KJSCE/IT/B.Tech./SEMVII/BDA/2020-21



**Experiment No.1**

**Title: Understanding the Big Data**

KJSCE/IT/B.Tech./SEMVII/BDA/2020-21

**Batch:B1** **Roll No.:1824008** **Experiment No.:1**

**Title: Understanding the Big data**

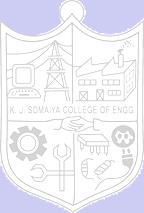
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed: Reference books, Internet \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Theory:**

**Big Data:**

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data [1][2].



**Importance of Big data:**

When big data is effectively and efficiently captured, processed, and analysed, companies are able to gain a more complete understanding of their business, customers, products, competitors, etc. which can lead to efficiency improvements, increased sales, lower costs, better customer service, and/or improved products and services[2].

Big Data will help to create new growth opportunities and entirely new categories of companies, such as those that aggregate and analyse industry data. Many of these will be companies that sit in the middle of large information flows where data about products and services, buyers and suppliers, consumer preferences and intent can be captured and analysed[2].

Big Data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services. Sophisticated analytics can substantially improve decision-making, minimise risks, and unearth valuable insights that would otherwise remain hidden. Big Data can be used to develop the next generation of products and services [3].

Big data expands the possible domains of application for algorithms and machine-mediated analysis. At some manufacturers, for example, algorithms analyze sensor data from production lines, creating self-regulating processes that cut waste, avoid costly (and sometimes dangerous) human interventions, and ultimately lift output [3].

Page No:

KJSCE/IT/B.Tech./SEMVII/BDA/2020-21

**Procedure:**

**Students should explore the answers of following questions and submit as write up:**

1. **What is the Big data?**

Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

## Characteristics Of Big Data

**(i) Volume –** The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data.

**(ii) Variety –** The next aspect of Big Data is its **variety**.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

**(iii) Velocity –** The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors,[Mobile](https://www.guru99.com/mobile-testing.html)devices, etc. The flow of data is massive and continuous.

1. **What are the sources of the Big data?**

**Social networks:** Arguably, the primary source of all big data that we know of today is the social networks that have proliferated over the past 5-10 years. This is by and large unstructured data that is represented by millions of social media postings and other data that is generated on a second-by-second basis through user interactions on the web across the world. Increase in access to the internet across the world has been a self-fulfilling act for the growth of data in social networks.

**Media:** Largely a result of the growth of social networks, media represents the millions, if not billions, of audio and visual uploads that take place on a daily basis. Videos uploaded on YouTube, music recordings on SoundCloud, and pictures posted on Instagram are prime examples of media, whose volume continues to grow in an unrestrained manner.

**Data warehouses**: Companies have long invested in specialized data storage facilities commonly known as data warehouses. A DW is essentially collections of historical data that companies wish to maintain and catalog for easy retrieval, whether for internal use or regulatory purposes. As industries gradually shift toward the practice of storing data in platforms such as Hadoop and NoSQL, more and more companies are moving data from their pre-existing data warehouses to some of the newer technologies. Company emails, accounting records, databases, and internal documents are some examples of DW data that is now being offloaded onto Hadoop or Hadoop-like platforms that leverage multiple nodes to provide a highly-available and fault-tolerant platform.

**Sensors:** A more recent phenomenon in the space of big data has been the collection of data from sensor devices. While sensors have always existed and industries such as oil and gas have been using drilling sensors for measurements at oil rigs for many decades, the advent of wearable devices, also known as the Internet Of Things such as Fitbit and Apple Watch, meant that now each individual could stream data at the same rate at which a few oil rigs used to do just 10 years back

1. **What are the variants of the Big data?**

## Structured, Semi-structured and Unstructured data

|  |  |  |  |
| --- | --- | --- | --- |
| **Factors** | **Structured data** | **Semi-structured data** | **Unstructured data** |
| Flexibility | It is dependent and less flexible | It is more flexible than structured data but less than flexible than unstructured data | It is flexible in nature and there is an absence of a schema |
| Transaction Management | Matured transaction and various concurrency technique | The transaction is adapted from DBMS not matured | No transaction management and no concurrency |
| Query performance | Structured query allow complex joining | Queries over anonymous nodes are possible | An only textual query is possible |
| Technology | It is based on the relational database table | It is based on RDF and XML | This is based on character and library data |

**Example Structured Data:**

Human-generated structured data mainly includes all the data a human input into a computer, such as his name and other personal details. When a person clicks a link on the internet, or even makes a move in a game, data is created- this can be used by companies to figure out their customer behavior and make the appropriate decisions and modifications.

Top 3 players who have scored most runs in international T20 matches are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Player | Country | Scores | No of Matches played |
| Brendon McCullum | New Zealand | 2140 | 71 |
| Rohit Sharma | India | 2237 | 90 |
| Virat Kohli | India | 2167 | 65 |

**Examples of unstructured data** include text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery – the list goes on and on.

**Examples of semi-structured** : CSV but  XML and JSON documents are semi structured documents,  NoSQL databases are considered as semi structured.

1. **Differentiate between Veracity and Variety of Big Data.**

**\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
| **Veractiy** | **Variety** |
| Data in a doubt | Data in many forms |
| Uncertanity due to data inconsistency and incompleteness,ambiguities,latency, deception and model apporoximations | Structure, unstructured text and multimedia |
| Quality,realiablity,uncertainty and meaning in data itself | Different data types and souces example  Relations, documents, web, xml files, grahps, multimedias, IOT, dark data,open data, external data, Data Integreation ETL ELT data viertualization |
| Data must be able to verified based on both accuracy and context | Data comes from all types of formats.This can include data generated within an organization as well asdata created from external sources, including publicely available data |
| An example of a high veracity data set would be data from a medical experiment or trial. | An example of high variety data sets would be the CCTV audio and video files that are generated at various locations in a city. |
| Another example weather forecast data. | An another example is past cricket matches data. |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Outcomes:**

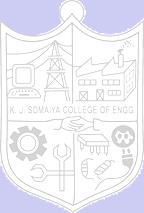
Understand what Big Data Analytics is.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Conclusion: (Conclusion to be based on the objectives and outcomes achieved)**

Learned more about the big data. Understood characteristics of the Big data. Able to written difference between the Variety and Veracity of the big data.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**



**Grade: AA / AB / BB / BC / CC / CD /DD**

**Signature of faculty in-charge with date**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**References:**

**Books/ Journals/ Websites:**

[1]Paul Zikopoulos, Chris Eaton, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data’, McGraw Hill Education

[2]<http://www.navint.com/images/Big.Data.pdf>

[3][http://iveybusinessjournal.com/publication/why-big-data- is-the-new-competitive-advantage/](http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/)

Page No: