INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 1)
Available online at: www.ijariit.com

A study on Big Data analytics, approach and applications

Anurag Singh Bisht
anuragbisht04@gmail.com
Netaji Subhas University of Technology, New Delhi, Delhi

Swadhin Kumar Nayak
swadhinkumarnayak.1993@gmail.com
Netaji Subhas University of Technology, New Delhi, Delhi

ABSTRACT

With the organizations going online, the consumers now have the ease to utilize the online resources for their convenience. However, this ever-increasing online consumer population has led to a huge growth in the amount of information that the organizations need to manage. How to utilize that data is another big question that organizations are often faced with. This is where Big Data comes into the picture. Huge volumes of data can be more effectively utilized with the use of platforms available in Big Data. It not just provides storage space for a large amount of data but can be used to extract value out of that data. Over the past few years, Big Data has emerged as the next step in technology, where to engage Big Data or not is the question to think. Organizations are contemplating how to use Big Data. Insurance, telecom, healthcare banking and many more sectors are actively taking up Big Data. Purpose of this review paper is to know the concepts of big data analytics which can be beneficial for beginners who are interested in this technology.

Keywords— *Big Data, V's, Structured data*

1. INTRODUCTION

Every day in our day to day life, we create 2.5 quintillion bytes of data. Today 90 percent of the data in the world 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.

Big Data is a term for data sets that are rapidly growing or so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, search, sharing, storage, transfer, visualization and querying information privacy.

'Big-data' is similar to 'Small-data', but bigger and have data bigger consequently requires different approaches techniques, tools and architectures. It is used to solve new problems and Old problems in a better way. Data are collected to perform various analysis. The kind of data each and every industry collects may differ from each other.

Thus Big data is nothing but the large volume of data where:

- It is very difficult to categorize.
- Different format of data.
- Data coming from Variety of sources.
- There is no upper limit and lower limit in terms of the volume of data.
- Improper frequency in the receipt of data etc.

Big data is becoming one of the most important technology trends that have the potential for dramatically changing the way organizations use the information to enhance the customer experience and transform their business models. Big data is not a single market. Rather, it is a combination of data-management technologies that have evolved over time. Big data enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time to gain the right insights. The key to understanding big data is that data has to be managed so that it can meet the business requires a given solution is designed to support [1].

2. DATA AT A GLANCE AND WHY BIG DATA

Here is a comparative scale of bytes wonderfully presented in infographic form for easy understanding and reference [3].

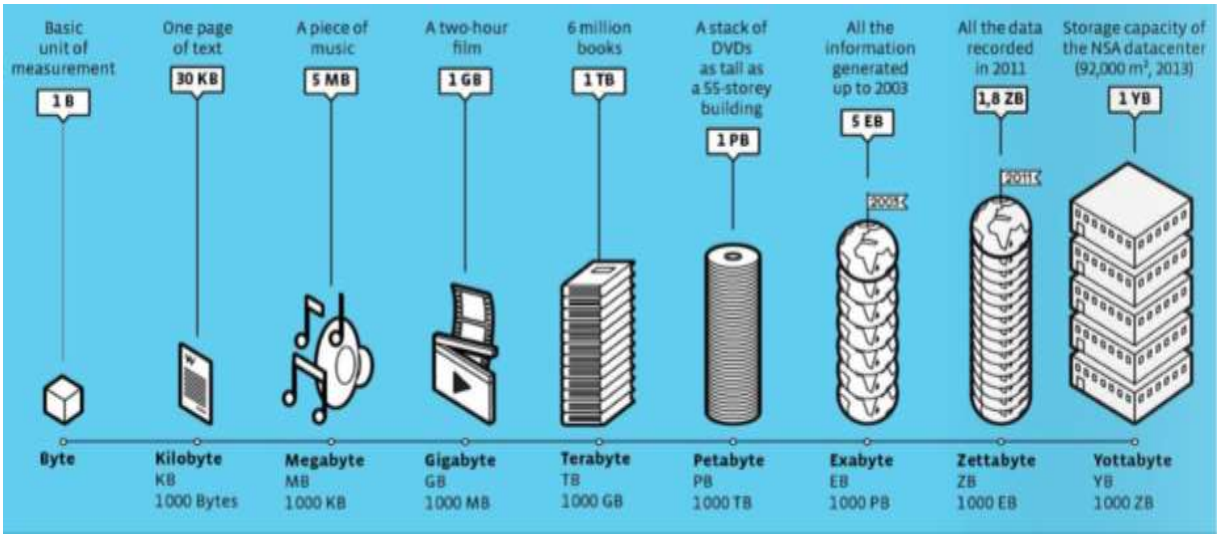


Fig. 1: Comparative scale of bytes

Fig. 2: Comparison with Big data

So low cost and high performance is the key factor that promotes the use of Big data analysis. Key enablers for the appearance and growth of 'Big-Data.

- Increase in storage capabilities
- Increase in processing power
- Availability of data
- Operational optimization
- Actionable intelligence
- Accurate predictions
- Fault and fraud detection
- More detailed records
- Improved decision-making
- Scientific discoveries
- Faster, better decision making

3. DATA AND ITS IMPORTANCE

The data processed by Big Data solutions can be human generated or machine generated. Human-generated data is the result of human interaction with systems, such as online services and digital devices, on the other hand, machine-generated data is generated by software programs and hardware devices in response to real-world events.

The primary types of data are:

(a) Structured data-It conforms to a data model or schema and is often stored in tabular form. It is used to capture relationships between different entities and is, therefore, most often stored in relational database. Structured data is frequently generated by enterprise applications and information systems like ERP and CRM systems. Due to the abundance of tools and databases that natively support structured data, it rarely requires special consideration in regards to processing or storage. Examples of this type of data include banking transactions, invoices, and customer records.

(b) Unstructured data-Data that does not conform to a data model or data schema is known as unstructured data. It is estimated that unstructured data makes up 80 per cent of the data within any given enterprise. Unstructured data has a faster growth rate than structured data. Video, image and audio files are all types of unstructured data. Special purpose logic is usually required to process and store unstructured data.

(c) Semi-structured data –It has a defined level of structure and consistency but is not relational in nature, instead, it is hierarchical or graphs based. Examples of common sources of semi-structured data include electronic data interchange (EDI) files, spreadsheets and sensor data. Semi-structured data often have special pre-processing and storage requirements [2].

Thus data is a crucial part of the process; it is important due to the following reasons:

- To understand the business.
- analyze the trend in markets
- To study the requirements/Feedbacks from customers
- To understand what our competitors are performing in markets etc

4. BIG DATA CHARACTERISTICS

Big Data characteristics can be used to help differentiate data categorized as 'Big' from other forms of data. It is initially classified as volume, velocity and variety also known as 3V's characteristics.
These characteristics are shown as below [5].

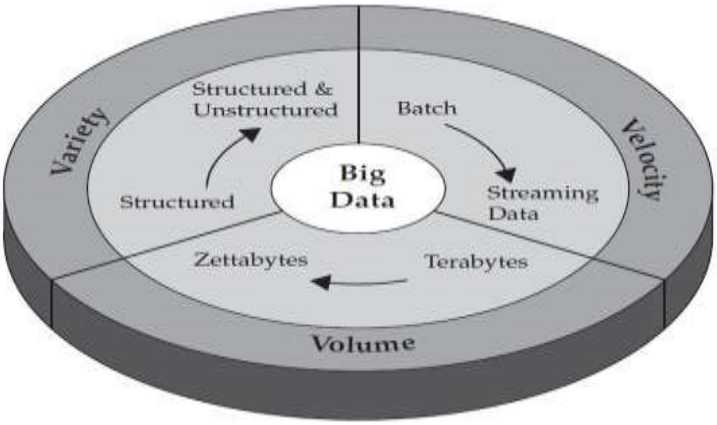


Fig. 3: Big data characteristics

5. CONCEPT OF 4V'S

Apart from the above characteristics, a new 'V' named as 'veracity' has been added as fourth 'V' (shown in the figure below [4]) in order to account for the lower signal to noise ratio of unstructured data as compared to structured data sources.

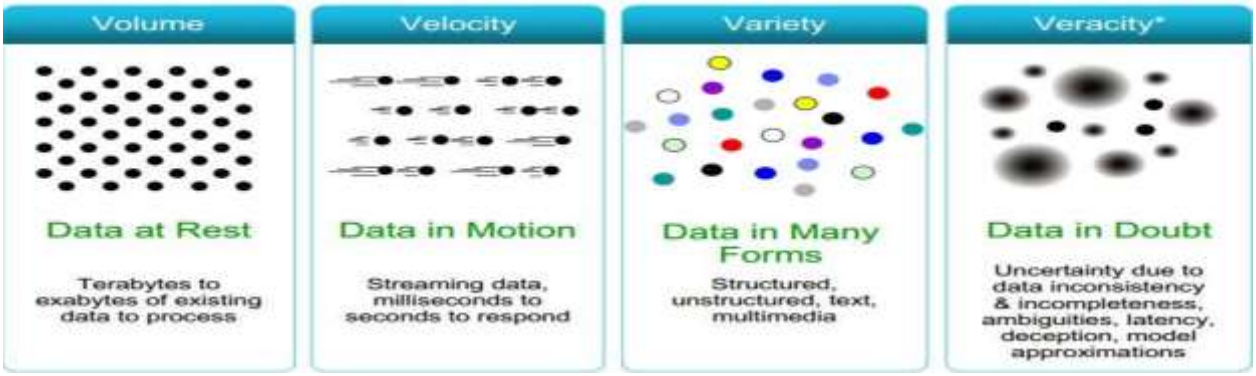


Fig. 4: The concept of 4V's

5.1 Volume

It is data size, enterprises are awash with ever-growing data of all types, easily amassing terabytes, even petabytes of information. The anticipated volume of data that is processed by Big data solutions is substantial and ever-growing. High data volumes impose distinct data storage and processing demands, as well as additional data preparation and management processes. Typical data sources that are responsible for generating high data volumes can include:

- Online transactions such as point of sale and banking
- Sensors such as GPS sensors, RFIDs, smart meters [2]. (Convert 350 billion annual meter readings to better predict power consumption)
- Social media such as Twitter and Facebook [2]. (Turn 12 terabytes of Tweets created each day into improved product sentiment analysis)

5.2 Velocity

Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into the enterprise in order to maximize its value.

For example:

- Scrutinize 5 million trade events created each day to identify potential fraud.
- Analysis of 500 million daily call detail records in real-time to predict customer churn faster.

In Big data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time. From an enterprise's point of view, the velocity of data translates into the amount of time it takes for the data to be processed once it enters the enterprise's perimeter. Coping with the fast inflow of data requires the enterprise to design highly elastic and available data processing solutions and corresponding data storage capabilities [2].

5.3 Variety

Big data is any type of data like structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.

For example:

- Monitor 100's of live video feeds from surveillance cameras to target points of interest.
- Exploit the 80% data growth in images, video and documents to improve customer satisfaction.

Thus data variety refers to the multiple formats and types of data that need to be supported by big data solutions. Data variety brings challenges for enterprises in terms of data integration, transformation, processing and storage [2].

5.4 Veracity

Most of the business leaders don't trust the information they use to make decisions. How can you act upon information if you don't trust it? Establishing trust in big data presents a huge challenge as the variety and number of sources grow.

Veracity refers to the quantity or fidelity of data. Data that enters Big Data environments need to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise. In relation to veracity, data can be part of the signal or noise of a dataset. Noise is data that cannot be converted into information and thus has no value, whereas signals have value and lead to meaningful information, Data with a high signal to noise ratio has more veracity than data with a lower ratio [2].
Apart from above 4V's a 5th V (value) has also been included as integral characteristics of Big Data.

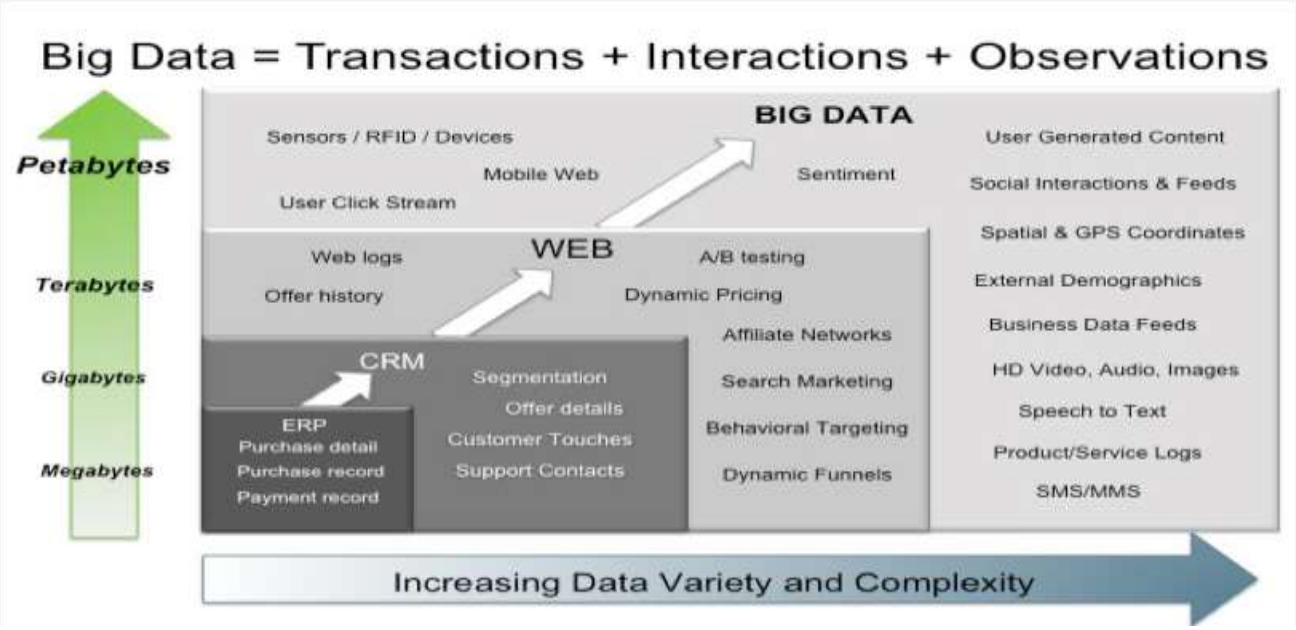
5.5 Value

Value starts and ends with the business use case. The business must define the analytic application of the data and its potential associated value to the business.

Value is defined as the usefulness of data for an enterprise. Value is also dependent on how long data processing takes because analytics results have a shelf life, for example, a 20 minute delayed stock quote has little to no value for making a trade compared to a quote that is 20 milliseconds old[2].

6. INCREASING AREAS OF BIG DATA

Big Data is the combination of Transactions, Interactions and Observations. The following graphic illustrates denotes the increasing areas o big data [6]



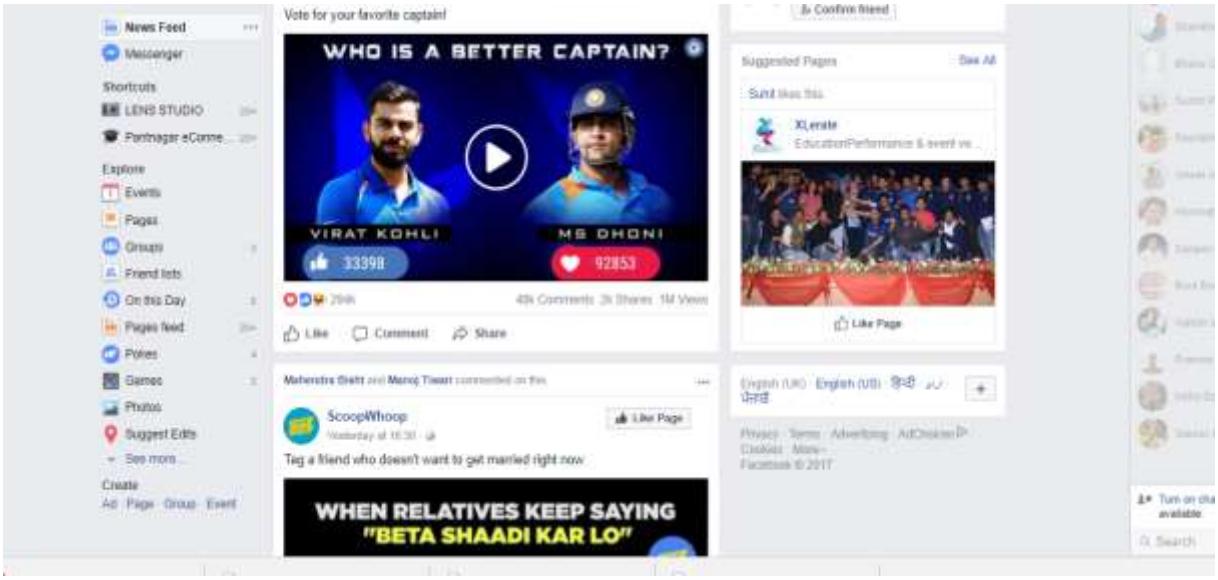
ERP, CRM, and transactional Web applications are classic examples of systems processing Transactions. Highly structured data in these systems is typically stored in SQL databases. Interactions are about how people and things interact with each other or with your business. Web Logs, User Click Streams, Social Interactions & Feeds, and User-Generated Content are classic places to find Interaction data. Observational data tends to come from the "Internet of Things". Sensors for heat, motion, pressure and RFID and GPS chips within such things as mobile devices, ATM machines, and even aircraft engines provide just some examples of "things" that output Observation data [6].

6.1 The easy to understand example for Big data

Almost the entire young population is now on social media. It is one of the best examples where the data will be coming in a different manner.

- There may be audio files
- There may be images
- Videos
- Text messages
- Documents
- Any others.

In the below figure a general home page of social media is shown that how it contains, images, post, videos i.e. a variety of data which is in huge amount for one account, like that there are different users having their own account with different data, here Big Data comes into the picture.



7. NEED FOR BIG DATA

7.1 Case 1

Let us assume a scenario where a Multinational Bank wanted to make a product on loan area. Usually, these types of Banks will have their presence in all over the world. So if they wanted to design a new loan product, then their simple requirements may be as below:

- The potential growth of each small unit of Geography where the Bank has its presence.
- The regulatory guidelines of each different Geography
- The percentage of the sale of such type of loan products in those Geographies.
- The percentage of defaulters of Loan products in a Geography
- Details about the different competitor's loan products in that area
- The upcoming risks in that Geography such as political conditions

The above requirements are not even a few percentages of the entire requirements. Imagine how a Bank will analyze all these data. Also, there will not be any defined format of the data for each of the above requirement. The historical data may come from multiple different sources and it will have multiple different formats and also the complexity.

7.2 Case 2

In another scenario, where a Bank wanted to process the loan request of a customer, the bank may need at least the below data.

- The credit history of the customer
- The past transactions of that customer
- The average account balance of that customer
- Checking for fraudulent attempts
- The number of proper closures of any other loans
- The capability of the customer to repay the EMI
- For a single customer, if this much of data needed, then how the banks are doing thousands of loan processing in a day.

Here is the need for Big data. If the tenure of the loan is only 5 years, then the Bank may check for the historical data of at least 5 years of that customer. But imagine if the loan request for this customer is to start a new business where he is in a need of 5 billion pounds and the repayment tenure is 20 years. So here is the need for big data.

8. PRACTICAL APPROACH ON BIG DATA APPLICATION

Big Data has changed the way we manage, analyze and leverage data in any industry. Most promising areas where big data can be applied to make a change is healthcare and partner companies.

In this approach we can interact with our routine health-related data which is available on the basis of our day to day physical activity, thus on monthly/quarterly/half yearly/yearly accumulation of this data which is huge in amount and can be used for health-related precautionary measures.

Below process can be followed to achieve this.

Step 1: These days we are having lots of smartwatches which can be used as a fitness tracker. These fitness trackers rely on sensors that track the movement of the users. These watches are useful in recording our daily routine activity like heartbeat rate, sleeping pattern, calories burn, steps and running activities and so on. We can get the data of all such activities and on the basis of these data, it can be analyzed a pattern on a healthy lifestyle. The pattern can be made once we have a big and huge amount of data that can be significant to get an overall output.

Step 2: Once we have the data, then it can be used in synchronization with the supporting applications, which are readily available in market (from play store we can install the apps in our phone where it will require to fill the basic details like name and mail address followed by pairing of smartwatch to the app via Bluetooth option) so these apps in coordination with the data from smartwatches can generate the overall report. Here if we are more active then we can get the reward points as an appreciation.

Step 3: This report can be received by our medical insurance company, and they can send these reports to the relevant medical specialist (only after the pre-approval from customer like us) and accordingly as per the data available on the report medical specialists can further provide proper consultation for improvement on consistent health activities.

Medical companies can share a growth model so that they can have partners in different areas like airlines, hotels, retail stores, sports companies. So based on the reward points earned by the user during health activities they can get a huge discount on the services/product of partner companies. Thus in a business point of view, it will not only enhance the profit and wealth of organization but also make coordination among them.

Above is the health activity data of one user that is utilizing in the whole process, like that if we have lots of users then this data (which is big and huge now) can be useful in different business aspects including healthcare.

9. ADVANTAGES OF BIG DATA ANALYTICS AND HOW IT APPLIED ON DIFFERENT SCENARIO

- It is apparent that many companies in India and abroad have started using Big Data to solve real-world problems, to get an edge in the marketplace and to bring efficiency in their factories. The enormous amount of Big Data being churned out is giving rise to numerous opportunities for both companies as well as their clients. Big Data is expected to be a huge market by 2020. It offers good business potential to companies that are operating in the infrastructure, services and analytics space. Because Big Data analysis can provide all sorts of insights, it also offers a competitive edge to companies in any industry or sector that figures out how to harness and use the data properly.
- Now, the question that arises is - what kind of uses can Big Data be put to in practical terms? It has been used by companies in vastly different sectors– from a bank that wanted to give customized messages to people drawing money at ATMs to a city corporation that wanted to predict when its buses were likely to fail.
- It is been suspected that in elections, Big Data will also be used by our political parties to recognize people preferences and plan strategies. In the United States, this approach was used during the Presidential elections. The campaign management team collated data from various aspects like polling, fundraising, volunteers, and social media into a central database. Then they were able to assess individual voters' online activities and ascertain whether campaign tactics were producing results. Based on the analysis, the campaign team targeted messaging and communications at individual voter levels, which prompted exceptionally high turnout: this was considered one of the critical factors in Obama's re-election. Product Innovation. Not all Big Data is new data. There is a wealth of information sitting unused within the corporate data repositories or at least not used effectively.
- That is the power of Big Data but that is just one way in which it can be useful. Like individuals, governments and organizations increasingly transact digitally and store data from sources as diverse as bank transactions, telecom calls, credit card usage, social media posts, retail sales, factory and machine outputs etc.

10. CHALLENGES IN HANDLING BIG DATA

- Data may be structured, Semi or un-structured it's hard to collect, Manage, Search, Store, Process & Analytics, Visualizing, Skills, Tools, Cost etc
- Imagine Google process about 24 petabytes of Data every day [1 petabyte is 1000 terabyte]

- Managing this much vast data is not possible in Traditional Databases
- Google came up with a solution for this using Google File system (GFS)
- GFS: Google file system a distributed parallel processing file system which splits the data and loads into its clusters, its fault-tolerant and easily scalable
- Big Table: Replacement for Relational Database Tables
- MapReduce: Framework to process Large Data sets

- Google published papers on these which leads to Evolution of Hadoop
- Hadoop: Apache Hadoop is an open source distributed software platform for storing and processing data. Written in Java, it runs on a cluster of industry-standard servers configured with direct-attached storage. Using Hadoop, one can store petabytes of data reliably on tens of thousands of servers while scaling performance cost-effectively by merely adding inexpensive nodes to the cluster.
- MapReduce: It is a software framework to write applications to process huge datasets in a parallel distributed environment on Hadoop clusters
- Hbase: Database similar to BigTable, sits on Top of HDFS
- HDFS: Hadoop distributed file system is similar to the Google file system.

11. BIG DATA – THE SAVIOR

- Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information and difficult to analyze and handle using common database management tools.
- The storage industry is continuously challenged as Big Data increases exponentially. While the physical storage can be enhanced with more terabyte drive arrays, the software infrastructure must be flexible enough to quickly and economically accommodate ever greater volumes of transactions and queries.
- The analytical challenge is deriving meaningful information from data in petabyte and exabyte volumes. Big Data analytics breaks down the data sets into smaller chunks for efficient processing and employs parallel computing to derive intelligence for effective decision-making.

12. MYTHS ABOUT BIG DATA

- **Big Data is only about massive data volume**: Volume is just one key element in defining Big Data, and it is arguably the least important of three elements. The other two are variety and velocity.
- **Big Data means Hadoop**: Hadoop is the Apache open-source software framework for working with Big Data. It was derived from Google technology and put to practice by Yahoo and others. But, Big Data is too varied and complex for a one-size-fits-all solution. While Hadoop has surely captured the greatest name recognition, it is just one of three classes of technologies well suited to storing and managing Big Data. The other two classes are NoSQL and Massively Parallel Processing (MPP) data stores.
- **Big Data Means unstructured data:** The term “unstructured" is imprecise and doesn’t account for the many varying and subtle structures typically associated with Big Datatypes. Also, Big Data may well have different data types within the same set that do not contain the same structure. Therefore, Big Data is probably better termed “multi-structured” as it could include text strings, documents of all types, audio and video files, metadata, web pages, email messages, social media feeds, form data, and so on
- **Big Data is for social media feeds and sentiment analysis:** Simply put, if your organization needs to broadly analyze web traffic, IT system logs, customer sentiment, or any other type of digital shadows being created in record volumes each day, Big Data offers a way to do this. Even though the early pioneers of Big Data have been the largest, web-based, social media companies- Google, Yahoo, Facebook- it was the volume, variety, and velocity of data generated by their services that required a radically new solution rather than the need to analyze social feeds or gauge audience sentiment.

13. CONCLUSION

Big Data is similar to Rubik’s cube, in which aim of all the firm and expert is same to get maximum useful content from data but the way and the initial approaches are different for each organization. Big data is a complex matter so one needs to know the business domain and requirements. As organizations are studying and architecting big data solutions they are also getting the ways and opportunities which are associated with Big Data. There is not a unique solution to this technology as well there is not a unique vendor which can claim to know all about Big Data. Honestly, Big Data is a too deep concept and there are many players, tools, architectures, vendors and procedures, so it is essential to select the appropriate findings as per the need.

14. REFERENCES

[1] Big Data for Dummies, A Wiley Brand
[2] Big data Fundamentals Concepts, Drivers and Techniques by Thomas Erl
[3] http://www.cnrs.fr/fr/pdf/cim/28/index.html#/22/
[4] https://www.datasciencecentral.com/profiles/blogs/data-veracity
[5] https://apandre.wordpress.com/2013/11/19/datawatch/
[6] https://hortonworks.com/blog/7-key-drivers-for-the-big-data-market/

