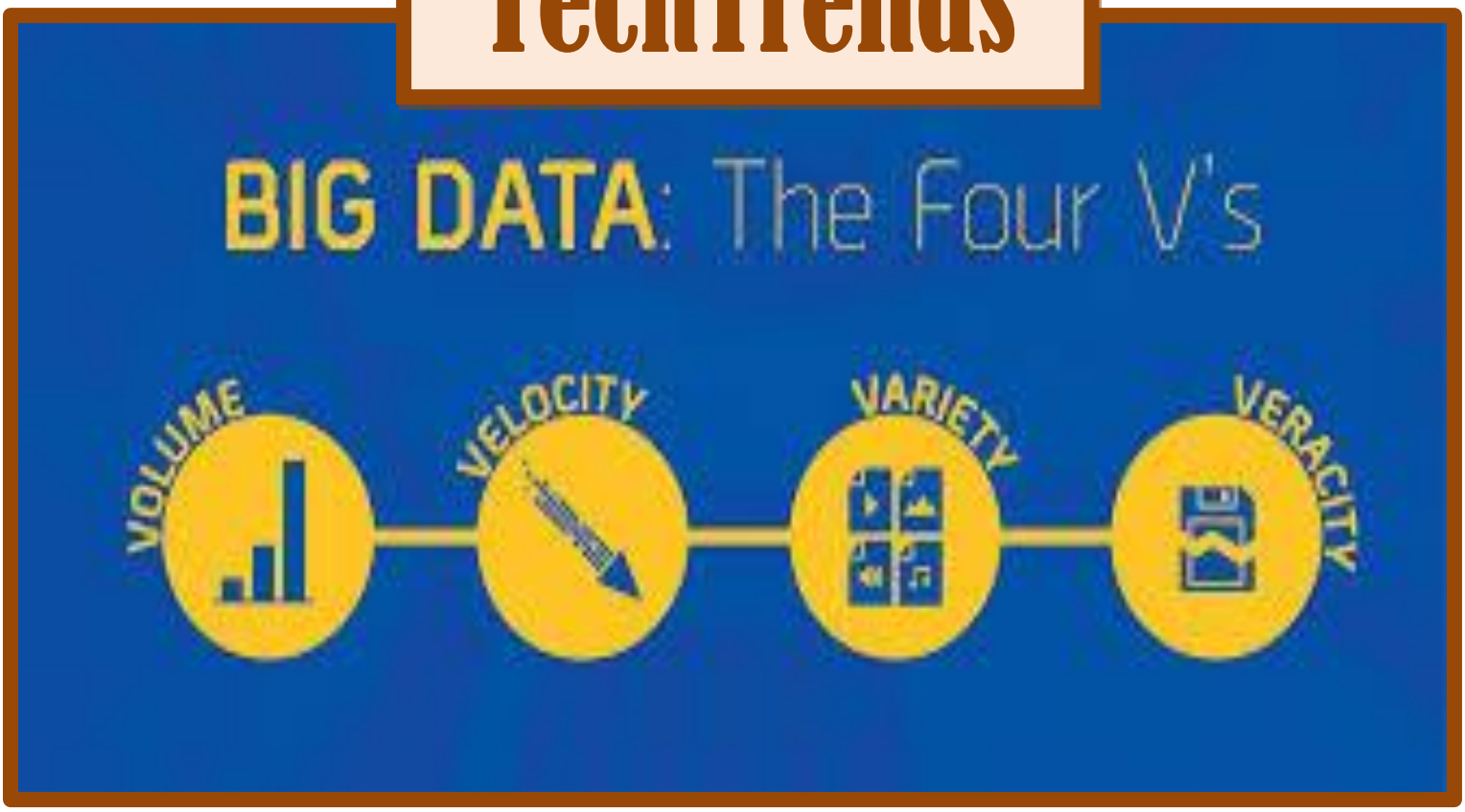
Department of Computer Engineering

Government Polytechnic for Girls, Surat

December-2020 Vol.-15

TechTrends

E-Newsletter



Vision:

To empower girls of diploma computer engineering to excel in IT Industries and serve the society.

Mission:

- To strive for academic excellence and professional competence among students and staff.
- To encourage innovative ideas among students to enhance their entrepreneurship skills.
- To provide high tech educational resources and supportive infrastructure.

Follow us on

gpgdceenewsletter@gmail.com

gpgdceenewsletter@gmail.com

What is Big Data?

Big Data as a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

In simple terms, "Big Data" consists of very large volumes of heterogeneous data that is being generated, often, at high speeds. These data sets cannot be managed and processed using traditional data management tools and applications at hand. Big Data requires the use of a new set of tools, applications and frameworks to process and manage the data.

Kum. C. D. Engineer
Lecturer,
Department of
Computer Engineering

Evolution of Big Data

Data has always been around and there has always been a need for storage, processing, and management of data, since the beginning of human civilization and human societies. However, the amount and type of data captured, stored, processed, and managed depended then and even now on various factors including the necessity felt by humans, available tools/technologies for storage, processing, management, effort/cost, and ability to gain insights into the data, make decisions, and so on.

Going back a few centuries, in the ancient days, humans used very primitive ways of capturing/storing data like carving on stones, metal sheets, wood, etc. Then with new inventions and advancements a few centuries in time, humans started capturing the data on paper, cloth, etc. As time progressed, the medium of capturing/storage/management became punching cards followed by magnetic drums, laser disks, floppy disks, magnetic tapes, and finally today we are storing data on various devices like USB Drives, Compact Discs, Hard Drives, etc.

In fact the curiosity to capture, store, and process the data has enabled human beings to pass on knowledge and research from one generation to the next, so that the next generation does not have to re-invent the wheel.

As we can clearly see from this trend, the capacity of data storage has been increasing exponentially, and today with the availability of the cloud infrastructure, potentially one can store unlimited amounts of data. Today Terabytes and Petabytes of data is being generated, captured, processed, stored, and managed.

Characteristic of Big Data

(Three V's of Big Data)

When do we say we are dealing with Big Data? For some people 1TB might seem big, for others 10TB might be big, for others 100GB might be big, and something else for others. This term is qualitative and it cannot really be quantified. Hence we identify Big Data by a few characteristics which are specific to Big Data. These characteristics of Big Data are popularly known as Three V's of Big Data.

The three v's of Big Data are Volume, Velocity, and Variety as shown below.

Volume

Volume refers to the size of data that we are working with. With the advancement of technology and with the invention of social media, the amount of data is growing very rapidly. This data is spread across different places, in different formats, in large volumes ranging from Gigabytes to Terabytes, Petabytes, and even more. Today, the data is not only generated by humans, but large amounts of data is being generated by machines and it surpasses human generated data. This size aspect of data is referred to as Volume in the Big Data world.

Velocity

Velocity refers to the speed at which the data is being generated. Different applications have different latency requirements and in today's competitive world, decision makers want the necessary data/information in the least amount of time as possible. Generally, in near real time or real time in certain scenarios. In different fields and different areas of technology, we see data getting generated at different speeds. A few examples include trading/stock exchange data, tweets on Twitter, status updates/likes/shares on Facebook, and many others. This speed aspect of data generation is referred to as Velocity in the Big Data world.

Variety

Variety refers to the different formats in which the data is being generated/stored. Different applications generate/store the data in different formats. In today's world, there are large volumes of unstructured data being generated apart from the structured data getting generated in enterprises. Until the advancements in Big Data technologies, the industry didn't have any powerful and reliable tools/technologies which can work with such voluminous unstructured data that we see today. In today's world, organizations not only need to rely on the structured data from enterprise databases/warehouses, they are also forced to consume lots of data that is being generated both inside and outside of the enterprise like clickstream data, social media, etc. to stay competitive. Apart from the traditional flat files, spreadsheets, relational databases etc., we have a lot of unstructured data stored in the form of images, audio files, video files, web logs, sensor data, and many others. This aspect of varied data formats is referred to as Variety in the Big Data world.

Sources of Big Data

Just like the data storage formats have evolved, the sources of data have also evolved and are ever expanding. There is a need for storing the data into a wide variety of formats. With the evolution and advancement of technology, the amount of data that is being generated is ever increasing. Sources of Big Data can be broadly classified into six different categories.

Enterprise Data

There are large volumes of data in enterprises in different formats. Common formats include flat files, emails, Word documents, spreadsheets, presentations, HTML pages/documents, pdf documents, XMLs, legacy formats, etc. This data that is spread across the organization in different formats is referred to as Enterprise Data.

Transactional Data

Every enterprise has some kind of applications which involve performing different kinds of transactions like Web Applications, Mobile Applications, CRM Systems, and many more. To support the transactions in these applications, there are usually one or more relational databases as a backend infrastructure. This is mostly

structured data and is referred to as Transactional Data.

Social Media

This is self-explanatory. There is a large amount of data getting generated on social networks like Twitter, Facebook, etc. The social networks usually involve mostly unstructured data formats which includes text, images, audio, videos, etc. This category of data source is referred to as Social Media.

Activity Generated

There is a large amount of data being generated by machines which surpasses the data volume generated by humans. These include data from medical devices, censor data, surveillance videos, satellites, cell phone towers, industrial machinery, and other data generated mostly by machines. These types of data are referred to as Activity Generated data.

Public Data

This data includes data that is publicly available like data published by governments, research data published by research institutes, data from weather and meteorological departments, census data, Wikipedia, sample open source data feeds, and other data which is freely available to the public. This type of publicly accessible data is referred to as ***Public Data***.

Archives

Organizations archive a lot of data which is either not required anymore or is very rarely required. In today's world, with hardware getting cheaper, no organization wants to discard any data, they want to capture and store as much data as possible. Other data that is archived includes scanned documents, scanned copies of agreements, records of ex-employees/completed projects, banking transactions older than the compliance regulations. This type of data, which is less frequently accessed, is referred to as ***Archive Data***.

What is Big Data Analytics?

Big Data Analytics examines large and different types of data in order to uncover the hidden patterns, insights, and correlations. Basically, Big Data Analytics is helping large companies facilitate their growth and development. And it majorly includes applying various data mining algorithms on a certain dataset

Tools for Big Data Analytics

Apache Hadoop

Big Data Hadoop is a framework that allows you to store big data in a distributed environment for parallel processing.

Apache Pig

Apache Pig is a platform that is used for analyzing large datasets by representing them as data flows. Pig is basically designed in order to provide an abstraction over MapReduce which reduces the complexities of writing a MapReduce program.

Apache HBase

Apache HBase is a multidimensional, distributed, open-source, and NoSQL database written in Java. It runs on top of HDFS providing Bigtable-like capabilities for Hadoop.

Apache Spark

Apache Spark is an open-source general-purpose cluster-computing framework. It provides an interface for programming all clusters with implicit data parallelism and fault tolerance.

Talend

Talend is an open-source data integration platform. It provides many services for enterprise application integration, data integration, data management, cloud storage, data quality, and Big Data.

Splunk

Splunk is an American company that produces software for monitoring, searching, and analyzing machine-generated data using a Web-style interface.

Apache Hive

Apache Hive is a data warehouse system developed on top of Hadoop and is used for interpreting structured and semi-structured data.

Kafka

Apache Kafka is a distributed messaging system that was initially developed at LinkedIn and later became part of the Apache project. Kafka is agile, fast, scalable, and distributed by design.

Big Data Statistics

- 100 Terabytes of data is uploaded to Facebook every day
- Facebook Stores, Processes, and Analyzes more than 30 Petabytes of user generated data
- Twitter generates 12 Terabytes of data every day
- LinkedIn processes and mines Petabytes of user data to power the "People You May Know" feature
- YouTube users upload 48 hours of new video content every minute of the day
- Decoding of the human genome used to take 10 years. Now it can be done in 7 days
- 500+ new websites are created every minute of the day

References :

- https://en.m.wikipedia.org/wiki/Big_data
- https://www.bigdataframework.org/big-data-architecture/

QUIZ (15)

Quiz-1

Find the next in the sequence

5, 21, 69, 213, 645, ?

a. 1935 b. 1815 c. 1941 d. 1290

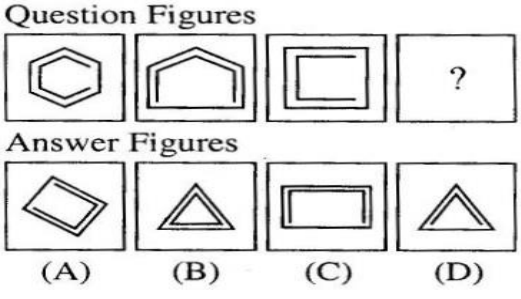
Quiz-2

In a code language QUEEN is written as OVCFL, then KING is written as

a. IJLH b. MKOF c. PHIK d. FOKM

Quiz-3

Out of the given answer figures, which is the correct one to replace the question mark?



a. A b. C c. B d. D

Answer of Last Quiz (14)

Q. 1 Answer: 2 Explanation: 7*6 = 4 2 9*9 = 8 1 5*3 = 1 5 6*2 = 1 2

Q. 2 Answer: 5 Explanation: From the given data,1 rabbit is going towards river not the six elephants. And these 6 elephants saw 2 monkeys are going towards river. Each monkey is holding 1 tortoise. Hence, **number of animals going towards river are 1 rabbit, 2 monkeys and 2 tortoise**= 1 + 2 + 2**= 5.**

Q. 3 Answer: 111,111,111

Explanation: 11 * 11 = 121 and 111 * 111 = 12321. Here, what we observe is that the square of the number follows a pattern: 2 1's gives 121. 3 1's gives 12321.
so to get 12345678987654321 we need to get square of 9 1's(111,111,111).

Student Corner:

WHY WE ARE LUCKY?

GOD HAS GIFTED US A LIFE AS A HUMAN BEING SO WE ARE LUCKY.....

WE HAVE A GOD AS A PARENTS SO WE ARE LUCKY.....

WE CAN DO EVERYTHING WHAT WE CAN WISH,

WE HAVE DERMINATION, WILL POWER, DESTINY SO WE ARE LUCKY...

WE CAN FEEL LOVE, SORROW, SADNESS, UPS AND DOWNS,

TRADEGY BUT..WE CAN OVERCOME FROM THEM SO WE ARE LUCKY.....

WE HAVE FAMILY, RELATIVES, FRIENDS WHOM WE SHARE ALL THE THINGS,

WE ENJOY EVERY MOMENTS OF THE LIFE WITH THEM SO WE ARE LUCKY.....

NOTHING IS IMPOSSIBLE TO A WILLING MIND...

Kum. Jayshree S. Patil
Enrollment No.: 186150307549
Div: 5C
Department of Computer Engineering