SUBJECT CODE: 305 **SUBJECT NAME: BIGDATA**

UNIT 1: INTRODUCTION TO BIGDATA

Big Data?

Big data is the term for 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.

Digital data?

Define: it is defined as the data that is stored on digital format maybeintheformofapicture, documentorvideo etc. it is the data that is not physical but stored in digital form.

Structured Data?

Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. *Example:* Relational data.

Characteristics and Advantages of Structured Data.

Characteristics –

- Data conforms to a data model and has easily identifiable structure
- Data is stored in the form of rows and columns
  Example : Database
- Data is well organised so, Definition, Format and Meaning of data is explicitly known
- Data resides in fixed fields within a record or file
- Similar entities are grouped together to form relations or classes
- Entities in the same group have same attributes
- Easy to access and query, so data can be easily used by other programs
- Data elements are addressable, so efficient to analyse and process

Advantages of Structured Data -

- Structured data have a well-defined structure that helps in easy storage and access of data
- Data can be indexed based on text string as well as attributes. This makes search operation hassle-free
- Data mining is easy i.e. knowledge can be easily extracted from data
- Operations such as Updating and deleting is easy due to well-structured form of data
- Business Intelligence operations such as Data warehousing can be easily undertaken
- Easily scalable in case there is an increment of data
- Ensuring security to data is easy

Sources of Structured Data?

- SQL Databases
- Spreadsheets such as Excel
- OLTP Systems
- Online forms
- Sensors such as GPS or RFID tags
- Network and Web server logs
- Medical devices

Semi-structured data?

Semi-structured data is the data which does not conforms to a data model but has some structure. It lacks a fixed or rigid schema. It is the data that does not reside in a rational database but that have some organisational properties that make it easier to analyse. With some process, we can store them in the relational database.

Characteristics and Advantages of semi-structured Data?

Characteristics –

- Data does not conform to a data model but has some structure.
- Data cannot be stored in the form of rows and columns as in Databases
- Semi-structured data contains tags and elements (Metadata) which is used to group data and describe how the data is stored
- Similar entities are grouped together and organised in a hierarchy
- Entities in the same group may or may not have the same attributes or properties

- Does not contains sufficient metadata which makes automation and management of data difficult
- Size and type of the same attributes in a group may differ
- Due to lack of a well defined structure, it can not used by computer programs easily

Advantages of Semi-structured Data:

- The data is not constrained by a fixed schema
- Flexible i.e. Schema can be easily changed.
- Data is portable
- It is possible to view structured data as semi-structured data
- Its supports users who can not express their need in SQL
- It can deal easily with the heterogeneity of sources.

Sources of Semi-Structured Data?

- E-mails
- XML and other markup languages
- Binary executables
- TCP/IP packets
- Zipped files
- Integration of data from different sources
- Web pages

Unstructured Data?

Unstructured data is the data which does not conforms to a data model and has no easily identifiable structure such that it can not be used by a computer program easily. Unstructured data is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database.

Characteristics of Unstructured Data?

- Data neither conforms to a data model nor has any structure.
- Data can not be stored in the form of rows and columns as in Databases
- Data does not follows any semantic or rules
- Data lacks any particular format or sequence
- Data has no easily identifiable structure
- Due to lack of identifiable structure, it can not used by computer programs easily

Sources of Unstructured Data?

- Web pages
- Images (JPEG, GIF, PNG, etc.)
- Videos
- Memos
- Reports
- Word documents and PowerPoint presentations
- Surveys

advantages and disadvantages of Unstructured Data?

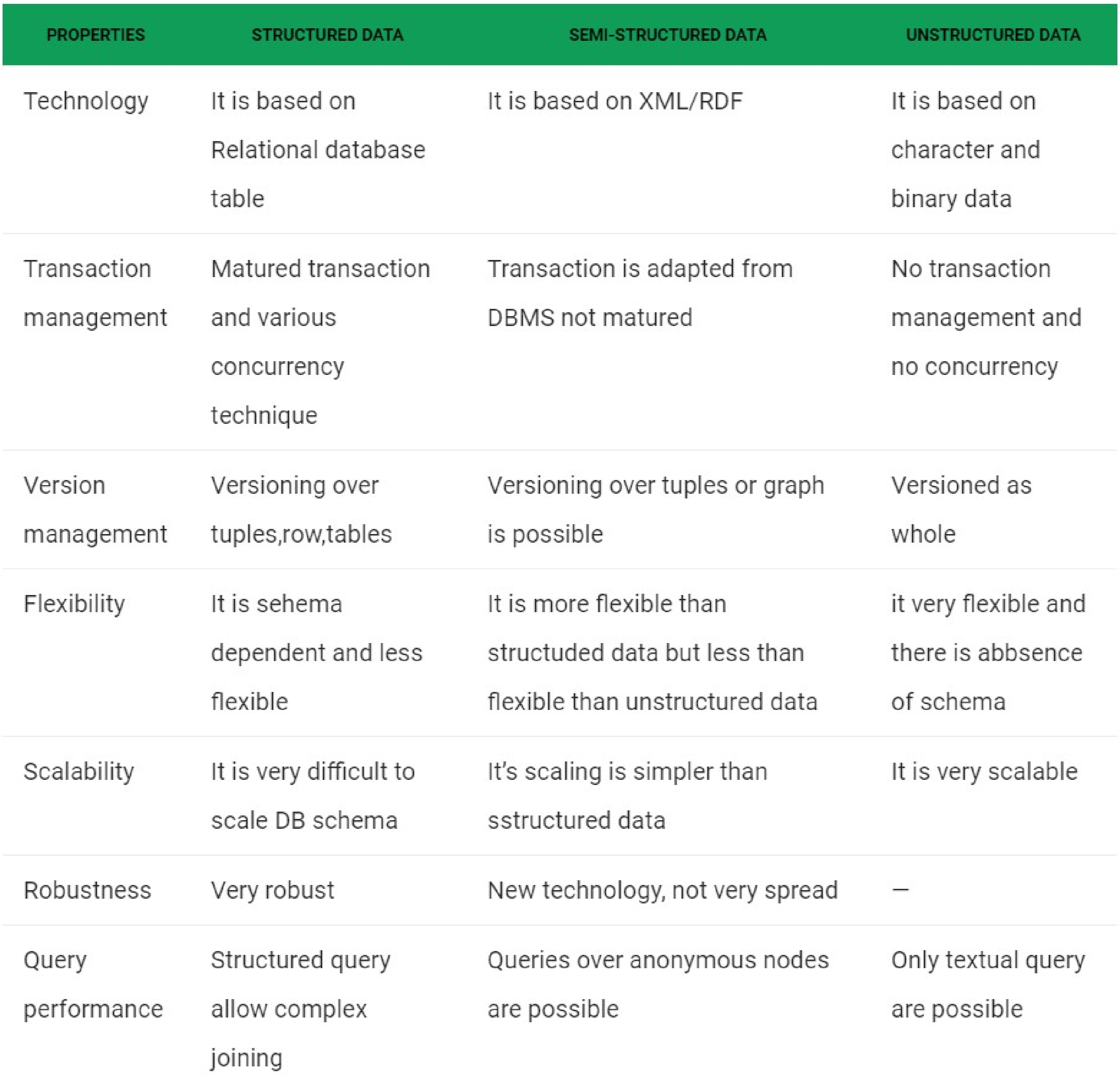
Advantages of Unstructured Data:

- Its supports the data which lacks a proper format or sequence
- The data is not constrained by a fixed schema
- Very Flexible due to absence of schema.
- Data is portable
- It is very scalable
- It can deal easily with the heterogeneity of sources.
- These type of data have a variety of business intelligence and analytics applications.

Disadvantages of Unstructured data:

- It is difficult to store and manage unstructured data due to lack of schema and structure
- Indexing the data is difficult and error prone due to unclear structure and not having pre-defined attributes. Due to which search results are not very accurate.
- Ensuring security to data is difficult task.

Differences between Structured, Semi-structured and Unstructured data.



PROPERTIES	STRUCTURED DATA	SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Technology	It is based on Relational database table	It is based on XML/RDF	It is based on character and binary data
Transaction management	Matured transaction and various concurrency technique	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as whole
Flexibility	It is sehema dependent and less flexible	It is more flexible than structuded data but less than flexible than unstructured data	it very flexible and there is abbsence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than sstructured data	It is very scalable
Robustness	Very robust	New technology, not very spread	–
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual query are possible

BIG DATA ANALYTICS ?

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data.
The analysis of structured data evolves due to the variety and velocity of the data manipulated. Therefore, it is no longer enough to analyze data and produce reports, the wide variety of data means that the systems in place must be capable of assisting in the analysis of data. The analysis consists of automatically determining, within a variety of rapidly changing data, the correlations between the data in order to help in the exploitation of it.

Big Data Anlytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of
integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways.

TYPES OF BIG DATA ANALYTICS ? Explain it's in details.

a) Descriptive Analytics

It consists of asking the question: What is happening?
It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.

b) Diagnostic Analytics

It consists of asking the question: Why did it happen?

Diagnostic analytics looks for the root cause of a problem. It is used to determine why something happened. This type attempts to find and understand the causes of events and behaviors.

c) Predictive Analytics

It consists of asking the question: What is likely to happen?

It uses past data in order to predict the future. It is all about forecasting.

Predictive analytics uses many techniques like data mining and artificial intelligence to analyze current data and make scenarios of what might happen.

d) Prescriptive Analytics

It consists of asking the question: What should be done?

It is dedicated to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.

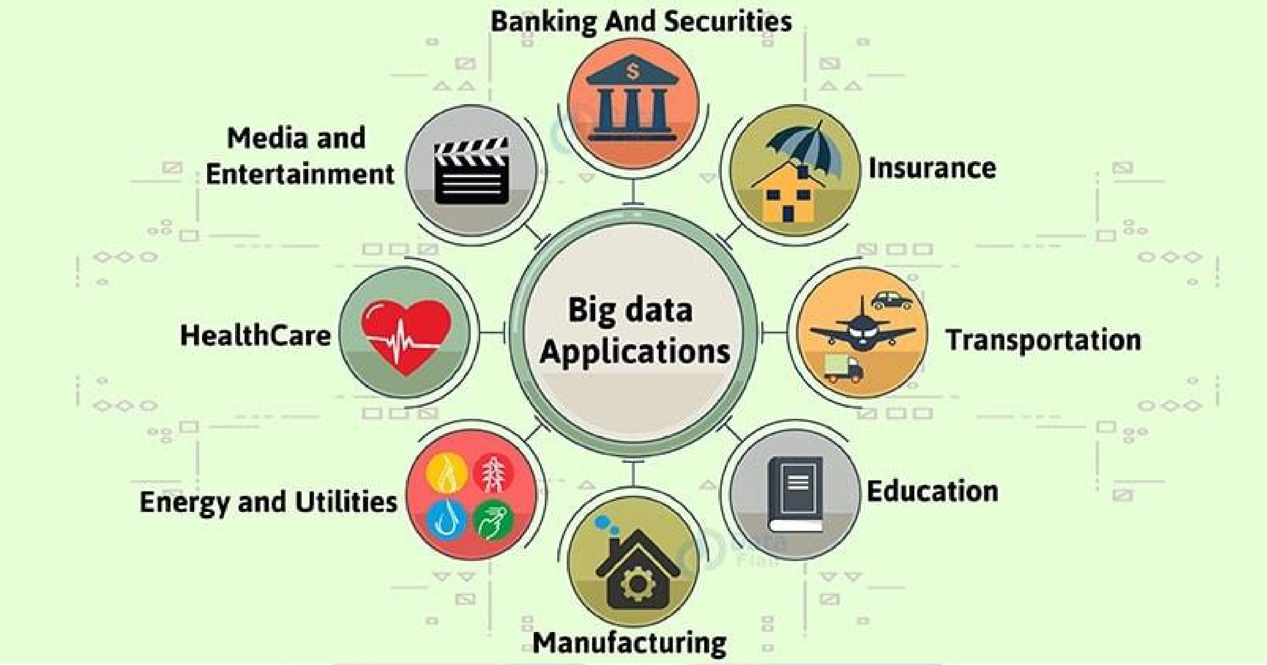
difference between Big Data & Data Analytics?

Think of Big Data like a library that you visit when the information to answer your question is not readily available. Whilst, data analytics is like the book that you pick up and sift through to find answers to your question. This is the basic difference between them.

Data analytics is generally more focused than big data because instead of gathering huge piles of unstructured data, data analysts have a specific goal in mind and sort through relevant data to look for ways to gain support. On the other hand, big data is a collection of a huge volume of data that requires a lot of filtering out to derive useful insights from it.

Another notable difference between the two is that Big data employs complex technological tools like parallel computing and other automation tools to handle the “big data”. Data analytics use predictive and statistical modelling with relatively simple tools.

applications in big Data?



Big data has found many applications in various fields today. The major fields where big data is being used are as follows.

- ## Government

Big data analytics has proven to be very useful in the government sector. Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign. The Indian Government utilizes numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation.

- ## Social Media Analytics

The advent of social media has led to an outburst of big data. Various solutions have been built in order to analyze social media activity like IBM's Cognos Consumer Insights, a point solution running on IBM's BigInsights Big Data platform, can make sense of the chatter. Social media can provide valuable real-time insights into how the market is responding to products and campaigns. With the help of these insights, the companies can adjust their pricing, promotion, and campaign placements accordingly. Before utilizing the big data there needs to be some preprocessing to be done on the big data in order to derive some intelligent and valuable results. Thus to know the consumer mindset the application of intelligent decisions derived from big data is necessary.

- ## Technology

The technological applications of big data comprise of the following companies which deal with huge amounts of data every day and put them to use for business decisions as well. For example, eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay"s 90PB data warehouse. Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005, they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB. Facebook handles 50 billion photos from its user base. Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

- ## Fraud detection

For businesses whose operations involve any type of claims or transaction processing, fraud detection is one of the most compelling Big Data application examples. Historically, fraud detection on the fly has proven an elusive goal. In most cases, fraud is discovered long after the fact, at which point the damage has been done and all that's left is to minimize the harm and adjust policies to prevent it from happening again. Big Data platforms that can analyze claims and transactions in real time, identifying large-scale patterns across many transactions or detecting anomalous behavior from an individual user, can change the fraud detection game.

- ## Call Center Analytics

Now we turn to the customer-facing Big Data application examples, of which call center analytics are particularly powerful. What's going on in a customer's call center is often a great barometer and influencer of market sentiment, but without a Big Data solution, much of the insight that a call center can provide will be overlooked or discovered too late. Big Data solutions can help identify recurring problems or customer and staff behavior patterns on the fly not only by making sense of time/quality resolution metrics but also by capturing and processing call content itself.

- ## Banking

The use of customer data invariably raises privacy issues. By uncovering hidden connections between seemingly unrelated pieces of data, big data analytics could potentially reveal sensitive personal information. Research indicates that 62% of bankers are cautious in their use of big data due to privacy issues. Further, outsourcing of data analysis activities or distribution of customer data across departments for the generation of richer insights also amplifies security risks. Such as customers' earnings, savings, mortgages, and insurance policies ended up in the wrong hands. Such incidents reinforce concerns about data privacy and discourage customers from sharing personal information in exchange for customized offers.

- ## Agriculture

A biotechnology firm uses sensor data to optimize crop efficiency. It plants test crops and runs simulations to measure how plants react to various changes in condition. Its data environment constantly adjusts to changes in the attributes of various data it collects, including temperature, water levels, soil composition, growth, output, and gene sequencing of each plant in the test bed. These simulations allow it to discover the optimal environmental conditions for specific gene types.

- ## Marketing

Marketers have begun to use facial recognition software to learn how well their advertising succeeds or fails at stimulating interest in their products. A recent study published in the Harvard Business Review looked at what kinds of advertisements compelled viewers to continue watching and what turned viewers off. Among their tools was "a system that analyses facial expressions to reveal what viewers are feeling." The research was designed to discover what kinds of promotions induced watchers to share the ads with their social network, helping marketers create ads most likely to "go viral" and improve sales.

- ## Smart Phones

Perhaps more impressive, people now carry facial recognition technology in their pockets. Users of I Phone and Android smartphones have applications at their fingertips that use facial recognition technology for various tasks. For example, Android users with the remember app can snap a photo of someone, then bring up stored information about that person based on their image when their own memory lets them down a potential boon for salespeople.