BIG DATA ANALYTICS

Objective:

To impart knowledge in Fundamentals, Big Data Analytics, Technologies and databases, Hadoop and Map Reduce Fundamentals

Unit I

Introduction to big data: Data, Characteristics of data and Types of digital data: Unstructured, Semi-structured and Structured, Sources of data, Working with unstructured data, Evolution and Definition of big data, Characteristics and Need of big data, Challenges of big data, Data environment versus big data environment

Unit II

Big data analytics: Overview of business intelligence, Data science and Analytics, Meaning and Characteristics of big data analytics, Need of big data analytics, Classification of analytics, Challenges to big data analytics, Importance of big data analytics, Basic terminologies in big data environment

Unit III

Big data technologies and Databases: Introduction to NoSQL, Uses, Features and Types, Need, Advantages, Disadvantages and Application of NoSQL, Overview of NewSQL, Comparing SQL, NoSQL and NewSQL, Introduction to MongoDB and its needs, Characteristics of MongoDB, Introduction of apache cassandra and its needs, Characteristics of Cassandra

Unit IV

Hadoop foundation for analytics: History, Needs, Features, Key advantage and Versions of Hadoop, Essential of Hadoop ecosystems, RDBMS versus Hadoop, Key aspects and Components of Hadoop, Hadoop architectures

Unit V

HadoopMapReduce and YARN framework: Introduction to MapReduce, Processing data with Hadoop using MapReduce, Introduction to YARN, Components, Need and Challenges of YARN, Dissecting YARN, MapReduce application, Data serialization and Working with common serialization formats, Big data serialization formats

Text Book

Seema Acharya and Subhashini Chellappan, “Big Data and Analytics”, Wiley India Pvt. Ltd., 2016

Reference Books

1. 1.“Big Data” by Judith Hurwitz, Alan Nugent, Dr. Fern Halper and Marcia Kaufman, Wiley Publications, 2014.
2. 2.“Big Data Imperatives : Enterprise Big Data Warehouse, BI Implementations and Analytics” by Soumendra Mohanty, Madhu Jagadeesh and Harsha Srivatsa, Apress Media, Springer Science + Business Media New York, 2013
3. “Mining of Massive Datasets”, Anand Rajaraman, Jure Leskovec, Jeffery D. Ullman, Springer, July 2013.
4. “Hadoop: The definitive Guide”, Tom White, O'Reilly Media, 2010.

BIGDATA ANALYTICS

UNIT-I

Big data:

What is Big Data?

Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. For example, the volume of data Facebook or Youtube need require it to collect and manage on a daily basis, can fall under the category of Big Data. However, Big Data is not only about scale and volume, it also involves one or more of the following aspects − Velocity, Variety, Volume, and Complexity.

Characteristics of Big Data

3 'V's of Big Data –

Variety,
Velocity,
Volume.
\

1) Variety

Variety of Big Data refers to structured, unstructured, and semistructured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more.

2) Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

3) Volume

We already know that Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data are stored in data warehouses.

Types of Big Data

Structured

By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. **For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc.,** will be present in an organized manner.

Unstructured

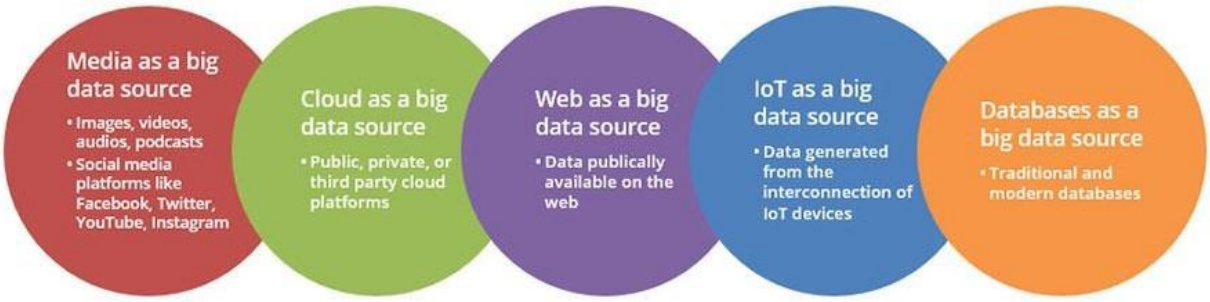
Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data.

Semi-structured

Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

Sources of big data

Voluminous amounts of big data make it crucial for businesses to differentiate, for the purpose of effectiveness, the disparate big data sources available



Media as a big data source

Media is the most popular source of big data, as it provides valuable insights on consumer preferences and changing trends. Since it is self-broadcasted and crosses all physical and demographical barriers, it is the fastest way for businesses to get an in-depth overview of their target audience, draw patterns and conclusions, and enhance their decision-making. Media includes social media and interactive platforms, like Google, Facebook, Twitter, YouTube, Instagram, as well as generic media like images, videos, audios, and podcasts that provide quantitative and qualitative insights on every aspect of user interaction.

Cloud as a big data source

Today, companies have moved ahead of traditional data sources by shifting their data on the cloud. Cloud storage accommodates structured and unstructured data and provides business with real-time information and on-demand insights. The main attribute of cloud computing is its flexibility and scalability. As big data

can be stored and sourced on public or private clouds, via networks and servers, cloud makes for an efficient and economical data source.

The web as a big data source

The public web constitutes big data that is widespread and easily accessible. Data on the Web or 'Internet' is commonly available to individuals and companies alike. Moreover, web services such as Wikipedia provide free and quick informational insights to everyone. The enormity of the Web ensures for its diverse usability and is especially beneficial to start-ups and SME's, as they don't have to wait to develop their own big data infrastructure and repositories before they can leverage big data.

IoT as a big data source

Machine-generated content or data created from IoT constitute a valuable source of big data. This data is usually generated from the sensors that are connected to electronic devices. The sourcing capacity depends on the ability of the sensors to provide real-time accurate information. IoT is now gaining momentum and includes big data generated, not only from computers and smartphones, but also possibly from every device that can emit data. With IoT, data can now be sourced from medical devices, vehicular processes, video games, meters, cameras, household appliances, and the like.

Databases as a big data source

Businesses today prefer to use an amalgamation of traditional and modern databases to acquire relevant big data. This integration paves the way for a hybrid data model and requires low investment and IT infrastructural costs. Furthermore, these databases are deployed for several business intelligence purposes as well. These databases can then provide for the extraction of insights that are used to drive business profits. Popular databases include a variety of data sources, such as MS Access, DB2, Oracle, SQL, and Amazon Simple, among others.

Working with unstructured data

The process of extracting and analyzing data amongst extensive big data sources is a complex process and can be frustrating and time-consuming. These complications can be resolved if organizations encompass all the necessary considerations of big data, take into account relevant data sources, and deploy them in a manner which is well tuned to their organizational goals.

Before the modern day ubiquity of online and mobile applications, databases processed straightforward, structured data. Data models were relatively simple and described a set of relationships between different data types in the database.

Unstructured data, in contrast, refers to data that doesn't fit neatly into the traditional row and column structure of relational databases. Examples of unstructured data include: emails, videos, audio files, web pages, and social media messages. In today's world of Big Data, most of the data that is created is unstructured with some estimates of it being more than 95% of all data generated.

As a result, enterprises are looking to this new generation of databases, known as NoSQL, to address unstructured data. MongoDB stands as a leader in this movement with over 10 million downloads and hundreds of thousands of deployments. As a document database with flexible schema, MongoDB was built specifically to handle unstructured data. MongoDB's flexible data model allows for development without a predefined schema which resonates particularly when most of the data in your system is unstructured.

The Evolution of Big Data

To truly understand the implications of Big Data analytics, one has to reach back into the annals of computing history, specifically business intelligence (BI) and scientific computing. The ideology behind Big Data can most likely be tracked back to the days before the age of computers, when unstructured data were the norm (paper records) and analytics was in its infancy. Perhaps the first Big Data challenge came in the form of the 1880 U.S. census, when the information concerning approximately 50 million people had to be gathered, classified, and reported on.

With the 1880 census, just counting people was not enough information for the U.S. government to work with—particular elements, such as age, sex, occupation, education level, and even the "number of insane people in household," had to be accounted for. That information had intrinsic value to the process, but only if it could be tallied, tabulated, analyzed, and presented. New methods of relating the data to other data collected came into being, such as associating occupations with geographic areas, birth rates with education levels, and countries of origin with skill sets.

The 1880 census truly yielded a mountain of data to deal with, yet only severely limited technology was available to do any of the analytics. The problem of Big Data could not be solved for the 1880 census, so it took over seven years to manually tabulate and report on the data.

With the 1890 census, things began to change, ...

Challenges of Big Data

It must be pretty clear by now that while talking about big data one can't ignore the fact that there are some obvious challenges associated with it. So moving forward in this blog, let's address some of those challenges.

- **Quick Data Growth**

Data growing at such a quick rate is making it a challenge to find insights from it. There is more and more data generated every second from which the data that is actually relevant and useful has to be picked up for further analysis.

- **Storage**

Such large amount of data is difficult to store and manage by organizations without appropriate tools and technologies.

- **Syncing Across Data Sources**

This implies that when organisations import data from different sources the data from one source might not be up to date as compared to the data from another source.

- **Security**

Huge amount of data in organisations can easily become a target for advanced persistent threats, so here lies another challenge for organisations to keep their data secure by proper authentication, data encryption, etc.

- **Unreliable Data**

We can't deny the fact that big data can't be 100 percent accurate. It might contain redundant or incomplete data, along with contradictions.

- **Miscellaneous Challenges**

These are some other challenges that come forward while dealing with big data, like **the integration** of data, **skill** and **talent availability, solution expenses** and **processing a large amount of data** in time and with accuracy so that the data is available for data consumers whenever they need it.

Data Environment versus Big Data Environment

Below are the lists of points, describe the comparisons between Small Data and Big Data.

BasisOf Comparison	**Small Data**	**Big Data**
Definition	Data that is 'small' enough for human comprehension.In a volume and format that makes it accessible, informative and actionable	Data sets that are so large or complex that traditional data processing applications cannot deal with them
Data Source	● Data from traditional enterprise systems like ○ Enterprise resource planning ○ Customer relationship management(CRM) ● Financial Data like general ledger data ● Payment transaction data from website	● Purchase data from point-of-sale ● Clickstream data from websites ● GPS stream data – Mobility data sent to a server ● Social media – Facebook, Twitter
Volume	Most cases in a range of tens or hundreds of GB.Some case few TBs (1 TB=1000 GB)	More than a few Terabytes (TB)
Velocity	● Controlled and steady data flow	● Data can arrive at very fast speeds.

	• Data accumulation is slow	• Enormous data can accumulate within very short periods of time
Variety	Structured data in tabular format with fixed schema and semi-structured data in JSON or XML format	High variety data sets which include Tabular data,Text files, Images, Video, Audio,XML,JSON,Logs,Sensor data etc.
Veracity (Quality of data)	Contains less noise as data collected in a controlled manner.	Usually, the quality of data not guaranteed. Rigorous data validation is required before processing.
Value	Business Intelligence, Analysis, and Reporting	Complex data mining for prediction, recommendation, pattern finding, etc.
Time Variance	Historical data equally valid as data represent solid business interactions	In some cases, data gets older soon(Eg fraud detection).
Data Location	Databases within an enterprise, Local servers, etc.	Mostly in distributed storages on Cloud or in external file systems.
Infrastructure	Predictable resource allocation.Mostly vertically scalable hardware	More agile infrastructure with a horizontally scalable architecture. Load on the system varies a lot.

******