19IT601 - DATA SCIENCE & ANALYTICS

UNIT-I - INTRODUCTION TO DATA SCIENCE AND BIG DATA

Data Science – Fundamentals and Components – Data Scientist – Terminologies Used in Big Data Environments – Types of Digital Data – Classification of Digital Data – Introduction to Big Data – Characteristics of Data – Evolution of Big Data – Big Data Analytics – Classification of Analytics – Top Challenges Facing Big Data – Importance of Big Data Analytics – Data Analytics Tools.

Data Science – Fundamentals and Components

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

Data science uses complex machine learning algorithms to build predictive models.

Data science encompasses

- preparing data for analysis and processing,
- performing advanced data analysis, and
- presenting the results to reveal patterns and
- enable stakeholders to draw informed conclusions

- Data preparation can involve cleansing, aggregating, and manipulating it to be ready for specific types of processing.
- Analysis requires the development and use of algorithms, analytics and AI models.
- It's driven by software that combs through data to find patterns within to transform these patterns into predictions that support business decision-making.
- And the results should be shared through the skillful use of data visualization tools that make it possible for anyone to see the patterns and understand trends.

Data Science Life Cycle

Data Science encompasses the following phases

- Capture
- Prepare and Maintain
- Preprocess or Process

- Analyze
- Communicate

Capture: This is the gathering of raw structured and unstructured data from all relevant sources via just about any method—from manual entry and web scraping to capturing data from systems and devices in real time.

Prepare and maintain: This involves putting the raw data into a consistent format for analytics or machine learning or deep learning models. This can include everything from cleansing, deduplicating, and reformatting the data, to using ETL (extract, transform, load) or other data integration technologies to combine the data into a data warehouse, data lake, or other unified store for analysis.

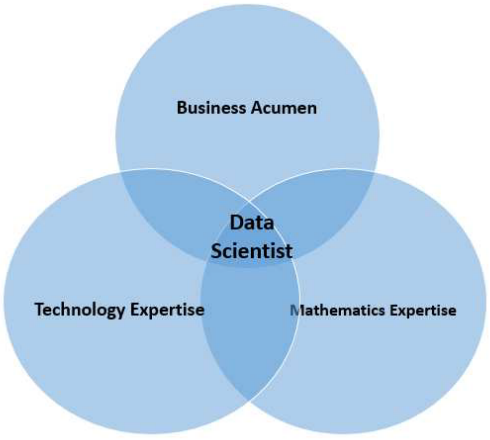
Preprocess or process: Here, data scientists examine biases, patterns, ranges, and distributions of values within the data to determine the data's suitability for use with predictive analytics, machine learning, and/or deep learning algorithms (or other analytical methods).

Analyze: This is where the discovery happens—where data scientists perform statistical analysis, predictive analytics, regression, machine learning and deep learning algorithms, and more to extract insights from the prepared data

Communicate: Finally, the insights are presented as reports, charts, and other data visualizations that make the insights—and their impact on the business—easier for decision-makers to understand.

Data Scientist

A data scientist analyzes business data to extract meaningful insights. In other words, a data scientist solves business problems through a series of steps



Business Acumen Skills

A data scientist should have business acumen skills to counter the pressure of business:

- Understanding of domain
- Business strategy

- Problem solving
- Communication
- Presentation
- Inquisitiveness

Technology Expertise Skills

A data scientist should be technology expert to convert the business into business logic:

- Good database knowledge such as RDBMS.
- Good NoSQL database knowledge such as MongoDB, Cassandra, HBase, etc.
- Programming languages such as Java, Python, etc.
- Open-source tools such as Hadoop, R.
- Datawarehousing, Datamining.
- Visualization such as Tableau, Flare, Google visualization APIs, etc.

Mathematics Expertise Skills

A data scientist should be mathematics expert to formulize and analyze data:

- Mathematics.
- Statistics.
- Artificial Intelligence (AI).
- Machine learning.
- Pattern recognition.
- Natural Language Processing

What Does a Data Scientist Do?

- A data scientist analyzes business data to extract meaningful insights. In other words, a data scientist solves business problems through a series of steps, including:
- Before tackling the data collection and analysis, the data scientist determines the problem by asking the right questions and gaining understanding.
- The data scientist then determines the correct set of variables and data sets.
- The data scientist gathers structured and unstructured data from many disparate sources—enterprise data, public data, etc.
- Once the data is collected, the data scientist processes the raw data and converts it into a format suitable for analysis. This involves cleaning and validating the data to guarantee uniformity, completeness, and accuracy.
- After the data has been rendered into a usable form, it's fed into the analytic system—ML algorithm or a statistical model.
- This is where the data scientists analyze and identify patterns and trends. When the data has been completely rendered, the data scientist interprets the data to find opportunities and solutions.

- The data scientists finish the task by preparing the results and insights to share with the appropriate stakeholders and communicating the results.

Terminologies Used in Big Data Environments

1. In-Memory Analytics
2. In-Database Processing
3. Massively Parallel Processing
4. Parallel System
5. Distributed System
6. Shared Nothing Architecture

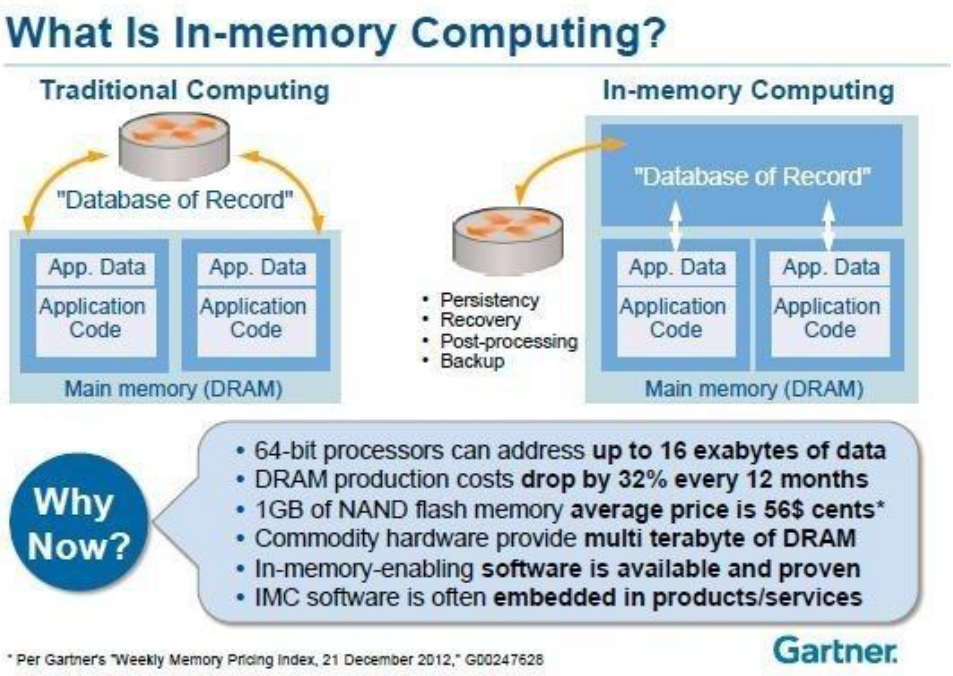
In-Memory Analytics

Data access from non-volatile storage such as hard disk is a slow process. One way to combat this challenge is to pre-process and store data (cubes, aggregate tables, query sets, etc.) so that the CPU has to fetch a small subset of records.

But this requires thinking in advance. This problem has been addressed using in-memory analytics. Here all the relevant data is stored in RAM.

The advantage is

- faster access,
- rapid deployment,
- better insights, and
- minimal IT involvement.



In-Database Processing

In-database processing is also called as in-database analytics. It works by fusing data warehouses with analytical systems.

Typically the data from various enterprise OLTP systems after cleaning up (de-duplication, scrubbing, etc.) through the process of ETL is stored in the Enterprise Data Warehouse (EDW) or data marts.

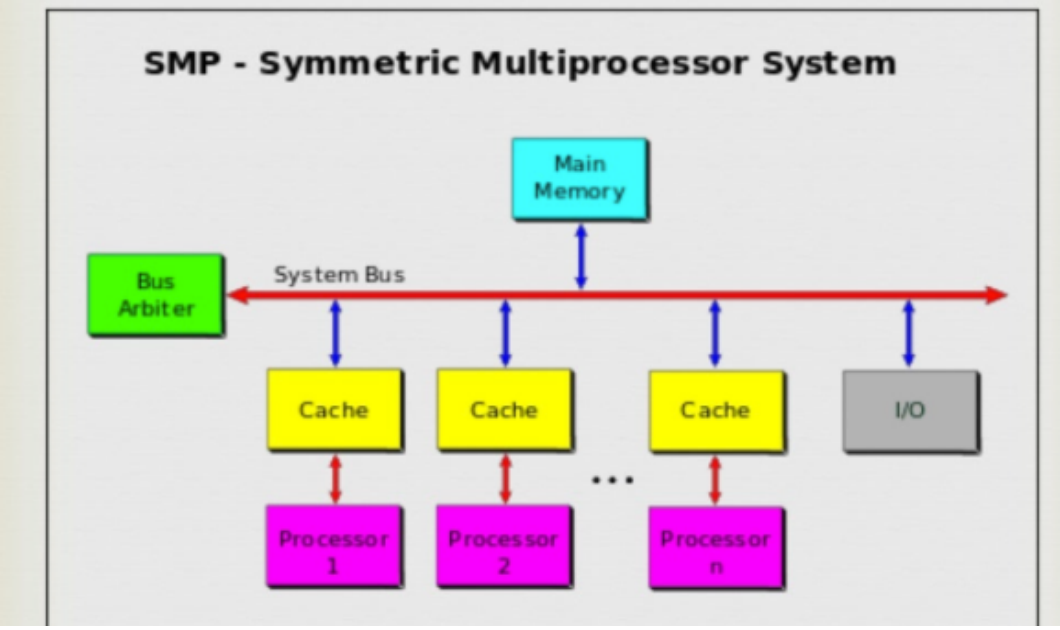
With in-database processing, the database program itself can run the computations eliminating the need for export and thereby saving on time.

Symmetric Multiprocessor System

In SMP, there is a single common main memory that is shared by two or more identical processors.

The processors have full access to all I/O devices and are controlled by a single operating system instance.

SMP are tightly coupled multiprocessor systems. Each processor has its own high-speed memory, called cache memory and are connected using a system bus.

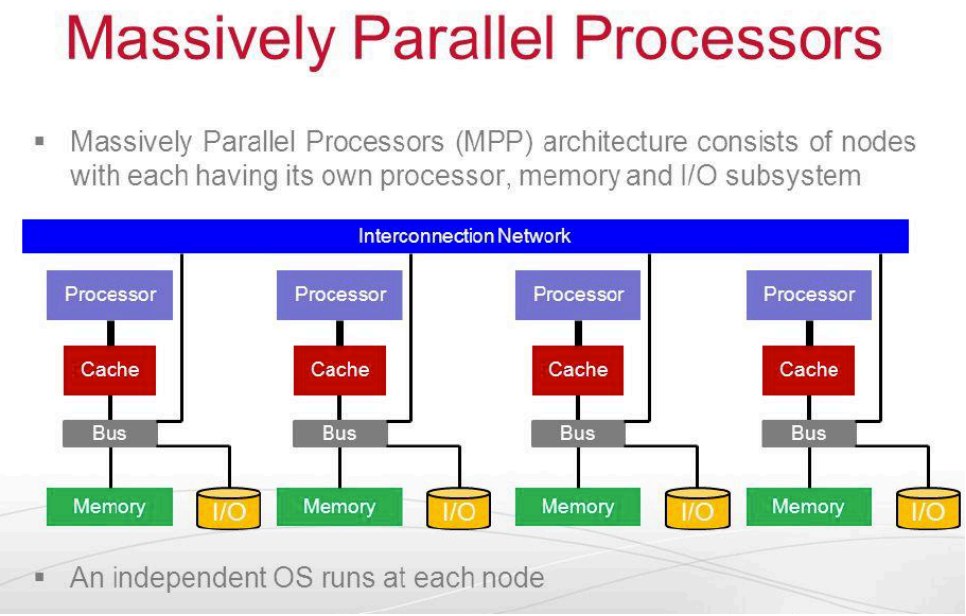


Massively Parallel Processing

MPP refers to the coordinated processing of programs by a number of processors working parallel.

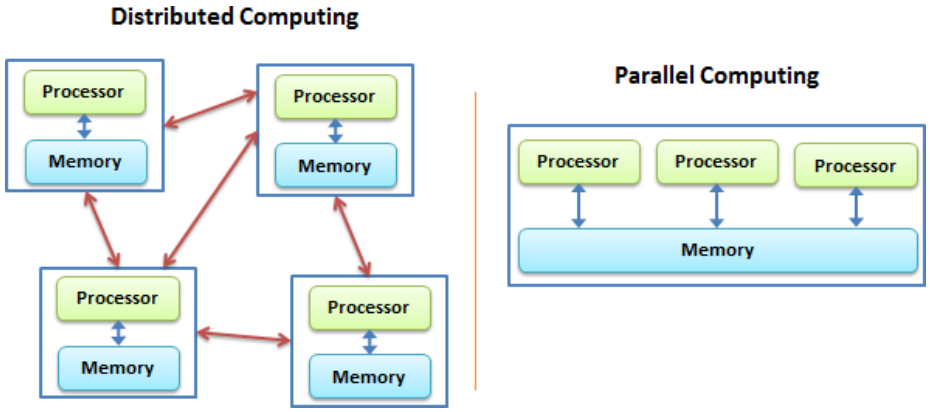
The processors, each have their own operating systems and dedicated memory.

They work on different parts of the same program. The MPP processors communicate using some sort of messaging interface.



Parallel and Distributed System

- A parallel database system is a tightly coupled system. The processor, co-operate for query processing. The user is unaware of the parallelism.
- Distributed database systems are known to be loosely coupled and are composed by individual machines. Each of the machines can run their individual application and serve their own respective users. The data is usually distributed across several machines.



Shared Nothing Architecture

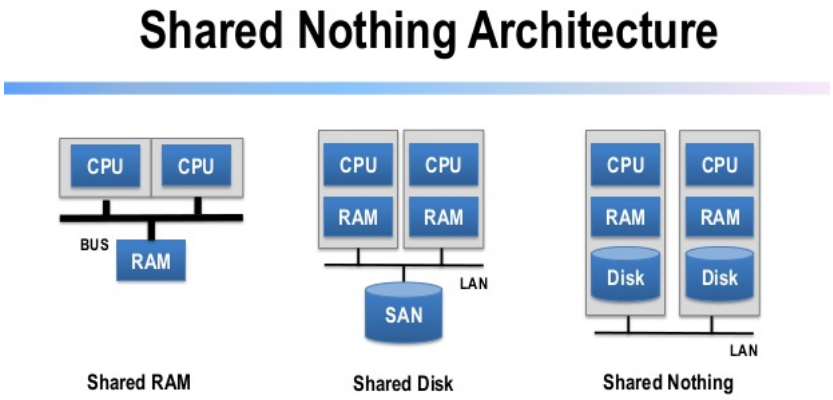
In shared nothing architecture, neither memory nor disk is shared among multiple processors.

Advantages:

- Fault Isolation – It provides benefit of isolating fault. A fault in a single node is contained and confined.
- Scalability - The disk is shared resource. Different nodes will have to take turns to access the critical data. A distributed shared disk system thus compromising on scalability.

The three most common type of architecture for multiprocessor high transaction rate system

- Shared Memory – A common central memory is shared by multiple processors
- Shared Disk – Multiprocessors share a common collection of disks while having their own private memory.
- Shared Nothing – If neither memory nor disk is shared among multiple processor.



Types of Digital Data – Classification of Digital Data

Digital data is classified into the following categories:

- Structured data
- Semi-structured data
- Unstructured data

Structured Data

It owns a dedicated data model. It also has a well defined structure, it follows a consistent order and it is designed in such a way that it can be easily accessed and used by person or a computer. Structured data is usually stored in well defined columns and databases.

Example : DBMS, RDBMS

When a data conforms to a pre-defined schema / structure we say it is structured data.

Sources of structured data

- Databases: Oracle Corp-Oracle, IBM-DB2, Microsof-Microsoft SQL Server, EMC-Greenplum, Teradata-Teradata, MySQL, PostgreSQL.
- Spreadsheets : MS Excel, Google sheets
- On-Line Transaction Processing (OLTP) Systems

Ease of Working with Structured Data

The ease is with respect to the following:

- **Insert/update/delete**: The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.
- **Security**: There are available check encryption and tokenization solutions to warrant the security of information throughout its lifecycle.
- **Indexing:** An index is a data structure that speeds up the data retrieval operations (primarily the SELECT DML statement) at the cost of additional writes and storage space.
- **Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server .
- **Transaction processing**: RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction.

Semi-Structured Data

This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program; for example, en XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

Semi-structured data is also referred to as self-describing structure. It has the following features

- It does not conform to the data models that one typically associates with relational databases or any other form of data tables.
- It uses tags to segregate semantic elements.
- Tags are also used to enforce hierarchies of records and fields within data.
- There is no separation between the data and the schema. The amount of structure used is dictated by the purpose at hand.
- In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes.

Sources of Semi-structured data

Amongst the sources for semi-structured data, the front runners are XML and JSON

- XML: eXtensible Markup Language (XML) is hugely popularized by web services developed utilizing the Simple Object Access Protocol (SOAP) principles.
- JSON: Java Script Object Notation (JSON) is used to transmit data between a server and a web application using REST architecture.
- MongoDB and Couchbase (originally known as Membase, store data natively in JSON format.

Unstructured Data

Unstructured data does not conform to a data model or is not in a form which can be used easily by a computer program.

Unstructured data is completely different of which neither has a structure nor obeys to follow formal structural rules of data models. It does not even have a consistent format and it found to be varying all the time.

About 80–90% data of an organization is in this format.

Sources of Unstructured Data

Web Pages, Images, Free-Form Text, Audios, Videos, Body of Email, Text, Messages, Chats, Social Media data, Word Document.

Issues with terminology

- Structure can be implied despite not being formerly defined.
- Data with some structure may still be labeled unstructured if the structure doesn't help with processing task at hand
- Data may have some structure or may even be highly structured in ways that are unanticipated or unannounced.

Dealing with Unstructured Data

- Data Mining
- Natural Language Processing (NLP)
- Text Analytics
- Noisy Text Analytics

Data Mining:

First, we deal with large data sets. Second, we use methods at the intersection of artificial intelligence, machine learning, statistics, and database systems to unearth consistent patterns in large data sets and/or systematic relationships between variables.

Few popular data mining algorithms are as follows:

- Association rule mining,
- Regression analysis
- Collaborative filtering

Natural language processing (NLP):

It is related to the area of human computer interaction. It about enabling computers to understand human or natural language input.

Text Analytics or Text Mining

Text mining is the process of gleaning high quality and meaningful information (through devising of patterns and trends by means of statistical pattern learning) from text.

It includes tasks such as text categorization, text clusterirg, sentiment analysis, concept/entity extraction, etc.

Noisy Text Analytics

Noisy text analytics: It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message-boards, text messages, etc.

Introduction to Big Data

Definition 1

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency.

Definition 2

Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them.

Definition 3

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Characteristics of Data

- **Composition:** The composition of data deals with the structure of data, that is, the sources of data the granularity, the types, and the nature of data as to whether it is static or real-time streaming.
- **Condition:** The condition of data deals with the state of data, that is, "Can one use this data as is for analysis?" or "Does it require cleansing for further enhancement and enrichment?"
- **Context:** The context of data deals with "Where has this data been generated?" "Why was this data generated?" "How sensitive is this data?" "What are the events associated with this data?" and so on.

Characteristics of Big Data

Big data characteristics are mere word that explain the remarkable potential of big data. In early stages development of big data and related terms there were only 3 V's (Volume, Variety, Velocity) considered as potential characteristics.

But ever growing technology and tools and variety of sources where information being received has potentially increased these 3 V's into 5 V's and still evolving.

The 5 V's are

- Volume
- Variety
- Velocity
- Veracity
- Value

Volume

Volume refers to the unimaginable amounts of information generated every second. This information comes from variety of sources like social media, cell phones, sensors, financial records, stock market etc.

Variety

Variety refers to the many types of data that are available. A reason for rapid growth of data volume is that the data is coming from different sources in various formats.

Big data extends beyond structured data to include unstructured data of all varieties: text, sensor data, audio, video, click streams, log files and more.

The variety of data is categorized as follows:

- Structured – RDBMS
- Semi Structured – XML, HTML, RDF, JSON
- Unstructured- Text, audio, video, logs, images

Velocity

Velocity essentially refers to the speed at which data is being created in real- time. It is the fast rate at which data is received and (perhaps) acted on. In other words it is the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development

Veracity

Data veracity, in general, is how accurate or truthful a data set may be. More specifically, when it comes to the accuracy of big data, it's not just the quality of the data itself but how trustworthy the data source, type, and processing of it is.

Value

Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights.

Mine the data, i.e., a process to turn raw data into useful data. Value represents benefits of data to your business such as in finding out insights, results, etc. which were not possible earlier.

Evolution of Big Data

Common Eras of Evolution

- 1970s and before was the era of mainframes. The data was essentially primitive and structured.
- Relational databases evolved in 1980s and 1990s. The era was of data intensive applications.
- 2000 and beyond: The World Wide Web (WWW) and the Internet of Things (IoT) have led to an onslaught of structured, unstructured, and multimedia data.

Brief History

- In 1960: Charles W. Bachman designed the Integrated Database System, the "first" DBMS.
- IBM created a database system of their own, known as IMS.
- In 1971 : evolved a standardization of a language for data base management called Common Business Oriented Language (COBOL)
- In 1974 :IBM to develop SQL, which was more advanced .
- In 1980s-90s : RDBM Systems like Oracle, MS SQL, DB2, My SQL and Teradata became very popular leading to development of enterprise resource planning systems (ERP), CRM, RDBMS were efficient to store and process structured data.
- In 2000s and beyond: due to explosion of internet processing speeds were required to be faster, and "unstructured" data (art, photographs, music, etc.) became much more common place.
- Unstructured data is both non-relational and schema-less, and Relational Database Management Systems simply were not designed to handle this kind of data. NoSQL database are primarily called as non-relational or distributed database.

Evolution of Terminologies

- The word Big Data is Coined by John Mashey in 1998. He introduced the term in article Big Data… and the Next Wave of Infrastress.
- In 2000, Francis Diebold presented a paper titled "' Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting" to the Eighth World Congress of the Econometric Society.
- Doug Laney in 2001 coined 3 V's Analyst with the Meta Group (Gartner), in his paper "3D Data Management: Controlling Data Volume, Velocity, and Variety."
- The 3V's have become the most accepted dimensions for defining big data.

Big Data Analytics

Definitions

Big data analytics is the often complex process of examining large volume of data to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

Big data analytics is the use of advanced analytic techniques against very large, diverse big data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Big data analytics describes the process of uncovering insights such as trends, hidden patterns, and correlations in large amounts of raw data to help making better informed decisions. It can be used for better decision making, preventing fraudulent activities, among other things.

Why BDA?

Technology-enabled analytics: Quite a few data analytics and visualization tools are available in the market today from leading vendors such as IBM, Tableau, SAS, R Analytics, Statistica, World Programming Systems (WPS), etc. to help process and analyze your big data.

About gaining a meaningful, deeper, and richer insight into your business to steer it in the right direction, understanding the customers demographics to cross-sell and up-sell to them.

BDA involves a tight handshake between three communities: IT, business users, and data scientists

Classification of Analytics

There are basically two schools of thought:

- Those that classify analytics into basic, operationalized, advanced, and monetized.
- Those that classify analytics into analytics 1.0, analytics 2.0, and analytics 3.0

First School of Thought

Basic analytics: This primarily is slicing and dicing of data to help with basic business insights.

Operationalized analytics: It is operationalized analytics if it gets woven into the enterprise's business processes.

Advanced analytics: This largely is about forecasting for the future by way of predictive and

prescriptive modeling.

Monetized analytics: This is analytics in use to derive direct business revenue

Second School of Thought

Analytics 1.0 - Era: mid 1950s to 2009

- Descriptive statistics (report on events, occurrences, etc. of the past)
- Relational databases

Analytics 2.0 - 2005 to 2012

- Descriptive statistics + predictive statistics (use data from the past to make predictions for the future)
- Database appliances, Hadoop clusters, SQL to Hadoop environments, etc.

Analytics 3.0 - 2012 to present

- Descriptive + predictive + prescriptive statistics (use data from the past to make prophecies for the future and at the same time make recommendations to leverage the situation to one's advantage)
- In memory analytics, in database processing, agile analytical methods, machine learning techniques, etc.

There are four types of analytics,

1. Descriptive Analytics
2. Diagnostic Analytics
3. Predictive Analytics
4. Prescriptive Analytics

Descriptive Analytics

- Describing or summarising the existing data using existing business intelligence tools to better understand what is going on or what has happened.
- Descriptive analytics shuffles raw data from various data sources to give meaningful insights into the past, i.e., it helps you understand the impact of past actions.
- It looks at the past performance and understands the performance by mining historical data to understand the cause of success or failure in the past.
- The two main techniques involved are **data aggregation and data mining** stating that this method is purely used for understanding the underlying behavior and not to make any estimations.
- By mining historical data, companies can analyze the consumer behaviors and engagements with their businesses that could be helpful in targeted marketing, service improvement, etc.

- It's used to identify and address the areas of strengths and weaknesses. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.
- Tools used - MS Excel, MATLAB (MaTrix LABoratory), STATA, etc.

Diagnostic Analysis

- Focus on past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.
- Diagnostic analytics is used to determine why something happened in the past. Diagnostic analytics helps identify anomalies and determine casual relationships in data.
- It is characterized by techniques such as **drill-down, data discovery, data mining and correlations.**
- Diagnostic analytics takes a deeper look at data to understand the root causes of the events. It is helpful in determining what factors and events contributed to the outcome.
- It mostly uses probabilities, likelihoods, and the distribution of outcomes for the analysis.
- A few techniques that uses diagnostic analytics include **attribute importance, principle components analysis, sensitivity analysis, and conjoint analysis**.
- Training algorithms for classification and regression also fall in this type of analytics

Predictive Analytics

- Emphasizes on predicting the possible outcome using statistical models and machine learning techniques.
- Predictive analytics is used to predict future outcomes. However, it is important to note that it cannot predict if an event will occur in the future; it merely forecasts what are the probabilities of the occurrence of the event.
- A predictive model builds on the preliminary descriptive analytics stage to derive the possibility of the outcomes. It uses findings of descriptive and diagnostic analytics to detect clusters and exceptions, and to predict future trends, which makes it a valuable tool for forecasting.
- The essence of predictive analytics is to devise models such that the existing data is understood to extrapolate the future occurrence or simply, predict the future data.
- Example application is sentiment analysis where all the opinions posted on social media are collected and analyzed (existing text data) to predict the person’s sentiment on a particular subject as being- positive, negative or neutral.

- Hence, predictive analytics includes building and validation of models that provide accurate predictions.
- Predictive analytics relies on machine learning algorithms like **random forests, SVM**, etc. and statistics for learning and testing the data.
- Usually, companies need trained data scientists and machine learning experts for building these models.
- The most popular tools for predictive analytics include Python, R, RapidMiner, etc.
- Predictive analytics is used by companies such as Walmart, Amazon, and other retailers to recognize sales patterns based on customer buying patterns, forecast consumer actions, forecast stock levels, and predict the sales revenue at the end of each quarter or year

Prescriptive Analytics:

- It is a type of predictive analytics that is used to recommend one or more course of action on analyzing the data.
- It can suggest all favorable outcomes according to a specified course of action and also suggest various course of actions to get to a particular outcome.
- Hence, it uses a strong feedback system that constantly learns and updates the relationship between the action and the outcome.
- Prescriptive analytics utilizes emerging technologies and tools, such as **Machine Learning, Deep Learning, and Artificial Intelligence algorithms**, making it modern to execute and oversee.
- Furthermore, this cutting edge data analytics type requires internal as well as external past data to provide users with favorable outcomes.
- For example, while calling for a cab online, the application uses GPS to connect you to the correct driver from among a number of drivers found nearby.
- Hence, it optimises the distance for faster arrival time. Recommendation engines also use prescriptive analytics.
- Prescriptive analytics is an advanced analytics concept based on –
    - Optimization that helps achieve the best outcomes.
    - Stochastic optimization helps understand how to achieve the best outcome and identify data uncertainties to make better decisions.

Top Challenges Facing Big Data

There are mainly seven challenges of big data: Scale, Security, Schema, Continuous availability, Consistency, Partition tolerant and data quality.

Scale: Storage (RDBMS (Relational Database Management System) or NoSQL (Not only SQL)) is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the attack of large volume, velocity and variety of big data. Should you scale vertically or should you scale horizontally?

Security: Most of the NoSQL big data platforms have poor security mechanisms (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data. A spot that cannot be ignored given that big data carries credit card information, personal information and other sensitive data.

Schema: Rigid schemas have no place. We want the technology to be able to fit our big data and not the other way around. The need of the hour is dynamic schema. Static (pre-defined schemas) are obsolete.

Continuous availability: The big question here is how to provide 24/7 support because almost all RDBMS and NoSQL big data platforms have a certain amount of downtime built in.

Consistency: Should one opt for consistency or eventual consistency?

Partition tolerant: How to build partition tolerant systems that can take care of both hardware and software failures?

Data quality: How to maintain data quality- data accuracy, completeness, timeliness, etc.? Do we have appropriate metadata in place?

Other Challenges

- Need For Synchronization Across Disparate Data Sources
- Acute Shortage Of Professionals Who Understand Big Data Analysis
- Getting Meaningful Insights Through The Use Of Big Data Analytics
- Getting Voluminous Data Into The Big Data Platform
- Uncertainty Of Data Management Landscape

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

Acute Shortage Of Professionals Who Understand Big Data Analysis

- With the exponential rise of data, a huge demand for big data scientists and Big Data analysts has been created in the market.
- 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.

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.

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.

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

Importance of Big Data Analytics

The various approaches to analysis of data and what it leads to.

Reactive - Business Intelligence: It allows the businesses to make faster and better decisions by providing the right information to the right person at the right time in the right format. It is about analysis of the past or historical data and then displaying the findings of the analysis or reports in the form of enterprise dashboards, alerts, notifications, etc. It has support for both pre-specified reports as well as ad hoc querying.

Reactive - Big Data Analytics: Here the analysis is done on huge datasets but the approach is still reactive as it is still based on static data.

Proactive - Analytics: This is to support futuristic decision making by use of data mining predictive modelling, text mining, and statistical analysis on. This analysis is not on big data as it still the traditional database management practices on big data and therefore has severe limitations on the storage capacity and the processing capability.

Proactive - Big Data Analytics: This is filtering through terabytes, petabytes, exabytes of information to filter out the relevant data to analyze. This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data.

Organizations can use big data analytics systems and software to make data-driven decisions that can improve business-related outcomes.

The benefits may include

- more effective marketing,
- new revenue opportunities,
- customer personalization and
- improved operational efficiency

The analytical accuracy will lead a greater positive impact in terms of

- Enhancing operational efficiencies,
- reducing cost and time, and
- originating new products, new services, and
- optimizing existing services
- Higher profits

Data Analytics Tools

Data Science Tools

The data science profession is challenging, but fortunately, there are plenty of tools available to help the data scientist succeed at their job.

- Data Analysis: SAS, Jupyter, R Studio, MATLAB, Excel, RapidMiner
- Data Warehousing: Informatica/ Talend, AWS Redshift
- Data Visualization: Jupyter, Tableau, Cognos, RAW, Matlibpro, Power BI, Zoho Analytics
- Machine Learning: Spark MLib, Mahout, Azure ML studio, Scikilearn, Pytorch, Tensorflow

Hadoop Environment - Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.

1. Massive Storage
2. Faster Processing

Core Components of Hadoop

- Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.
- Hadoop MapReduce: This is YARN-based system for parallel processing of large data sets.
- Hadoop Yet Another Resource Negotiator (YARN): This is a framework for job scheduling and cluster resource management.

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

- Apache Cassandra
- Simple DB – Amazon
- Google BigTable
- MongoDB
- HBase

Visualization Tools

Tableau is an end-to-end data analytics platform that allows you to prep, analyze, collaborate, and share your big data insights. Tableau excels in self-service visual analysis, allowing people to ask new questions of governed big data and easily share those insights across the organization.

Microsoft Power BI

The Microsoft Power BI is the data visualization tool that is used for business intelligence type of data. It is and can be used for reporting, self-service analytics, and predictive analytics.

JupyteR A web-based application, JupyteR, is one of the top-rated data visualization tools that enable users to create and share documents containing visualizations, equations, narrative text, and live code. JupyteR is ideal for data cleansing and transformation, statistical modeling, numerical simulation, interactive computing, and machine learning.

RAW RAW, better-known as RawGraphs, works with delimited data such as TSV file or CSV file. It serves as a link between data visualization and spreadsheets. Featuring a range of non-conventional and conventional layouts, RawGraphs provides robust data security even though it is a web-based application.

Python:

This is one of the most versatile programming languages that is rapidly being deployed for various applications including Machine Learning.

SAS:

SAS is an advanced analytical tool that is being used for working with huge volumes of data and deriving valuable insights from it.

R Studio

R Programming: R is the Number 1 programming language that is being used by Data Scientists for the purpose of statistical computing and graphical applications alike.

RapidMiner is a software package that allows data mining, text mining and predictive analytics. Rapidminer is a comprehensive data science platform with visual workflow design and full automation