact on our lives & why organizations are hunting for pr fessionals with ***Big Data skills***? In this Big Data Tutorial, I will give you a complete insight about Big Data.

Bel w are the topics which I will cover in this Big Data Tutorial:

- Story of Big Data
- Big Data Driving Factors
- What is Big Data?
- Big Data Characteristics
- Types of Big Data
- Examples of Big Data
- Applications of Big Data
- Challenges with Big Data

Let me start this Big Data Tutorial with a short story.

Story of Big Data

In ancient days, people used to travel from one village to another village on a horse driven cart, but as the time passed, villages became towns and people spread out. The distance to travel from one town to the other town also increased. So, it became a problem to travel between towns, along with the luggage. Out of the blue, one smart fella suggested, we should groom and feed a horse more, to solve this problem. When I look at this solution, it is not that bad, but do you think a horse can become an elephant? I don't think so. Another smart guy said, instead of 1 horse pulling the cart, let us have 4 horses to pull the same cart. What do you guys think of this solution? I think it is a fantastic solution. Now, people can travel large distances in less time and even carry more luggage.

The same concept applies on Big Data. Big Data says, till today, we were okay with storing the data into our servers because the volume of the data was pretty limited, and the amount of time to process this data was also okay. But now in this current technological world, the data is growing too fast and people are relying on the data a lot of times. Also the speed at which the

The quantity of data on planet earth is growing exponentially for many reasons. Various sources and our day to day activities generates lots of data. With the invent of the web, the whole world has gone online, every single thing we do leaves a digital trace. With the smart objects going online, the data growth rate has increased rapidly. The major sources of Big Data are social media sites, sensor networks, digital images/videos, cell phones, purchase transaction records, web logs, medical records, archives, military surveillance, eCommerce, complex scientific research and so on. All these information amounts to around some Quintillion bytes of data. By 2020, the data volumes will be around 40 Zettabytes which is equivalent to adding every single grain of sand on the planet multiplied by seventy-five.

What is Big Data?

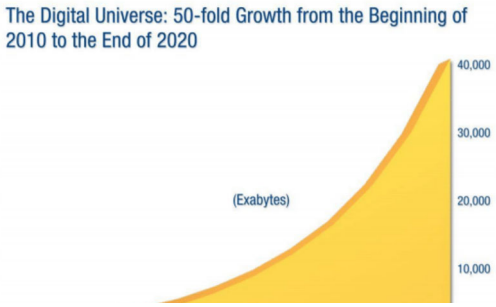
Big Data is a term used for a collection of data sets that are large and complex, which is difficult to store and process using available database management tools or traditional data processing applications. The challenge includes capturing, curating, storing, searching, sharing, transferring, analyzing and visualization of this data.

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.



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.

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

   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.

Go through our Big Data video below to know more about Big Data:

Big Data Tutorial For Beginners | What Is Big Data | Edureka

Big Data Tutorial For Beginners | What Is Big Data | Big Data Tutorial | Hac

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.

Till now, I have just covered the introduction of Big Data. Furthermore, this Big Data tutorial talks about examples, applications and challenges in Big Data.

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

- **Smarter Healthcare**: Making use of the petabytes of patient's data, the organization can extract meaningful information and then build applications that can predict the patient's deteriorating condition in advance.
- **Telecom**: Telecom sectors collects information, analyzes it and provide solutions to different problems. By using Big Data applications, telecom companies have been able to significantly reduce data packet loss, which occurs when networks are overloaded, and thus, providing a seamless connection to their customers.
- **Retail**: Retail has some of the tightest margins, and is one of the greatest beneficiaries of big data. The beauty of using big data in retail is to understand consumer behavior. Amazon's recommendation engine provides suggestion based on the browsing history of the consumer.
- **Traffic control**: Traffic congestion is a major challenge for many cities globally. Effective use of data and sensors will be key to managing traffic better as cities become increasingly densely populated.
- **Manufacturing**: Analyzing big data in the manufacturing industry can reduce component defects, improve product quality, increase efficiency, and save time and money.
- **Search Quality**: Every time we are extracting information from google, we are simultaneously generating data for it. Google stores this data and uses it to improve its search quality.

Someone has rightly said: ***"Not everything in the garden is Rosy!"***. Till now in this Big Data tutorial, I have just shown you the rosy picture of Big Data. But if it was so easy to leverage Big data, don't you think all the organizations would invest in it? Let me tell you upfront, that is not the case. There are several challenges which come along when you are working with Big Data.

Now that you are familiar with Big Data and its various features, the next section of this blog on Big Data Tutorial will shed some light on some of the major challenges faced by Big Data.

Challenges with Big Data

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.

Hadoop to the Rescue

We have a savior to deal with Big Data challenges – its ***Hadoop***. Hadoop is an open source, Java-based programming framework that supports the storage and processing of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

Hadoop with its distributed processing, handles large volumes of structured and unstructured data more efficiently than the traditional enterprise data warehouse. Hadoop makes it possible to run applications on systems with thousands of commodity hardware nodes, and to handle thousands of terabytes of data. Organizations are adopting Hadoop because it is an open source software and can run on commodity hardware (your personal computer). The initial cost savings are dramatic as commodity hardware is very cheap. As the organizational data increases, you need to add more & more commodity hardware on the fly to store it and hence, Hadoop proves to be economical. Additionally, Hadoop has a robust Apache community behind it that continues to contribute to its advancement.

As promised earlier, through this blog on Big Data Tutorial, I have given you the maximum insights in Big Data. This is the end of Big Data Tutorial. Now, the next step forward is to know and learn Hadoop. We have a *series of Hadoop tutorial* blogs which will give in detail knowledge of the complete Hadoop ecosystem.

All the best, Happy Hadooping!

*Now that you have understood what is Big Data, check out the **Big Data training** by Edureka, a trusted online learning company with a network of more than 250,000 satisfied learners spread across the globe. The Edureka Big Data Hadoop Certification Training course helps learners become expert in HDFS, Yarn, MapReduce, Pig, Hive, HBase, Oozie, Flume and Sqoop using real-time use cases on Retail, Social Media, Aviation, Tourism, Finance domain.*

Got a question for us? Please mention it in the comments section and we will get back to you.