Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 5): pp. 9548-9554
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.1040

Classification Of Big Data In The Present Era With Challenges

Ms. Rohini D. Dhage VMV Commerce, JMT Arts and JJP Science College,Wardhaman Nagar, Nagpur. rohinidhananjaydhage@yahoo.co.in

Dr. Vaibhav R. Bhedi VMV Commerce, JMT Arts and JJP Science College,Wardhaman Nagar, Nagpur. vaibhav_bhedi@rediffmail.com

Abstract:

Big data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Big data philosophy encompasses unstructured, semi-structured, and structured data; however, the main focus is on structured data. The "big data" philosophy is based on the use of a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of massive scale. In the big data era, data is created in real-time or near-real-time, and data can be stored, processed, and managed across databases that are dotted around anywhere in the world. This paper focuses on the importance of big data in the context of the Internet and how it can be used in the future.

Keyword: Big Data, Volume, Variety, Velocity.

I. Introduction:

The "big data"isthedata whose scale,diversity,and complexity require newarchitectures,techniques,algorithms,andanalysisto manage it and extract value and hidden knowledge from it.[1] Data has alwaysexistedand there has been a needto store, process,andmanage datasince human civilization and society began.However, the amount and type of data captured, stored,processed,and managed depended on variousfactors,including theneed forhumans, availabletools and technologiesfor storage, processing, management,effort, and costs, theability to gaindata insights,make decisions,etc[2].

The term has been in use since the 1990s, with some giving credit to John Mashey for coining or at least making it popular.Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big Data philosophy encompasses unstructured, semi-structured, and structured data, however, the main focus is on unstructured data.Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.Big data requires a set of techniques and technologies with new

forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale[3].

Processing information like this illustrates why big data has become so important:

- Most data collected now is unstructured and requires different storage and processing than that found in traditional relational databases.
- Available computational power is skyrocketing, meaning there are more opportunities to process big data.
- The Internet has democratized data, steadily increasing the data available while also producing more and more raw data.

It is important to realize that big data comes in many shapes and sizes. It also has many different uses – real-time fraud detection, web display advertising and competitive analysis, call center optimization, social media, and sentiment analysis, intelligent traffic management, and smart power grids, to name just a few. All of these analytical solutions involve significant (and growing) volumes of both multi-structured and structured data. Many of these analytical solutions were not possible previously because they were too costly to implement, or because analytical processing technologies were not capable of handling the large volumes of data involved promptly. In some cases, the required data simply did not exist in an electronic form.

II. Characteristics of 'Big Data':

The original three 'V' Dimension Characteristics of Big Data identified in 2001 are in Fig 1:

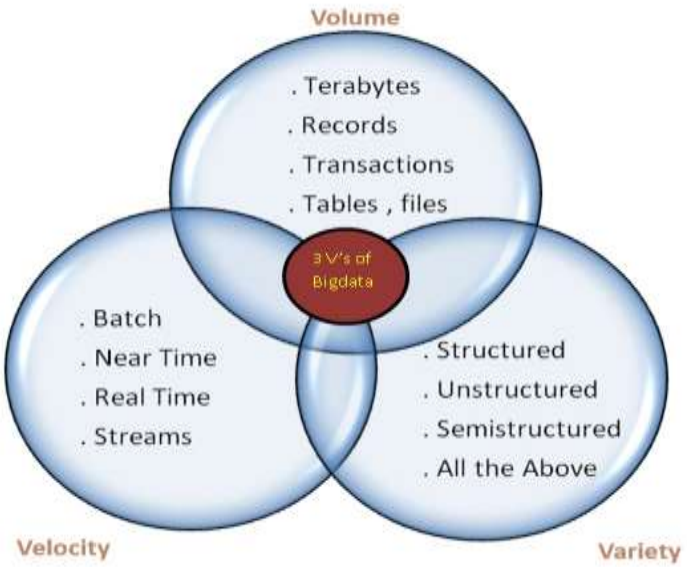

Fig 1: Big Data Identification

1) Volume (amount of data and the size of the data set)

Volume Refers to the vast amounts of data generated every second. We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. This makes most data sets too large to store and analyze using traditional database technology. New big data tools use distributed systems so that we can store and analyze data across databases that are dotted around anywhere in the world.[4]

2) Velocity (speed of data in and out or data in motion)

Velocity Refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyze the data while it is being generated (sometimes referred to as inmemory analytics), without ever putting it into databases.

Velocity is the speed at which the data is created, stored, analyzed, and visualized. In the past, when batch processing was common practice, it was normal to receive an update from the database every night or even every week. Computers and servers require substantial time to process the data and update the databases. In the big data era, data is created in real-time or near real-time. With the availability of Internet-connected devices, wireless or wired, machines, and devices can pass on their data the moment it is created.

3) Variety (range of data types, domains, and sources)

Variety Refers to the different types of data we can now use. In the past, we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. 80% of the world's data is unstructured (text, images, video, voice, etc.) With big data technology, we can now analyze and bring together data of different types such as messages, social media conversations, photos, sensor data, and video or voice recordings.

In the past, all data that was created was structured data, it neatly fitted in columns and rows but those days are over. Nowadays, 90% of the data that is generated by an organization is unstructured data. Data today comes in many different formats: structured data, semi-structured data, unstructured data, and even complex structured data. The wide variety of data requires a different approach as well as different techniques to store all raw data.

There are many different types of data and each of those types of data require different types of analyses or different tools to use. Social media like Facebook posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are.It can be structured, semi-structured, or unstructured.

Big Data can be **structured, unstructured, and semi-structured** and are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will come in array forms, that are **PDFs, Emails, audio, SM posts, photos, videos,** etc.

The data is categorized as below:

Structured data: In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

Semi-structured: In Semi-structured, the schema is not appropriately defined, e.g., JSON, XML, CSV, TSV, and email. OLTP (Online Transaction Processing) systems are built to work with semi-structured data. It is stored in relations, i.e., tables.

Unstructured Data: All the unstructured files, log files, audio files, and image files are included in the unstructured data. Some organizations have much data available, but they do not know how to derive the value of data since the data is raw.

Quasi-structured Data: The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools [5].

III. The Big Data Challenge:

Storage

With vast amounts of data generated daily, the greatest challenge is storage (especially when the data is in different formats) within legacy systems. Unstructured data cannot be stored in traditional databases.

Processing

Processing big data refers to the reading, transforming, extraction, and formatting of useful information from raw information. The input and output of information in unified formats continue to present difficulties.

Security

Security is a big concern for organizations. Non-encrypted information is at risk of theft or damage by cyber-criminals. Therefore, data security professionals must balance access to data against maintaining strict security protocols.

Finding and Fixing Data Quality Issues

Many of you are probably dealing with challenges related to poor data quality, but solutions are available. The following are four approaches to fixing data problems:

- Correct information in the original database.

- Repairing the original data source is necessary to resolve any data inaccuracies.
- You must use highly accurate methods of determining who someone is.

Scaling Big Data Systems

Database sharding, memory caching, moving to the cloud, and separating read-only and write-active databases are all effective scaling methods. While each one of those approaches is fantastic on its own, combining them will lead you to the next level.

Evaluating and Selecting Big Data Technologies

Companies are spending millions on new big data technologies, and the market for such tools is expanding rapidly. In recent years, however, the IT industry has caught on to big data and analytics potential. The trending technologies include the following:

- Hadoop Ecosystem
- Apache Spark
- NoSQL Databases
- R Software
- Predictive Analytics
- Prescriptive Analytics

Big Data Environments

In an extensive data set, data is constantly being ingested from various sources, making it more dynamic than a data warehouse. The people in charge of the big data environment will quickly forget where and what each data collection came from.

Real-Time Insights

The term "real-time analytics" describes the practice of performing analyses on data as a system is collecting it. Decisions may be made more efficiently and with more accurate information thanks to real-time analytics tools, which use logic and mathematics to deliver insights on this data quickly.

Data Validation

Before using data in a business process, its integrity, accuracy, and structure must be validated. The output of a data validation procedure can be used for further analysis, BI, or even to train a machine learning model.

Healthcare Challenges

Electronic health records (EHRs), genomic sequencing, medical research, wearables, and medical imaging are just a few examples of the many sources of health-related big data.

Barriers to Effective Use Of Big Data in Healthcare

- The price of implementation
- Compiling and polishing data
- Security
- Disconnect in communication [6].

IV. Big Data Architecture:

The first thing that comes to mind about Big Data is MapReduce and Distributed File System (DFS). Each enterprise derives and uses a different model because Big Data has not only one architecture. The first MapReduce was developed by Google in 2002, Hadoop was developed by Yahoo in 2006, and Hive by Facebook in early 2008. Hive, Impala, Spark, Cassandra, Pig, and HBase all use the MapReduce model. So, here we will explain MapReduce and HDFS [7].

Hadoop Distributed File System (HDFS):

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets [8].

MapReduce:

MapReduce has emerged as a popular way to harness the power of large clusters of computers. MapReduce allows programmers to think in a data-centric fashion: they focus on applying transformations to sets of data records, and allow the details of distributed execution, network communication, and fault tolerance to be handled by the MapReduce framework [9].

V. Conclusion:

Big data is data that exceeds the processing capacity of conventional database systems. In this paper fundamental concepts about Big Data are presented. This paper describes the new concept of Big data, its importance, and the existing projects. There is no doubt that the industries are going ablaze with the huge eruption of data. None of the sectors have remained untouched by this drastic change in a decade. Technology has crept into each business arena and hence, it has become an essential part of every processing unit. Talking about the IT industry specifically, software and automation are the bare essential terms and are used in every phase of a process cycle.

VI. References:

1. CS525: Special Topics in DBs Large-Scale Data Management. Spring 2013 WPI, Mohamed Eltabakh**.** http://web.cs.wpi.edu/~cs525/s13-MYE/
2. Big Data Basics - Part 1 - Introduction to Big Data. https://www.mssqltips.com/sqlservertip/3132/big-data-basics--part-1--introduction-to-big-data/
3. Wikipedia https://en.wikipedia.org/wiki/Big_data#Characteristics admin May 26, 2015
4. http://www.dataintensity.com/characteristics-of-big-data-part-one/
5. https://www.javatpoint.com/big-data-characteristics
6. https://www.simplilearn.com/challenges-of-big-data-article
7. Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2020). Big data characteristics, architecture, technologies, and applications. J Comput Sci, 16, 817-824.
8. Borthakur, D. (2010). HDFS architecture. Document on Hadoop Wiki. URL http://hadoop. apache. org/common/docs/r0, 20.
9. Condie, T., Conway, N., Alvaro, P., Hellerstein, J. M., Elmeleegy, K., & Sears, R. (2010, April). MapReduce online. In Nsdi (Vol. 10, No. 4, p. 20).