BIGDATA ANALYTICS

UNIT-II

Bigdata Anlytics:

Overview of Business Intelligence:

Business Intelligence (BI) applications are decision support tools that enable real-time, interactive access to and analysis of mission-critical corporate information. BI applications bridge the gaps between information silos in an organization. Sophisticated analytical capabilities have access to such corporate information resources as data warehouses, transaction processing applications, and enterprise applications like Enterprise Resource Planning (ERP). BI enables users to access and leverage vast amounts of data, providing valuable insight into potential opportunities and areas for business process refinement.

BI applications can be classified as follows:

- Personalized Dashboards for Process Monitoring and Highlighting Exceptions
- Decision Support with Drill-Down and "What-If" Analysis
- Data-Mining to Understand and Discover Patterns and Behaviors
- Automated Agents to Drive Rule-Based Business Strategy via Integrated Processes

Investments made in an EPM (Enterprise Process Management) implementation can be very expensive, so it is imperative that every asset is leveraged. Youngsoft's team of experienced professionals provide a wide range of services across the suite of EPM products, consistently delivering solutions that reduce costs, increase profits, and improve overall efficiency.

Benefits:

- Cost Reduction during the Implementation Process
- Retained Knowledge after Completion of the Implementation
- End User Step by Step Training during the Implementation Process
- Providing Tier-One Production Support during Post-Implementation Process
- Shadowing the Implementation

What is Data Science?

Data Science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing, and aligning the data.

In simple terms, it is the umbrella of techniques used when trying to extract insights and information from data.

Applications of Data Science

- **Internet Search**

  Search engines make use of data science algorithms to deliver the best results for search queries in a fraction of seconds.

- **Digital Advertisements**

  The entire digital marketing spectrum uses the data science algorithms - from display banners to digital billboards. This is the mean reason for digital ads getting higher CTR than traditional advertisements.

- **Recommender Systems**

  The recommender systems not only make it easy to find relevant products from billions of products available but also adds a lot to user-experience. A lot of companies use this system to promote their products and suggestions in accordance with the user's demands and relevance of information. The recommendations are based on the user's previous search results.

Need of Bigdata Analytics

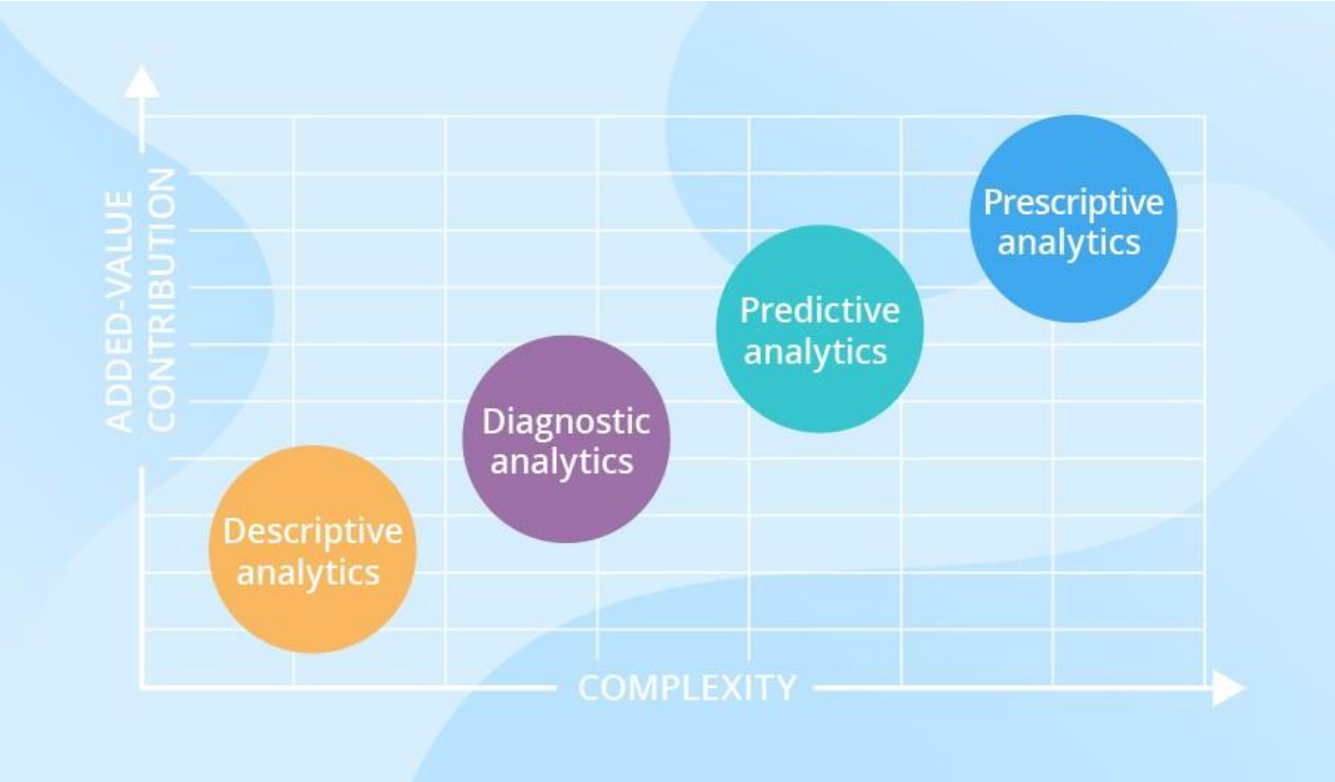
Why is big data analytics important?

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. In his report *Big Data in Big Companies*, IIA Director of Research Tom Davenport interviewed more than 50 businesses to understand how they used big data. He found they got value in the following ways:

1. **Cost reduction.** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.
3. **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Types of data analytics

There are 4 different types of analytics.



Descriptive analytics

Descriptive analytics answers the question of *what happened*. Let us bring an example from ScienceSoft's practice: having analyzed monthly revenue and income per product group, and the total quantity of metal parts produced per month, a manufacturer was able to answer a series of 'what happened' questions and decide on focus product categories.

Descriptive analytics juggles raw data from multiple data sources to give valuable insights into the past. However, these findings simply signal that something is wrong or right, without explaining why. For this reason, our data consultants don't recommend highly data-driven companies to settle for descriptive analytics only, they'd rather combine it with other types of data analytics.

Diagnostic analytics

At this stage, historical data can be measured against other data to answer the question of *why something happened*. For example, you can check ScienceSoft's BI demo to see how a retailer can drill the sales and

gross profit down to categories to find out why they missed their net profit target. Another flashback to our data analytics projects: in the healthcare industry, customer segmentation coupled with several filters applied (like diagnoses and prescribed medications) allowed identifying the influence of medications.

Diagnostic analytics gives in-depth insights into a particular problem. At the same time, a company should have detailed information at their disposal, otherwise, data collection may turn out to be individual for every issue and time-consuming.

Predictive analytics

Predictive analytics tells *what is likely to happen.* It uses the findings of descriptive and diagnostic analytics to detect clusters and exceptions, and to predict future trends, which makes it a valuable tool for forecasting. Check ScienceSoft's case study to get details on how advanced data analytics allowed a leading FMCG company to predict what they could expect after changing brand positioning.

Predictive analytics belongs to advanced analytics types and brings many advantages like sophisticated analysis based on machine or deep learning and proactive approach that predictions enable. However, our data consultants state it clearly: forecasting is just an estimate, the accuracy of which highly depends on data quality and stability of the situation, so it requires careful treatment and continuous optimization.

Prescriptive analytics

The purpose of prescriptive analytics is to literally prescribe *what action to take* to eliminate a future problem or take full advantage of a promising trend. An example of prescriptive analytics from our project portfolio: a multinational company was able to identify opportunities for repeat purchases based on customer analytics and sales history.

Prescriptive analytics uses advanced tools and technologies, like machine learning, business rules and algorithms, which makes it sophisticated to implement and manage. Besides, this state-of-the-art type of data analytics requires not only historical internal data but also external information due to the nature of algorithms it's based on. That is why, before deciding to adopt prescriptive analytics, ScienceSoft strongly recommends weighing the required efforts against an expected added value.

Big Data Analytics Challenges

Need For Synchronization Across Disparate Data Sources

As data sets are becoming bigger and more diverse, there is a big challenge to incorporate them into an analytical platform. If this is overlooked, it will create gaps and lead to wrong messages and insights.

2. Acute Shortage Of Professionals Who Understand Big Data Analysis

The analysis of data is important to make this voluminous amount of data being produced in every minute, useful. With the exponential rise of data, a huge demand for big data scientists and Big Data analysts has been created in the market. It is important for business organizations to hire a data scientist having skills that are varied as the job of a data scientist is multidisciplinary. Another major challenge faced by businesses is

the shortage of professionals who understand Big Data analysis. There is a sharp shortage of data scientists in comparison to the massive amount of data being produced.

3. Getting Meaningful Insights Through The Use Of Big Data Analytics

It is imperative for business organizations to gain important insights from Big Data analytics, and also it is important that only the relevant department has access to this information. A big challenge faced by the companies in the Big Data analytics is mending this wide gap in an effective manner.

4. Getting Voluminous Data Into The Big Data Platform

It is hardly surprising that data is growing with every passing day. This simply indicates that business organizations need to handle a large amount of data on daily basis. The amount and variety of data available these days can overwhelm any data engineer and that is why it is considered vital to make data accessibility easy and convenient for brand owners and managers.

5. Uncertainty Of Data Management Landscape

With the rise of Big Data, new technologies and companies are being developed every day. However, a big challenge faced by the companies in the Big Data analytics is to find out which technology will be best suited to them without the introduction of new problems and potential risks.

6. Data Storage And Quality

Business organizations are growing at a rapid pace. With the tremendous growth of the companies and large business organizations, increases the amount of data produced. The storage of this massive amount of data is becoming a real challenge for everyone. Popular data storage options like data lakes/ warehouses are commonly used to gather and store large quantities of unstructured and structured data in its native format. The real problem arises when a data lakes/ warehouse try to combine unstructured and inconsistent data from diverse sources, it encounters errors. Missing data, inconsistent data, logic conflicts, and duplicates data all result in data quality challenges.

7. Security And Privacy Of Data

Once business enterprises discover how to use Big Data, it brings them a wide range of possibilities and opportunities. However, it also involves the potential risks associated with big data when it comes to the privacy and the security of the data. The Big Data tools used for analysis and storage utilizes the data disparate sources. This eventually leads to a high risk of exposure of the data, making it vulnerable. Thus, the rise of voluminous amount of data increases privacy and security concerns.

The Importance of Big Data Analytics

Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including:

- New revenue opportunities
- More effective marketing
- Better customer service
- Improved operational efficiency
- Competitive advantages over rivals

Big data analytics applications enable big data analysts, data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional BI and analytics programs. This encompasses a mix of semi-structured and unstructured data -- for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the internet of things (IoT).

Basic Terminologies in big data environment

we will discuss the terminology related to Big Data ecosystem. This will give you a complete understanding of Big Data and its terms.

Over time, Hadoop has become the nucleus of the Big Data ecosystem, where many new technologies have emerged and have got integrated with Hadoop. So it's important that, first, we understand and appreciate the nucleus of modern Big Data architecture.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers, using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Components of the Hadoop Ecosystem

Let's begin by looking at some of the components of the Hadoop ecosystem:

Hadoop Distributed File System (HDFS™):

This is a distributed file system that provides high-throughput access to application data. Data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout the cluster. In this method, the map and reduce functions can be executed on smaller subsets of your larger data sets, and this provides the scalability needed for Big Data processing.

M apReduce:

MapReduce is a programming model specifically implemented for processing large data sets on Hadoop cluster. This is the core component of the Hadoop framework, and it is the only execution engine available for Hadoop 1.0.

The MapReduce framework consists of two parts:

1. A function called 'Map', which allows different points in the distributed cluster to distribute their work.

2. A function called 'Reduce', which is designed to reduce the final form of the clusters' results into one output.

The main advantage of the MapReduce framework is its fault tolerance, where periodic reports from each node in the cluster are expected as soon as the work is completed.

The MapReduce framework is inspired by the 'Map' and 'Reduce' functions used in functional programming. The computational processing occurs on data stored in a file system or within a database, which takes a set of input key values and produces a set of output key values.

Each day, numerous MapReduce programs and MapReduce jobs are executed on Google's clusters. Programs are automatically parallelized and executed on a large cluster of commodity machines.

Map Reduce is used in distributed grep, distributed sort, Web link-graph reversal, Web access log stats, document clustering, Machine Learning and statistical machine translation.

Pig:

Pig is a data flow language that allows users to write complex MapReduce operations in simple scripting language. Then Pig then transforms those scripts into a MapReduce job.

Hive:

Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism for querying the data using a SQL-like language called HiveQL. At the same time, this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.

Sqoop:

Enterprises that use Hadoop often find it necessary to transfer some of their data from traditional relational database management systems (RDBMSs) to the Hadoop ecosystem.

Sqoop, an integral part of Hadoop, can perform this transfer in an automated fashion. Moreover, the data imported into Hadoop can be transformed with MapReduce before exporting them back to the RDBMS. Sqoop can also generate Java classes for programmatically interacting with imported data. Sqoop uses a connector-based architecture that allows it to use plugins to connect with external databases.

Flume:

Flume is a service for streaming logs into Hadoop. Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

Storm:

Storm is a distributed, real-time computation system for processing large volumes of high-velocity data. Storm is extremely fast and can process over a million records per second per node on a cluster of modest size. Enterprises harness this speed and combine it with other data-access applications in Hadoop to prevent undesirable events or to optimize positive outcomes.

Kafka:

Apache Kafka supports a wide range of use cases such as a general-purpose messaging system for scenarios where high throughput, reliable delivery, and horizontal scalability are important. Apache Storm and Apache HBase both work very well in combination with Kafka.

Oozie:

Oozie is a workflow scheduler system to manage Apache Hadoop jobs. The Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions, whereas the Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability.

Oozie is integrated with the rest of the Hadoop stack and supports several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system-specific jobs (such as Java programs and shell scripts). Oozie is a scalable, reliable and extensible system.

Spark:

Apache Spark is a fast, in-memory data processing engine for distributed computing clusters like Hadoop. It runs on top of existing Hadoop clusters and accesses the Hadoop data store (HDFS).

Spark can be integrated with Hadoop's 2 YARN architecture, but cannot be used with Hadoop 1.0.

Apache Solr:

Apache Solr is a fast, open-source Java search server. Solr enables you to easily create search engines that search websites, databases, and files for Big Data

Apache Yarn:

Apache Hadoop YARN (Yet Another Resource Negotiator) is a cluster management technology. YARN is one of the key features in the second-generation Hadoop 2 version of the Apache Software Foundation's open-source distributed processing framework. Originally described by Apache as a redesigned resource manager, YARN is now characterized as a large-scale, distributed operating system for big data applications.

Tez:

Tez is an execution engine for Hadoop that allows jobs to meet the demands for fast response times and extreme throughput at petabyte scale. Tez represents computations as a dataflow graphs and can be used with Hadoop 2 YARN.

Apache Drill:

Apache Drill is an open-source, low-latency query engine for Hadoop that delivers secure, interactive SQL analytics at petabyte scale. With the ability to discover schemas on the go, Drill is a pioneer in delivering self-service data exploration capabilities on data stored in multiple formats in files or NoSQL databases. By adhering to ANSI SQL standards, Drill does not require a learning curve and integrates seamlessly with visualization tools.

Apache Phoenix:

Apache Phoenix takes your SQL query, compiles it into a series of HBase scans, and co-ordinates the running of those scans to produce a regular JDBC result set. Apache Phoenix enables OLTP and operational analytics in Hadoop for low-latency applications by combining the best of both worlds. Apache Phoenix is fully integrated with other Hadoop products such as Spark, Hive, Pig, Flume, and Map Reduce.

Cloud Computing:

Cloud Computing is a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. Cloud Computing is comparable to grid computing, a type of computing where the unused processing cycles of all computers in a network are harnessed to solve problems that are too processor-intensive for any single machine.

In Cloud Computing, the word cloud (also phrased as “the cloud”) is used as a metaphor for the Internet, hence the phrase cloud computing means “a type of Internet-based computing” in which different services such as servers, storage and applications are delivered to an organization’s computers and devices via the Internet.

NoSQL:

The NoSQL database, also called Not Only SQL, is an approach to data management and database design that’s useful for very large sets of distributed data. This database system is non-relational, distributed, open-

source and horizontally scalable. NoSQL seeks to solve the scalability and big-data performance issues that relational databases weren't designed to address.

Apache Cassandra:

Apache Cassandra is an open-source distributed database system designed for storing and managing large amounts of data across commodity servers. Cassandra can serve as both a real-time operational data store for online transactional applications and a read-intensive database for large-scale business intelligence (BI) systems.

SimpleDB:

Amazon Simple Database Service (SimpleDB), also known as a key value data store, is a highly available and flexible non-relational database that allows developers to request and store data, with minimal database management and administrative responsibility.

This service offers simplified access to a data store and query functions that let users instantly add data and effortlessly recover or edit that data.

SimpleDB is best used by customers who have a relatively simple data requirement, like data storage. For example, a business might use cookies to track visitors that visit its company website. Some applications might read the cookies to get the visitor's identifier and look up the feeds they're interested in. Amazon SimpleDB gives users the option to store tens of attributes for a million customers, but not thousands of attributes for a single customer.

Google BigTable:

Google's BigTable is a distributed, column-oriented data store created by Google Inc. to handle very large amounts of structured data associated with the company's Internet search and Web services operations.

BigTable was designed to support applications requiring massive scalability; from its first iteration, the technology was intended to be used with petabytes of data. The database was designed to be deployed on clustered systems and uses a simple data model that Google has described as "a sparse, distributed, persistent multidimensional sorted map." Data is assembled in order by row key, and indexing of the map is arranged according to row, column keys, and timestamps. Here, compression algorithms help achieve high capacity.

MongoDB:

MongoDB is a cross-platform, document-oriented database. Classified as a NoSQL database, MongoDB shuns the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster.

MongoDB is developed by MongoDB Inc. and is published as free and open-source software under a combination of the GNU Affero General Public License and the Apache License. As of July 2015, MongoDB is the fourth most popular type of database management system, and the most popular for document stores.

HBase:

Apache HBase (Hadoop DataBase) is an open-source NoSQL database that runs on the top of the database and provides real-time read/write access to those large data sets.

HBase scales linearly to handle huge data sets with billions of rows and millions of columns, and it easily combines data sources that use a wide variety of different structures and schema. HBase is natively integrated with Hadoop and works seamlessly alongside other data access engines through YARN.

Neo4j:

Neo4j is a graph database management system developed by Neo Technology, Inc. Neo4j is described by its developers as an ACID-compliant transactional database with native graph storage and processing. According to db-engines.com, Neo4j is the most popular graph database.

Couch DB:

CouchDB is a database that completely embraces the web. It stores your data with JSON documents. It accesses your documents and queries your indexes with your web browser, via HTTP. It indexes, combines, and transforms your documents with JavaScript.

CouchDB works well with modern web and mobile apps. You can even serve web apps directly out of CouchDB. You can distribute your data, or your apps, efficiently using CouchDB's incremental replication. CouchDB supports master-master setups with automatic conflict detection.

****