UNIT I INTRODUCTION TO BIG DATA

Introduction to Big Data Platform – Challenges of Conventional Systems - Intelligent data analysis –Nature of Data - Analytic Processes and Tools - Analysis Vs Reporting - Modern Data Analytic Tools- Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error.

ANALYTIC PROCESSES AND TOOLS

Big Data Analytics is a process of examining large and varied data sets, collectively known as big data, to uncover hidden patterns, unknown correlations, market trends, customer preferences and other valuable insights. The importance of Big data Analytics is hard to overstate. It transforms the way businesses and organizations operate, making them more efficient, more informed and more capable of making predictions. Through Big Data Analytics, we can better understand, predict customer behavior, allowing us to tailor our strategies and services to meet their needs. We can also improve our business strategies, making informed decisions that drive growth and innovation.

To harness the power of big data analytics we need to understand the process involved. We need to transform raw, unstructured data into meaningful insights. This is the analytics process of Big Data. The Process involves the following steps:

1. Data Collection- It's about gathering information from various sources, eg. social media, weblogs or IoT devices. The more diverse and vast the data; the better the potential insights. But Quality is needed more than quantity.
2. Data processing - We convert our raw data into a more usable format. We might use algorithms, or even machine learning models to transform and structure the data. It's like turning a jigsaw puzzle into a clear picture.
3. Data cleaning - It acts as detective combing through the data, identifying and rectifying errors, inconsistencies and redundancies. Clean data is like a well-oiled machine ready to deliver accurate results.
4. Data exploration - Cleaned data should be explored clearly. We use statistical methods and visual tools to understand patterns, trends, and potential relationships. Its intrepid explorer, charting the uncharted, seeking hidden treasure in the form of insights.
5. Model building - We use algorithms to create predictive models based on our data. Thus the data become insights and insights become actions, like a craftsman molding raw materials into a masterpiece.
6. Interpretation of results - The output of our models is translated into actionable insights. We use numbers, graphs and charts to inform business decisions.

These steps are iterative i.e. need to revisit the steps based on new data or evolving business needs. Each step is crucial to be accurate; one mis-lead step can lead to misguided insights.

Understanding these processes is essential, but the right tools can make or break the big data analytics journey. The right tools can simplify and enhance your big data analytics process. The top tools that are shaping the world of big data analytics are:

1. Hadoop - open source software framework is a powerhouse for storing and processing large data sets across clusters of computers - It is designed to scale up from a single server to thousands of machines, each offering local computation and storage.
2. Spark - Open source, distributed computing system excels at real-time processing. It's lightning fast and can handle both batch and streaming workloads and it's compatible with hadoop making it a versatile tool in big data analytics.
3. Flink - Stream processing framework provides high throughput, low latency and exactly-once semantics, making it ideal for event-driven applications. It's excellent for real time analytics and complex event processing.
4. Hive - Data warehouse software facilitates reading, writing and managing large datasets residing in distributed storage. It's fantastic for ad hoc querying and analysis of structured and semi-structured data.
5. Tableau - Business Intelligence software excels at data visualization. It helps to turn raw data into easily understandable visuals, making the analysis process more intuitive and accessible.

Each of these tools offers unique features that can propel your big data analytics to new heights.

MODERN DATA ANALYTIC TOOLS

As we're growing with the pace of technology, the demand to track data is increasing rapidly. Today, almost 2.5quintillion bytes of data are generated globally and it's useless until that data is segregated in a proper structure. It has become crucial for businesses to maintain consistency in the business by collecting meaningful data from the market today and for that, all it takes is the right data analytic tool and a professional data analyst to segregate a huge amount of raw data by which then a company can make the right approach.

There are hundreds of data analytics tools out there in the market today but the selection of the right tool will depend upon your business NEED, GOALS, and VARIETY to get business in the right direction. the top 10 analytics tools in big data are:

1. APACHE Hadoop

It's a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel. It can process both structured and unstructured data from one server to multiple computers. Hadoop also offers cross-platform support for its users. Today, it is the best big data analytic tool and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

Features of Apache Hadoop:

- Free to use and offers an efficient storage solution for businesses.
- Offers quick access via HDFS (Hadoop Distributed File System).
- Highly flexible and can be easily implemented with MySQL, and JSON.
- Highly scalable as it can distribute a large amount of data in small segments.
- It works on small commodity hardware like JBOD or a bunch of disks.

2. Cassandra

APACHE Cassandra is an open-source NoSQL distributed database that is used to fetch large amounts of data. It's one of the most popular tools for data analytics and has been praised by many tech companies due to its high scalability and availability without compromising speed and performance. It is capable of delivering thousands of operations every second and can handle petabytes of resources with almost zero downtime. It was created by Facebook back in 2008 and was published publicly.

Features of APACHE Cassandra:

- Data Storage Flexibility: It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
- Data Distribution System: Easy to distribute data with the help of replicating data on multiple data centers.
- Fast Processing: Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.
- Fault-tolerance: The moment, if any node fails, it will be replaced without any delay.

3. Qubole

It's an open-source big data tool that helps in fetching data in a value of chain using ad-hoc analysis in machine learning. Qubole is a data lake platform that offers end-to-end service with reduced time and effort which are required in moving data pipelines. It is capable of configuring multi-cloud services such as AWS, Azure, and Google Cloud. Besides, it also helps in lowering the cost of cloud computing by 50%.

Features of Qubole:

- Supports ETL process: It allows companies to migrate data from multiple sources in one place.
- Real-time Insight: It monitors user's systems and allows them to view real-time insights
- Predictive Analysis: Qubole offers predictive analysis so that companies can take actions accordingly for targeting more acquisitions.
- Advanced Security System: To protect users' data in the cloud, Qubole uses an advanced security system and also ensures to protect any future breaches. Besides, it also allows encrypting cloud data from any potential threat.

4. Xplenty

It is a data analytic tool for building a data pipeline by using minimal codes in it. It offers a wide range of solutions for sales, marketing, and support. With the help of its interactive graphical interface, it provides solutions for ETL, ELT, etc. The best part of using Xplenty is its low investment in hardware & software and its offers support via email, chat, telephonic and virtual meetings. Xplenty is a platform to process data for analytics over the cloud and segregates all the data together.

Features of Xplenty:

- Rest API: A user can possibly do anything by implementing Rest API
- Flexibility: Data can be sent, and pulled to databases, warehouses, and salesforce.
- Data Security: It offers SSL/TSL encryption and the platform is capable of verifying algorithms and certificates regularly.
- Deployment: It offers integration apps for both cloud & in-house and supports deployment to integrate apps over the cloud.

5. Spark

APACHE Spark is another framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is capable of handling multi-petabytes of data as well. Recently, Spark made a record of processing 100 terabytes of data in just 23 minutes which broke the previous world record of Hadoop (71 minutes). This is the reason why big tech giants are moving towards spark now and is highly suitable for ML and AI today.

Features of APACHE Spark:

- Ease of use: It allows users to run in their preferred language. (JAVA, Python, etc.)
- Real-time Processing: Spark can handle real-time streaming via Spark Streaming
- Flexible: It can run on, Mesos, Kubernetes, or the cloud.

6. Mongo DB

Came in limelight in 2010, is a free, open-source platform and a document-oriented (NoSQL) database that is used to store a high volume of data. It uses collections and documents for storage and its document consists of key-value pairs which are considered a basic unit of Mongo DB. It is so popular among developers due to its availability for multi-programming languages such as Python, Jscript, and Ruby.

Features of Mongo DB:

- Written in C++: It's a schema-less DB and can hold varieties of documents inside.
- Simplifies Stack: With the help of mongo, a user can easily store files without any disturbance in the stack.
- Master-Slave Replication: It can write/read data from the master and can be called back for backup.

7. Apache Storm

A storm is a robust, user-friendly tool used for data analytics, especially in small companies. The best part about the storm is that it has no language barrier (programming) in it and can support any of them. It was designed to handle a pool of large data in fault-tolerance and horizontally scalable methods. When we talk about real-time data processing, Storm leads the chart because of its distributed real-time big data processing system, due to which today many tech giants are using APACHE Storm in their system. Some of the most notable names are Twitter, Zendesk, NaviSite, etc.

Features of Storm:

- Data Processing: Storm process the data even if the node gets disconnected
- Highly Scalable: It keeps the momentum of performance even if the load increases
- Fast: The speed of APACHE Storm is impeccable and can process up to 1 million messages of 100 bytes on a single node.

8. SAS

Today it is one of the best tools for creating statistical modeling used by data analysts. By using SAS, a data scientist can mine, manage, extract or update data in different variants from different sources. Statistical Analytical System or SAS allows a user to access the data in any format (SAS tables or Excel worksheets). Besides that it also offers a cloud platform for business analytics called SAS Viya and also to get a strong grip on AI & ML, they have introduced new tools and products.

Features of SAS:

- Flexible Programming Language: It offers easy-to-learn syntax and has also vast libraries which make it suitable for non-programmers
- Vast Data Format: It provides support for many programming languages which also include SQL and carries the ability to read data from any format.
- Encryption: It provides end-to-end security with a feature called SAS/SECURE.

9. Data Pine

Datapine is an analytical tool used for BI and was founded back in 2012 (Berlin, Germany). In a short period of time, it has gained much popularity in a number of countries and it’s mainly used for data extraction (for small-medium companies fetching data for close monitoring). With the help of its enhanced UI design, anyone can visit and check the data as per their requirement and offer in 4 different price brackets, starting from $249 per month. They do offer dashboards by functions, industry, and platform.

Features of Datapine:

- Automation: To cut down the manual chase, datapine offers a wide array of AI assistant and BI tools.

- Predictive Tool: datapine provides forecasting/predictive analytics by using historical and current data, it derives the future outcome.
- Add on: It also offers intuitive widgets, visual analytics & discovery, ad hoc reporting, etc.

10. Rapid Miner

It's a fully automated visual workflow design tool used for data analytics. It's a no-code platform and users aren't required to code for segregating data. Today, it is being heavily used in many industries such as ed-tech, training, research, etc. Though it's an open-source platform but has a limitation of adding 10000 data rows and a single logical processor. With the help of Rapid Miner, one can easily deploy their ML models to the web or mobile (only when the user interface is ready to collect real-time figures).

Features of Rapid Miner:

- Accessibility: It allows users to access 40+ types of files (SAS, ARFF, etc.) via URL
- Storage: Users can access cloud storage facilities such as AWS and dropbox
- Data validation: Rapid miner enables the visual display of multiple results in history for better evaluation.
