

A Review Paper on Big Data & Analytics

Shivashish Upadhyay

Department of Computer Science and Engineering, ABES Institute of Technology, Ghaziabad, India

Abstract – Technology has been changing very fast which is leading to explosion of data that can be termed as Big Data and this data is coming from various sources and therefore data is growing exponentially. Variety of enormous data is generated at an extremely fast speed in various sectors therefore analyzing big data has been extremely crucial and inevitable as a result big data analytics is being adopted all throughout the globe in order to gain numerous benefits from the data being produced. The main purpose of this paper is to discuss the evolution of big data, characteristics of big data, need of big data analytics, basic tools and technologies used in big data analytics, advantages and disadvantages of big data analytics and domains currently making use of big data analytics.

Key Words: Evolution of Big Data, Big Data, Characteristics of Big Data, Big Data Analytics, Applications

1. INTRODUCTION

Advancement of computer science and technology is taking day by day. As a result of which approximately 2.5 quintillion of data is created worldwide every day. This data comes from various sources whether it is from social media, banking sector and from various other institutions. This data is not in the same form as this is coming from various sources. A lot of data is generated from smartphones, smart watches, and various other smart products. The deal is this data is not in the format that primitive relational database can handle and apart from that the volume of data has increased exponentially. This enormous collection of data is Big Data. The Word "Big Data" is utilized by sociologist Mr. Charles Tilly in his article. later, "Big Data" is utilized by CNN in year of 2001 in news story.

2. Evolution of Big Data

There are various factors contributing to the growth of data. Some of those factors are listed below:

- Evolution of Technology leading to evolution of Big data
- Huge amount of data from IoT leading to evolution of Big Data
- Huge amount of data on social media leading to evolution of Big Data

2.1 Evolution of Technology leading to evolution of Big Data

Figure given below depicts how technology have been evolved. Earlier we have landline telephones but now we use android and iOS smartphones that are making our lives smarter. But there is one more truth that these smartphones are generating a large amount of data for every action even one video sends through messenger generates data. This is only an example we have no idea how much data we have been generating through these smartphones. Now the deal is this data is not structured to it cannot be handled with the help of traditional database management systems, so we make use of Big Data analytics to analyze useful information from this huge amount of data.

Now looking at next listed change, earlier we were using bulky desktops for processing very less amount of data usually in KBs and MBs. Initially floppies were used for storing very less data in KBs but then came hard disks storing TBs of data and now we are using cloud computing. Traditional computing is now replaced by cloud computing. In today's advanced word of computing a large amount of data is generated and stored over high speed disks and over high speed servers using cloud technology.

Similarly, now a days self-driving car have come out. These self-driving cars make use of various sensors for identifying every minute detail like size of the obstacle, distance of the obstacle and then it decides how to react. Therefore, these Self-driving cars also generate a large amount of data.

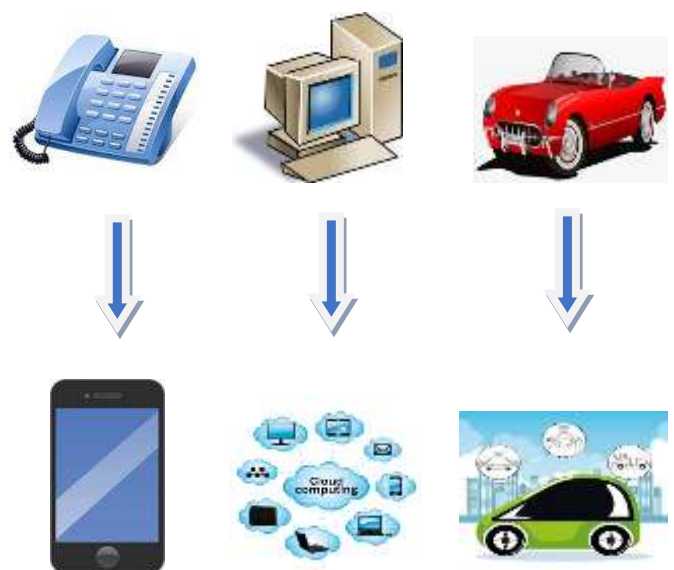


Figure 1: Evolution of Technology

2.2 Huge amount of data from IoT leading to evolution of Big Data

IoT stands for Internet of Things. As its name indicates IoT connects a physical device with internet and makes a device smarter. Now a days we have smart TVs, ACs, Clocks, Cars even smart houses. Taking example of smart ACs. Smart Air conditioners monitors the body temperature and outside temperature and accordingly decides what should be the temperature of the room so in order to do this it has to first accumulate data from various sources using sensors and using information available on internet. Now we can see that how these smart ACs generating huge amount of data. In 2020 there are more than 50 billion IoT devices. This number is expected to increase up to 125 million till 2030. So it can be clearly seen how these IoT devices generating large amount of data.



Figure 2: Internet of Things

2.3 Huge amount of data on social media leading to evolution of Big Data

Social media is one of the most important factors in the evolution of Big Data. Now a days everyone is using Facebook, Instagram, YouTube, Snapchat and lot of other social media websites. So, these social media sites have lots of data for example these have your personal details like your name, age and apart from that each picture that you like or react to also generates data. Even the Facebook pages that you go around liking also generates data. Now we can see that most people are sharing videos on Facebook so that is also generating huge amount of data. And the most challenging part here is the data generated is not structured. This presence of large volume of variety of data is one of the most important factors in evolution of Big data.



Figure 3: Data generated per minute on Social media

2.4 Some other factors responsible for evolution of Big Data

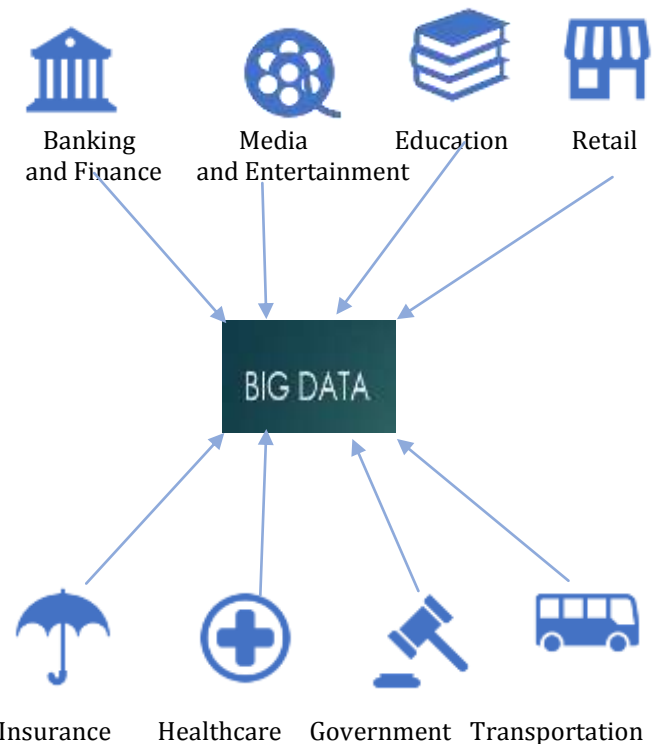


Figure 4: Some more Factors responsible for evolution of Big Data

Figure given above lists some of the more factors that are responsible for evolution of Big Data.

Talking about retail sector. As we know all of us visit online shopping sites and we search a lot items on those sites so all these search history of customers is stored along with the personal detail and this data is used for customizing the ads and suggesting suitable products to the customers. And there are numerous ways that we might don't know that we are

generating data on such sites. These sites were not available earlier so that time there was no way that such huge amount of data was generated. So online retailing sites are generating Big Data. Similarly, the data have evolved due to other reasons also like banking and finance, media and entertainment, education, insurance, healthcare, government, transportation, education etc.







2. Definition of Big Data

Big Data is the term used for referring huge amount of structured and unstructured data which is so large that it is difficult to process that data using conventional primitive data base management tools. In most cases the volume of data is too big and it moves too fast and exceeds current processing capacity.

Big data is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database system tools or traditional data processing applications.

3. Characteristics of Big Data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: Volume, Variety and Velocity. Over time, other Vs have been added to description of big data.

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					

3.1. Volume of Data

The amount of data that is created is significant in this specific circumstance. It is the size of the data which decides the worth and capability of the data viable and whether it can really be viewed as Big Data or not. The name 'Enormous Data' itself contains a term which is identified with size and subsequently the trademark.

3.2. Variety

Big Data is collection of different kinds of data from various sources. The variety here indicates about the different forms

of data structured, unstructured, semi-structured. This implies that the category to which big data belongs to is also most important fact that needs to be known by data scientists. It helps the data scientists to take more advantage from big data for their use.

3.3. Velocity

Velocity here refers to the speed of generation of data and how fast it has to be analyzed to meet the requirements. Increasing velocity of data can not be handled using traditional systems.

3.4. Veracity

Data is captured from various sources so there may be variation in quality of data. This data has a lot of inconsistency and missing values. The accuracy and efficiency in results of Big Data Analytics is dependent on veracity of data.

3.5. Value

It is mechanism to bring the correct meaning out of the data. Firstly, we need to mine only useful data from largely available data sets after that we performs certain analytics on that data that we have cleaned. The result of analysis should be of some value for us that is it can basically find out certain insights that were not possible earlier. We need to take care that whatever data have been generated it makes sense and it helps our business to grow and it has some value to it.

4. Need for Big Data Analytics

- Making smarter and more efficient organizations
- Optimizing Business operations by analyzing customer behavior
- Cost Reduction
- Next generation Products

5. Big Data Analytics

Big data analytics examine large and different types of data to uncover hidden patterns, correlations and other insights. Basically, Big Data analytics is helping large companies to facilitate growth and development. So, this majorly involve applying various data mining algorithms on a given set of data which will then aid these organizations in making better decisions.

6. Stages in Big Data Analytics

- Identifying Problem
- Designing Data Requirement
- Pre-processing Data

- Performing Analytics over Data
- Visualizing Data

7. Types of Big Data analytics

- **Descriptive Analysis:** It answers the question-"What is happening now?". It uses data aggregation and data mining techniques to provide insight into the past (can be 1min ago or few years past) and then it answers what is happening now based on the incoming data (E.g. Google Analysis tool).
- **Predictive Analysis:** It answers the question-"What is going to happen?". It uses statistical models and forecast techniques to understand the future and answer what could happen (E.g. South West airlines with sensors identifies patterns that indicate potential malfunctions and immediate repairing is done).
- **Prescriptive Analysis:** It answers the question-"What action should be taken?" (E.g. Google's Self Driving Cars).
- **Diagnostic Analysis:** It answers the question-"Why did it happened?". It is helpful in determining what kind of factors and events contributed towards a particular outcome (E.g. Time series data of sales of a company can be analyzed using this technique to know why sales decreased).

7. Tools used in Big Data Analytics

- **Hadoop:** Hadoop is a open source framework that allows to store data in a distributed fashion so that it can be processed parallelly.
- **Apache Pig:** Apache Pig is a platform that is majorly use for analyzing large data sets and then represent these data sets as data flows. Basically, Pig is used for scripting and the language is Pig Latin.
- **Kafka:** Kafka is a messaging system.
- **Apache Hive:** It is data warehousing tools and it allows us to perform big data analytics using Hive language which is similar to SQL.
- **Splunk:** Splunk is a log analysis tools. Logs contains information about every single transaction.
- **Talend:** It is an open source software integration platform which helps to analyze and turn the data into business insights.
- **Spark:** It is an in memory data processing engine that allows us to efficiently execute freeman, machine learning and SQL workloads and it require fast iterative access to data sets.
- **Apache Hbase:** It is NOSQL database that allows us to store unstructured and semi-structured data with ease and provide read and write access.

8. Applications of Big Data Analytics

- **Healthcare:** In healthcare big data analytics is used to reduce cost, predict epidemics, avoid preventable diseases and then improve the quality of life in general. The most widespread application of big data in healthcare is electronic health record for storing patient's records
- **Telecom:** Telecom industry is one of the most significant contributors to big data. So, telecom industry basically analyzes all our call data records in real time and then they identify fraudulent behavior and acts on them immediately. Now the marketing division of telecom industry modifies their campaigns to better target their customers and then use these insights for developing new products.
- **Insurance:** Insurance companies use big data analytics for risk assessment, marketing, customer insights, customer experience and much more.
- **Government:** Government across the world are also adopting big data analytics. For example, The Indian Government had used big data analytics to get an estimate of trade in the country.
- **Finance:** Banks and financial firms use analytics to differentiate fraudulent interactions from legitimate interactions.
- **Automobile:** Many automobile companies are using big data analytics. For example Rolls Royce. So Rolls Royce embraced big data by fitting hundreds of sensors into its engines and propulsion system and these sensors basically record every tiny change in the engine and propulsion system so the changes in the data in real time are reported to the engineers who will then decide the best course of action such as scheduling, maintenance or dispatching the engineering teams if problems arises.
- **Retail:** Online E commerce sites are widely using big data analytics for analyzing customer behavior and then suggesting the customers preferable goods.

9. CONCLUSION

Technology is changing day by day. Increase in population and advancement in technologies is leading to evolution of more and more data. Today's smart devices are generating a huge amount of data. This huge amount of data can be analyzed by using different analytics tools. And the result of analysis can be used for taking advantage in various business operations. Many domains are making use of big data analytics for providing better services and thus earning more profit. So, it can be concluded that as of now big data is very useful for all type of professionals. Businessman can use big data analytics for performing better in their business and

this field has great scope for computer science professionals as well for performing various data analytics operations

ACKNOWLEDGEMENT

I would also like to express my gratitude towards Department of Computer Science and Engineering ABES Institute of Technology, Ghaziabad for their kind co-operation and encouragement which help me in completion of this work. It helped me a lot to realize of what we study for.

Secondly, I would like to thank my parents who patiently helped me as i went through my work.

Edureka! Big Data lectures were extremely supporting to me in acquiring the required knowledge related to the topic.

Last but clearly not the least, I would thank The Almighty for giving me strength to complete my work on time.

REFERENCES

- [1] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.
- [2] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.
- [3] "Community cleverness required". Nature 455 (7209): 1. 4 September 2008. doi:10.1038/455001a.
- [4] "Sandia sees data management challenges spiral". HPC Projects. 4 August 2009.
- [5] META Group. "3D Data Management: Controlling Data Volume, Velocity, and Variety." February 2001.Xiaolong Jin
- [6] Big Data Seminar report available at <https://www.seminarsttopics.com/seminar/8722/big-data-seminar-report-pdf>
- [7] Big Data Analytics | Big Data Explained | Big Data Tools & Trends | Big Data Training | Edureka available at <https://www.youtube.com/watch?v=k7zu3NXEiGY&t=1725s>
- [8] Data Tutorial For Beginners | What Is Big Data | Big Data Tutorial | Hadoop Training Big | Edureka available at <https://www.youtube.com/watch?v=zez2Tv-bcXY&t=801s>

BIOGRAPHY



Shivashish Upadhyay is an undergraduate computer science and engineering student pursuing B. Tech at ABES Institute of Technology, Ghaziabad. His area of interest is Big Data, Machine Learning, AI. He has done training in python & machine learning