

# **MCA 304B: Big Data Analytics**

## **UNIT-I**

What is Big Data: Varieties of Data – Unstructured data – Trends in Data Storage – Basically Available Soft State Eventual Consistency (BASE) – Industry Examples of Big Data.

## **UNIT -II**

Big Data Technology: New and older approaches – Data Discovery – Terminologies used in Big Data Environments- Open Source technologies for Big Data Analytics – Cloud and Big Data – Big Data Foundation – Computation – Limitations – Big Data Emerging Technologies.

## **UNIT III**

Business Analytics –Consumption of Analytics – Creation to Consumption of Analytics – Data visualization by Organizations – 90/10 rule of critical thinking – Decision sciences and analytics – Learning over knowledge – Agility – Scale and convergence – Privacy and security in Big Data.

## **UNIT IV**

Predictive Analytics – Target Definition – Linear Regression – Logistic Regression – Decision trees – Neural Networks – Support Vector machines - Classification trees – Ensemble methods – Association Rules -Segmentation , Sequence Rules, Social Network analytics.

## **UNIT V**

Hadoop – Why Hadoop? – Why not RDBMS? – RDBMS Versus Hadoop – Components of Hadoop – Hadoop File System – Hadoop Technologies Stack – Managing Resources and Applications with Hadoop YARN – Data ware housing Hadoop Concepts – Applications of Hadoop using PIG,YARN,HIVE.

### **Text Books**

Big Data and Analytics, Seema Acharya ,Subhashini chellappan, Wiley publications Baesens, 2014, Analytics in a Big Data World: The Essential Guide to Data Science and Its

Applications, Wiley India Private Lim

### **REFERENCE BOOK**

“Big Data Analytics: Systems, Algorithms, Applications” Prabhu, C.S.R  
., Sreevallabh Chivukula ,Mogadala .A, Ghosh, R., Livingston

## **LECTURE NOTES**

### **UNIT -1**

#### **Big Data:**

Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.[2] Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and

data source. Big data was originally associated with three key concepts: volume, variety, and velocity.

The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Thus a fourth concept, veracity, refers to the quality or insightfulness of the data. Without sufficient investment in expertise for big data veracity, then the volume and variety of data can produce costs and risks that exceed an organization's capacity to create and capture value from big data.

Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem.

Big data also encompasses a wide variety of data types, including the following: structured data, such as transactions and financial records; unstructured data, such as text, documents and multimedia files; and. Semi -Structured

data, such as web server logs and streaming data from sensors.

Big data is the set of technologies created to store, analyze and manage this bulk data, a macro-tool created to identify patterns in the chaos of this explosion in information in order to design smart solutions. Today it is used in areas as diverse as medicine, agriculture, gambling and environmental protection. Big data is the set of technologies created to store, analyze and manage this bulk data, a macro-tool created to identify patterns in the chaos of this explosion in information in order to design smart solutions. Today it is used in areas as diverse as medicine, agriculture, gambling and environmental protection.

Almost an endless number of applications: GPS systems can detect traffic jams in the area checked by a user and suggest alternatives; a subscription streaming TV channel has created the characters and plot of its most successful series by analyzing the contents its viewers consume and prefer to watch; smart watches monitor the heart rate of millions of users and identify patterns that can anticipate to and prevent cardiovascular diseases; humidity sensors in crop fields plan the irrigation frequency, combining their data with the weather forecasts, and a long etcetera. Their applications have even reached the world of politics: Juan Verde, the Spanish adviser who worked on the political campaigns of the democrat party in the USA stated that: "These are not the TV elections anymore; they are the elections of big data".

### **Uses of Big data:**

## main uses and applications

#R&D #internet Nowadays, almost 6,500 million connected devices share information over the Internet. In 2025, this figure will rise up to 20,000 million. Big data analyses this "sea of data" to convert it into the information that is transforming our world.

The digital revolution is changing the economy, society and people. The data generated by thousands of millions of devices is in the center of this revolution. According to Gartner, there were close to 6,500 million devices in 2016 and this figure will rise to 20,000 million by 2025. Likewise, according to a top ICT solution provider, the Internet of Things will lead to sharp increase, with up to 100,000 million connected devices.

Big data is the set of technologies created to store, analyze and manage this bulk data, a macro-tool created to identify patterns in the chaos of this explosion in information in order to design smart solutions. Today it is used in areas as diverse as medicine, agriculture, gambling and environmental protection.

Almost an endless number of applications: GPS systems can detect traffic jams in the area checked by a user and suggest alternatives; a subscription streaming TV channel has created the characters and plot of its most successful series by analyzing the contents its viewers consume and prefer to watch; smart watches monitor the heart rate of millions of users and identify patterns that can anticipate to and prevent cardiovascular diseases; humidity sensors in crop fields plan

the irrigation frequency, combining their data with the weather forecasts, and a long etcetera. Their applications have even reached the world of politics: Juan Verde, the Spanish adviser who worked on the political campaigns of the democrat party in the USA stated that: "These are not the TV elections anymore; they are the elections of big data" .One of the main applications of advanced data analysis is the study of consumer patterns. Social networks, such as Facebook, Twitter or Instagram, are a tool used by brands to learn more about their consumers and connect with them. Companies have also started to gather data from their consumers. A company specializing in big data and retail intelligence has installed 15,000 sensors in the shopping areas of 25 countries. Thanks to the data gathered with these sensors, they have detected that 36.8% of customers entering a shop in Spain have bought something from the shop.

Digital transformation companies leads to the generation of huge volumes of data that organizations do not know how to use and manage. And this is already being portrayed in the labour market. According to Randstad, a labour service company, the big data specialist profile is one of the three most popular profiles in 2017. Companies are now asking their candidates to have international experience, strategic vision, analytical capacity and adaptation to change as the main requirements.

Big data is expected to create 900,000 jobs around the world in the next six years, and companies who manage to use data smartly will increase their productivity

## **Varieties of Big Data**

Big data is classified in three ways: Structured Data. Unstructured Data. Semi-Structured Data.

Variety of Big Data refers to structured, unstructured, and semi structured 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. Big data is a collection of data from many different sources and is often describe by five characteristics: volume, value, variety, velocity, and veracity. Big data Variety of Big Data refers to structured, unstructured, and semi structured 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, Variety of Big Data refers to structured, unstructured, and semi structured 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. Big data is a collection of data from many different sources and is often describe by five characteristics: volume, value, variety, velocity, and veracity. Big data” is a term relative to the available computing and storage power on the market — so in 1999, one gigabyte (1 GB) was considered big data. Today, it may consist of petabytes (1,024 terabytes) or exabytes

(1,024 petabytes) of information, including billions or even trillions of records from millions of people.

Structured data is the easiest to work with. It is highly organized with dimensions defined by set parameters. Variety of Big Data refers to structured, unstructured, and semi structured 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. Big data is a collection of data from many different sources and is often describe by five characteristics: volume, value, variety, velocity, and veracity. Big data” is a term relative to the available computing and storage power on the market — so in 1999, one gigabyte (1 GB) was considered big data. Today, it may consist of petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of information, including billions or even trillions of records from millions of people.

Structured data is the easiest to work with. It is highly organized with dimensions defined by set parameters . Photos, videos, audios, SM posts, and so Much more. Big data is a collection of data from many different sources and is often describe by five Characteristics: volume, value, variety, velocity, and veracity. Big data” is a term relative to the available computing and storage power on the market — so in 1999, one gigabyte (1 GB) was considered big data. Today, it may consist of petabytes (1,024 terabytes) Or exabytes (1,024 petabytes) of information, including billions or even trillions of records from millions of peoples. Structured data is



the easiest to work with. It is highly organized with Dimensions defined by set parameters.

- Billing
- Contact
- Address
- Expenses
- Debit/credit card numbers

Because structured data is already tangible numbers, it's much easier for a Program to sort through and collect data. Structured data follows schemas: essentially road maps to specific data points. These schemas outline where each datum is and what it means. A payroll database will lay out employee identification information, pay rate, Hours worked, how compensation is delivered, etc. The schema will define Each one of these dimensions for whatever application is using it. The program Won't have to dig into data to discover what it actually means, it can go Straight to work collecting and processing it. Working With It Structured data is the easiest type of data to analyze because it requires little To no preparation before processing. A user might need to cleanse data and Pare it down to only relevant points, but it won't need to be interpreted or Converted too deeply before a true inquiry can be performed. One of the major perks of using structured data is the streamlined process of Merging enterprise data with relational. Because pertinent data dimensions Are usually defined and specific elements are in a

uniform format, very little Preparation needs to be done to make all sources compatible.

The ETL process for structured data stores the finished product in what is Called a data warehouse. These databases are highly structured and filtered for The specific analytics purpose the initial data was harvested for. Relational databases are easily-queried datasets. They allow users to find External information and either study it standalone or integrate it with their Internal data for more context. Relational database management systems use Summary Applications data can be classified as structured, semi-structured, and unstructured data. Structured data is neatly organized and obeys a fixed set of rules. Semi-structured data doesn't obey any schema, but it has certain discernible features for an organization. Data serialization languages are used to convert data objects into a byte stream. These include XML, JSON, and YAML. Unstructured data doesn't have any structure at all. All these three kinds of data are present in an application. All three of them play equally important roles in developing resourceful and attractive applications.

## **UNSTRUCTURED DATA**

Unstructured data (or unstructured information) is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult

to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents. In 1998, Merrill Lynch said “unstructured data comprises the vast majority of data found In an organization, some estimates run as high as 80%.” Unstructured data just happens to Be in greater abundance than structured data is. Examples of unstructured data are: Rich Media. Media and entertainment data, surveillance data, geo-spatial data, audio, weather data. Unstructured data is information that is not arranged according to a preset data model or schema, And therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia Are two common types of unstructured content. Unstructured data is information that is not arranged according to a preset data model or schema, and therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia are Two common types of unstructured content. Unstructured data helps you improve customer experience Unstructured data offer the key to helping you really get to know your customers. You can come to Understand things like what trends they value on social media, what opinions they have, and, Ultimately, what they want from your brand Retailers, manufacturers and other companies analyze unstructured data to improve customer Experience and enable targeted marketing. They also do sentiment analysis to better understand Customers and identify attitudes about products, customer service and corporate brands Keep the business objective(s) in mind Define metadata for faster data access Choose the right

analytics techniques Exploratory data analysis techniques Qualitative data analysis techniques Artificial Intelligence (AI) and Machine Learning (ML) techniques Identify the right data sources Evaluate the technologies you'd want to use Get real-time data access.

Store and integrate data using data lake Wrangle the data to get the desired features In the next few sections, we'll discuss the various unstructured data analysis techniques and tips, Challenges in handling unstructured data, and suggestions for overcoming these challenges .difficult to understand using traditional programs as compared to data stored in fielded Form in databases or annotated (semantically tagged) in documents .In 1998, Merrill Lynch said "unstructured data comprises the vast majority of data found In an organization, some estimates run as high as 80%." Unstructured data just happens to Be in greater abundance than structured data is. Examples of unstructured data are: Rich Media. Media and entertainment data, surveillance data, geo-spatial data, audio, weather data .Unstructured data is information that is not arranged according to a preset data model or schema, And therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia Are two common types of unstructured content. Unstructured data is information that is not arranged according to a preset data model or schema, And therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia are Two common types of unstructured content. Unstructured data helps you improve customer experience Unstructured data offer the

key to helping you really get to know your customers. You can come to Understand things like what trends they value on social media, what opinions they have, and, Ultimately, what they want from your brand Retailers, manufacturers and other companies analyze unstructured data to improve customer

Experience and enable targeted marketing. They also do sentiment analysis to better understand Customers and identify attitudes about products, customer service and corporate brands Keep the business objective(s) in mind Define metadata for faster data access Choose the right analytics techniques Exploratory data analysis techniques Qualitative data analysis techniques Artificial Intelligence (AI) and Machine Learning (ML) techniques Identify the right data sources Evaluate the technologies you'd want to use Get real-time data access Store and integrate data using data lake Wrangle the data to get the desired features In the next few sections, we'll discuss the various unstructured data analysis techniques and tips,

Challenges in handling unstructured data, and suggestions for overcoming these challenges. difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents .In 1998, Merrill Lynch said "unstructured data comprises the vast majority of data found in an organization, some estimates run as high as 80%." Unstructured data just happens to be in greater abundance than structured data is. Examples of unstructured data are: Rich media. Media and entertainment data,

surveillance data, geo-spatial data, audio, weather data. Unstructured data is information that is not arranged according to a preset data model or schema, and therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia are two common types of unstructured content.

Unstructured data is information that is not arranged according to a preset data model or schema, and therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia are two common types of unstructured content. Unstructured data helps you improve customer experience. Unstructured data offers the key to helping you really get to know your customers. You can come to understand things like what trends they value on social media, what opinions they have, and, ultimately, what they want from your brand. Retailers, manufacturers and other companies analyze unstructured data to improve customer experience and enable targeted marketing. They also do sentiment analysis to better understand customers and identify attitudes about products, customer service and corporate brands. Keep the business objective(s) in mind. Define metadata for faster data access. Choose the right analytics techniques. Exploratory data analysis techniques. Qualitative data analysis techniques. Artificial Intelligence (AI) and Machine Learning (ML) techniques. Identify the right data sources. Evaluate the technologies you'd want to use. Get real-time data access. Store and integrate data using data lake. Wrangle the data to get the desired features.

In the next few sections, we'll discuss the various unstructured data analysis techniques and tips, challenges in handling unstructured data, and suggestions for overcoming these challenges difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents .In 1998, Merrill Lynch said "unstructured data comprises the vast majority of data found in an organization, some estimates run as high as 80%." Unstructured data just happens to be in greater abundance than structured data is. Examples of unstructured data are: Rich media. Media and entertainment data, surveillance data, geo-spatial data, audio, weather data.Unstructured data is information that is not arranged according to a preset data model or schema, and therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia are two common types of unstructured content. Unstructured data is information that is not arranged according to a preset data model or schema, and therefore cannot be stored in a traditional relational database or RDBMS. Text and multimedia are two common types of unstructured content. Unstructured data helps you improve customer experience Unstructured data offer the key to helping you really get to know your customers. You can come to understand things like what trends they value on social media, what opinions they have, and, ultimately, what they want from your brand Retailers, manufacturers and other companies analyze unstructured data to improve customer experience and enable targeted marketing. They also do sentiment analysis

to better understand customers and identify attitudes about products, customer service and corporate brands Keep the business objective(s) in mind Define metadata for faster data access.

Choose the right analytics techniques  
Exploratory data analysis techniques Qualitative data analysis techniques Artificial Intelligence (AI) and Machine Learning (ML) techniques Identify the right data sources Evaluate the technologies you'd want to use Get real-time data access Store and integrate data using data lake Wrangle the data to get the desired features In the next few sections, we'll discuss the various unstructured data analysis techniques and tips, challenges in handling unstructured data, and suggestions for overcoming these challenges

## **TRENDS IN DATA**

WWT storage industry experts dive into eight data storage trends, based on customer insights, Testing in the ATC and OEM partnerships. The new calendar year has gotten off to a fast start, and without fail, our customers continue to Work on improving their data strategies. In a world of cyber threats, customers are There focus on protecting their data, data sovereignty, data locality, data tiering and density .Explosive data growth and data placement remain challenges for decision makers in an Increasingly digital world. How does an organization accommodate data spread across a private Data center, colocation facility or a public cloud ?The last two years have introduced change at a pace the industry has never seen. Some customers Simply survived



the pandemic, while others thrived. Most organizations expanded their IT Footprint with new investments in areas supporting a virtual workforce like video collaboration And virtual desktop deployments. Each of these investments generates substantially more data. In this article, we will be diving into data storage industry trends, leveraging insights we've Uncovered from working in and around global service providers, global financials, healthcare and Enterprise accounts; testing these technologies in our Advanced Technology Center (ATC); and Working with our OEM partners, ranging from industry stalwarts to stealth startups. Customize storage solutions to the application Data conversations typically start with the use case and applications that will leverage the storage Subsystem. Primary use cases for shared storage still lean toward databases, high-performance Applications such as those for data science, enterprise resource planning (ERP) solutions and Large-scale hypervisors like VMware. These continue to make up the most significant amounts of Primary storage consumption inside the four walls of a traditional data center .What's changed ?Companies must plan for storage needs in various locations with unpredictable timelines. For Example, applications that reside on-premises today may or may not move to the public Cloud. Other organizations made storage decisions based on short-term supply chain constraints That may be sub-optimized for the application and require remediation as product availability Frees up. As things normalize in 2022 and cloud strategies come into focus, storage groups must work Again to customize storage to the

application, ensuring optimal performance, uptime, resiliency  
And, ultimately, a great end-user experience.

## Extend storage architecture to the public cloud

The new IT world is hybrid and complex with new areas emerging like Cloud Ops and FinOps to Optimize data storage across clouds. Public cloud providers like AWS, Azure and Google continue to expand, and we see independent software vendors (ISV) investing heavily to provide data services in the public cloud. As customers begin their journey to the cloud, we see traditional storage teams expanding their knowledge and contributions to applications in the public cloud. Organizations will consider leveraging and extending traditional ISV data management solutions to the public cloud versus leveraging cloud-native services. How a customer leverages storage solutions on-premises can and should be part of the decision-making process on how to plan for a shift to ISV or cloud-native deployments for storage in a public cloud. In addition to technical considerations, organizations must also understand how various IT and cloud groups interact. We saw dedicated cloud teams formed early in many organizations to move quickly to public cloud; now those organizations must work to ensure these teams align closely with traditional IT organizations to build out an integrated and comprehensive data strategy. Cloud strategy should not be an "either/or" conversation; it must be an "and" conversation. Storage fabric – unlock the performance of all flash 2016 was the year most storage manufacturers began delivering solutions with 100% solid-state flash drives. This was the "year of all flash," and the array manufacturers

were racing to hit 100,000 input/output operations (IOPs) at sub-millisecond (ms) response times in what's called a "three-tier stack"-- meaning separate compute resources and switching fabric (usually Fiber Channel at this stage of the game), coupled with an array-based storage platform.

The introduction of NAND flash into mainstream storage array products has since evolved to additional advancements in the storage media to "storage class memory" (SCM), which improves response times and delivers advancements in quad-level cell technology (QLC) to help drive down costs of the previous tri-level cell products still in heavy use today. Additionally, until the last few years, all these very modern storage types have ridden atop a small computer serial interface (SCSI) stack -- most recently via serial attached SCSI (SAS) -- that was designed for spinning media starting in the 1980s. This fabric must evolve as well to ensure more efficient use of the compute stack's integration with storage. Non-Volatile Memory Express (NV Me) and NV Me over Fabrics (NV Me-of) have been introduced in recent years to help solve that problem. Now that the server operating systems are being updated with native NV Me-of capabilities for FC, TCP and RoCEv2 (for example), we anticipate growth in adoption to new storage fabric technologies as customers look to get more out of their IT investments.

OPEX models picking up steam Some customers are looking for a cloudlike experience but have no interest in putting their data into a public cloud provider, while others have a cloud-first/cloud-native mentality. Whether your

organization's crown jewels are stored in an EPIC solution for healthcare, a database for an enterprise resource planning (ERP) system, or you have machine learning (ML) initiatives, we are seeing a shift from traditional capital expenditure (CAPEX) spending on IT infrastructure to a "pay by the drip" operational expense (OPEX) model. In the data storage space, this is referred to as storage as a service (STAAS). These models vary greatly from OEM to OEM, and it's important to understand the nuances of

each; they include but are not limited to:

- Dell APEX
- NetApp Keystone
- Pure as a Service
- HPE Green lake
- Hitachi Ever flex
- IBM Storage as a Service

Automation continues its march Automation continues to be top of mind for our customers. The most successful organizations Are selective and targeted with automation initiatives, prioritizing areas that accelerate storage Adoption, increase reliability and reduce operational costs. Many dive directly into picking tools and engineering a solution, but it's important to first Understand the drivers behind automation and map out a comprehensive approach inclusive of People, process and technology to maximize the impact. There are common patterns we see within organizations that have successful automation

Programs:

- Culture of collaboration
- Aligning teams around a “platform” concept
- Culture of learning

Collaboration ensures mutual understanding among the teams, both within the storage discipline And beyond, leading to a superior outcome for data consumers. This collaboration typically leads To an evolution within IT around a platform concept. Moving to a platform concept forces teams To think differently. Instead of focusing on silos, platform teams gravitate toward learning more About and understanding consumers of data and work to create solutions that help those teams Operate with more autonomy .While technology remains important, understanding desired outcomes and actively connecting People, process and technology to those outcomes is paramount.

Edge driving new data requirements Throughout the last 24 months, many organizations had to move very quickly to ensure they Could adhere to the new demands of remote work at the edge. As if keeping a remote workforce Up and running isn't challenging enough, IT organizations had to also produce new ways of Protecting the data generated at the edge while keeping that data secure. The downstream data Impacts to an IT organization come full circle when designing a solution to maintain workforce Productivity. In addition to a user's persistent data generated at the edge, the mere existence of a Virtual desktop to perform job

functions increases the need for performant primary storage solutions that can adhere to flexible working hours and parallel workstreams for many users Accessing the storage at once.

Cyber threats are impacting data strategies In the world we live in today, the threat of a cyber-attack is very real. Ransomware attacks are at An all-time high, and organizations must proactively protect themselves from bad actors. Storage Providers often discuss how their solution with immutability is a one-stop-shop to solving the Ransomware problem, but in reality, it is just one piece of the puzzle .There is no silver bullet. Enterprises must pursue a comprehensive strategy spanning the Business, application, enterprise architecture, security and IT teams to mitigate risk and improve Overall data security. Most organizations will pursue a turnkey cyber resilience program in the Long-term but will start with a cyber recovery initiative focused around a data vault or cyber Vault concept .With the above in mind, there are still things that IT and the storage teams can do to protect data. Understanding and tiering data according to application priority is a good place to start. Taking Advantage of encryption and immutability features with critical production data contributes to Mitigating exposure as well. And adding backups, data cleansing and local/remote storage Replication can contribute to a more robust data protection posture.

## **INDUSTRY OF BIG DATA**

Last, but not least, most organizations are developing cloud-native architectures based on Containers but are struggling to provide persistent storage to these ephemeral container Environments .In recent years, development groups were just dipping their toes in the water with Kubernetes (K8), but recently, the level of adoption has increased substantially with organizations moving Towards a DevOps mindset. This accelerated adoption has created the need for containers to Communicate with storage subsystems for block and file-level persistence. The container storage Interface (CSI), in the case of Kubernetes, gives the container the ability to retain their data Kubernetes code.This is important as applications like databases are not ephemeral, so if organizations try to run These in a container, the data must survive a reboot, or the application will be corrupted. CSI Comes in many deployments, from basic drivers and APIs to full orchestration suites. Most Storage manufacturers have written a driver and support CSI now.

How WWT can help We covered a lot of ground above, and there are also other areas we're tracking, such as next-Generation artificial intelligence, machine learning and associated storage solutions. To help you Keep up with the changing storage landscape, there are many ways WWT can assist in giving You insights into critical key performance indicators related to your infrastructure and storage Stack, including protecting your most valuable IT asset – your data. solutions that can adhere to flexible working hours and parallel workstreams for many users

accessing the storage at once.. Cyber threats are impacting data strategies

In the world we live in today, the threat of a cyber-attack is very real. Ransomware attacks are at

an all-time high, and organizations must proactively protect themselves from bad actors. Storage

providers often discuss how their solution with immutability is a one-stop-shop to solving the

ransomware problem, but in reality, it is just one piece of the puzzle.

There is no silver bullet. Enterprises must pursue a comprehensive strategy spanning the

business, application, enterprise architecture, security and IT teams to mitigate risk and improve

overall data security. Most organizations will pursue a turnkey cyber resilience program in the

long-term but will start with a cyber recovery initiative focused around a data vault or cyber

vault concept.

With the above in mind, there are still things that IT and the storage teams can do to protect data.

Understanding and tiering data according to application priority is a good place to start. Taking

advantage of encryption and immutability features with critical production data contributes to



mitigating exposure as well. And adding backups, data cleansing and local/remote storage

replication can contribute to a more robust data protection posture .Containerized applications need storage too

Last, but not least, most organizations are developing cloud-native architectures based on

containers but are struggling to provide persistent storage to these ephemeral container

environments .In recent years, development groups were just dipping their toes in the water with Kubernetes (K8), but recently, the level of adoption has increased substantially with organizations moving towards a DevOps mindset. This accelerated adoption has created the need for containers to communicate with storage subsystems for block and file-level persistence. The container storage interface (CSI), in the case of Kubernetes, gives the container the ability to retain their data

persistence without the need to touch the core Kubernetes code.This is important as applications like databases are not ephemeral, so if organizations try to run these in a container, the data must survive a reboot, or the application will be corrupted. CSI comes in many deployments, from basic drivers and APIs to full orchestration suites. Most storage manufacturers have written a driver and support CSI now.

How WWT can help We covered a lot of ground above, and there are also other areas we're tracking, such as

next-generation artificial intelligence, machine learning and associated storage solutions. To help you keep up with the changing storage landscape, there are many ways WWT can assist in giving you insights into critical key performance indicators related to your infrastructure and storage stack, including protecting your most valuable IT asset – your data. We have labs for testing next-generation protocols, including NV Me and NV Me-of, and offer workshops around block, file and object storage solutions that give an unbiased approach.

We also recommend exploring our Advanced Technology Center (ATC) to gain hands-on

experience with the latest technologies and cut your proof-of-concept time from months to weeks. WWT's deep-rooted relationships with major OEMs and our rigorous evaluation of recent technology providers can help streamline decision-making, testing and trouble shooting .For more information on primary storage, data protection, cyber resilience or any of the topics mentioned within the article, connect with one of our storage industry experts today.

## **SHORT QUESTIONS**

- 1.Discuss the varieties of data?
- 2.What is industry of big data?
- 3.Write about unstructured data?
- 4.What is big data?
- 5.Write about trends in data storage-industry?

## **LONG QUESTIONS**

- 1.What is big data? Explain varieties of big data?
- 2.Write about unstructured data? Explain about details?
- 3.What is trends in data storage-industry Examples of big data?
- 4.Write about industry of big data in details?

## **UNIT-2**

### **Big Data technology**

Big Data Technologies can be defined as software tools for analyzing, processing, and extracting data from an extremely complex and large data set with which traditional management tools can never deal. The bubble around Big Data has certainly started to burst, and the coming year awaits reasonable developments in the applications of the Big Data world. Well, most of us are now more than familiar with terms like Hadoop, Spark, NO-SQL, Hive, Cloud, etc. We know there are at least 20 NO-SQL databases and a number of other Big Data technologies emerging every month. But which of these Big Data technologies see prospects going forward? Big data is a specific indicator for the vast assembly of data, increasing enormously in size and exponentially with time. Big Data Technologies can be defined as software tools for analyzing, processing, and extracting data from an extremely complex and large data set with which traditional management tools can never deal.

Operational Big Data Technologies indicates the volume of data generated every day, such as online transactions, social media or any information from a particular company used for analysis by software based on big data technology. It acts as raw data to supply big data analysis technology. A few Operational Big Data Technologies cases include information on MNC management, Amazon, Flipkart, Walmart, online ticketing for movies, flights, railways, and more.

- Bookings for trains, flights, buses, movies, etc. online using an online booking system.
- Online trading or shopping on e-commerce websites such as Amazon, Ajo, Myntra, etc.
- Online data from social networking sites such as Instagram, Facebook, Messenger, WhatsApp, etc.
- Executive details or Employee data of MNCs.

### Analytical Big Data Technologies

Analytical Big Data Technologies concerns the advanced adjustment of Big Data Technologies, which is rather complicated than Operational Big Data. This category includes the real analysis of Big Data, which is essential to business decisions. Some examples in this area include stock marketing, weather forecasting, time series, and medical records analysis.

- Stock marketing data.

- Weather forecasting data.
  - Maintaining detailed space mission databases where each and every detail about a mission is mentioned.
  - Medical records allow doctors to monitor the health status of a patient
- Hadoop Ecosystem Hadoop Framework was developed to store and process data with a simple programming model in a

Distributed data processing environment. The data present on different high-speed and low-expense Machines can be stored and analyzed. Enterprises have widely adopted Hadoop as Big Data Technologies for their data warehouse needs in the past year. The trend seems to continue and grow in the coming year as well. Companies that have not explored Hadoop so far will most likely see its Advantages and applications.

Artificial Intelligence Artificial Intelligence is a broad bandwidth of computer technology that deals with the development of Intelligent machines capable of carrying out different tasks typically requiring human intelligence. AI is Developing fast, from Apple's Siri to self-driving cars. As an interdisciplinary branch of science, it takes into account a number of approaches, such as increased Machine Learning and Deep Learning to Make a remarkable shift in most tech industries. AI is revolutionizing the existing Big Data Technologies.

NoSQL Database NoSQL includes various Big Data Technologies in the database, which are developed to design Modern applications. It shows a non-SQL or non-relational

database providing a method for data Acquisition and recovery. They are used in Web and Big Data Analytics in real time. It stores Unstructured data and offers faster performance and flexibility while addressing various data types—For example, MongoDB, Redis, and Cassandra. It provides design integrity, easier horizontal scaling, And control over opportunities in a range of devices. It uses data structures that are different from Those concerning databases by default, which speeds up NoSQL calculations. Facebook, Google, Twitter, and similar companies store user data on terabytes daily.

R Programming R is one of the open-source Big Data Technologies and programming languages. The free software Is widely used for statistical computing, visualization, and unified development environments such as

- Stock marketing data.

- Weather forecasting data.
- Maintaining detailed space mission databases where each and every detail about a mission is mentioned.

- Medical records allow doctors to monitor the health status of a patient

Hadoop Ecosystem Hadoop Framework was developed to store and process data with a simple programming model in a distributed data processing environment. The data present on different high-speed and low-expense machines can be stored and analyzed.

Enterprises have widely adopted Hadoop as Big Data Technologies for their data warehouse needs in the past year. The trend seems to continue and grow in the coming year as well. Companies that have not explored Hadoop so far will most likely see its advantages and applications.

Artificial Intelligence is a broad bandwidth of computer technology that deals with the development of intelligent machines capable of carrying out different tasks typically requiring human intelligence. AI is developing fast, from Apple's Siri to self-driving cars. As an interdisciplinary branch of science, it takes into account a number of approaches, such as increased Machine Learning and Deep Learning to make a remarkable shift in most tech industries. AI is revolutionizing the existing Big Data Technologies.

NoSQL Database NoSQL includes various Big Data Technologies in the database, which are developed to design modern applications. It shows a non-SQL or non-relational database providing a method for data acquisition and recovery. They are used in Web and Big Data Analytics in real time. It stores unstructured data and offers faster performance and flexibility while addressing various data types—for example, MongoDB, Redis, and Cassandra. It provides design integrity, easier horizontal scaling, and control over opportunities in a range of devices. It uses data structures that are different from those concerning databases by default, which speeds up NoSQL calculations.

Facebook, Google, Twitter, and similar companies store user data on terabytes daily.

R Programming R is one of the open-source Big Data Technologies and programming languages. The free software Data . 4) EMERGING BIG DATA TECHNOLOGIES

Tensor Flow Tensor Flow has a robust, scalable ecosystem of resources, tools, and libraries for researchers, allowing them to quickly create and deploy powerful Machine Learning applications.

Beam Apache Beam offers a compact API layout to create sophisticated Parallel Data Processing pipelines through various Execution Engines or Runners. Apache Software Foundation developed these tools for Big Data in the year 2016.

Docker is one of the tools for Big Data that makes the development, deployment and running of container applications simpler. Containers help developers stack an application with all the necessary components, such as libraries and other dependencies.

Airflow Apache Airflow is a Process Management and Scheduling System for the management of data pipelines. Airflow utilizes job workflows made up of DAGs (Directed Acyclic Graphs) tasks. The code description of workflows makes it easy to manage, validate and version a large amount of Data.

Kubernetes Kubernetes is one of the open-source tools for Big Data developed by Google for vendor-agnostic



cluster and container management. It offers a platform for container systems' automation, deployment, escalation, and execution through host clusters.

Blockchain Blockchain is the Big Data technology that carries a unique data safe feature in the digital Bitcoin currency so that it is not deleted or modified after the fact is written. It's a highly secured environment and an outstanding option for numerous Big Data applications in various industries like banking, finance, insurance, medical, and retail, to name a few. Tools for Big Data will greatly minimize the effort of the end-users. Companies like Informatica have already shown innovations in this frontier. We can see more such Big Data technologies and more companies working towards such self-service solutions.

## **CONCLUSION**

To summarize, Big Data is still very much rising with more adoptions and more applications of the existing Big Data technologies and the launch of newer solutions related to Big Data security, Cloud integrations, data mining, etc.

## **New & older approaches**

Six big data analysis techniques

- A/B testing. ...
- Data fusion and data integration. ...
- Data mining. ...

- Machine learning. ...
- Natural language processing (NLP). ...
- Statistics
- Data analysis, or analytics (DA) is the process of examining data sets (within the form of text, audio and video), and drawing conclusions about the information they contain, more commonly through specific systems, software, and methods. Data analytics technologies are used on an industrial scale, across commercial business industries, as they enable organizations to make calculated, informed business decisions.<sup>5</sup>
- Globally, enterprises are harnessing the power of various different data analysis techniques and using it to reshape their business models.<sup>6</sup> As technology develops, new analysis software emerge, and as the Internet of Things (IoT) grows, the amount of data increases. Big data has evolved as a product of our increasing expansion and connection, and with it, new forms of extracting, or rather “mining”, data.
- Database technology eliminates many of the problems of traditional file organization by organizing data: centralizing data and controlling redundant data and serve many applications and different groups at the same time.
- include 1) the Command and Control approach, 2) the Traditional approach, and 3) the Non-Invasive approach. This article compares and contrasts the approaches and quickly summarizes each approach. Data Manipulation: providing a way to insert and update data in the database. Query

Execution: retrieving information from the data in the database. Data Integrity: ensuring that data stored is well-formed.

- While traditional data is based on a centralized database architecture, big data uses a distributed architecture. Computation is distributed among several computers in a network. This makes big data far more scalable than traditional data, in addition to delivering better performance and cost benefits.

- Big Data Analysis Techniques

- The global big data market revenues for software and services are expected to increase from \$42 billion to \$103 billion by year 2027.<sup>1</sup> Every day, 2.5 quintillion bytes of data are created, and it's only in the last two years that 90% of the world's data has been generated.<sup>2</sup> If that's any indication, there's likely much more to come.

- The world is driven by data, and it's being analysed every second, whether it's through your phone's Google Maps, your Netflix habits, or what you've reserved in your online shopping cart – in many ways, data is unavoidable and it's disrupting almost every known market.<sup>3</sup> The business world is looking to data for market insights and ultimately, to generate growth and revenue. Although data is becoming a game changer within the business arena, it's important to note that data is also being utilized by small businesses, corporate and creative alike. A global survey from McKinsey revealed that when organizations use data, it benefits the customer and the business by generating new data-driven

services, developing new business models and strategies, and selling data-based products and utilities.<sup>4</sup> The incentive for investing and implementing data analysis tools and techniques is huge, and businesses will need to adapt, innovate, and strategise for the evolving digital marketplace.

- Every day, 2.5 quintillion bytes of data are created, and it's only in the last two years that 90% of the world's data has been generated.

- What is data analysis?

- Data analysis, or analytics (DA) is the process of examining data sets (within the form of text, audio and video), and drawing conclusions about the information they contain, more commonly through specific systems, software, and methods. Data analytics technologies are used on an industrial scale, across commercial business industries, as they enable organizations to make calculated, informed business decisions.<sup>5</sup>

- Globally, enterprises are harnessing the power of various different data analysis techniques and using it to reshape their business models.<sup>6</sup> As technology develops, new analysis software emerge, and as the Internet of Things (IoT) grows, the amount of data increases. Big data has evolved as a product of our increasing expansion and connection, and with it, new forms of extracting, or rather “mining”, data.

- Six big data analysis techniques

- Big data is characterised by the three V's: the major volume of data, the velocity at which it's processed, and the wide

variety of data.<sup>7</sup> It's because of the second descriptor, velocity, that data analytics has expanded into the technological fields of machine learning and artificial intelligence. Alongside the evolving computer-based analysis techniques data harnesses, analysis also relies on the traditional statistical methods.<sup>9</sup> Ultimately, how data analysis techniques function within an organisation is twofold; big data analysis is processed through the streaming of data as it emerges, and then performing batch analysis' of data as it builds –to look for behavioural patterns and trends.<sup>10</sup> As the generation of data increases, so will the various techniques that manage it. As data becomes more insightful in its speed, scale, and

depth, the more it fuels innovation.

- The world is driven by data, and it's being analyzed every second, whether it's through your phone's Google Maps, your Netflix habits, or what you've reserved in your online shopping cart.
- McKinsey's big data report identifies a range of big data techniques and technologies, that draw from various fields such as statistics, computer science, applied mathematics, and economics.<sup>11</sup> As these methods rely on diverse disciplines, the analytics tools can be applied to both big data and other smaller datasets:
  - 1. A/B testing
  - This data analysis technique involves comparing a control group with a variety of test groups, in order to discern what treatments or changes will improve a given objective

variable. McKinsey gives the example of analysing what copy, text, images, or layout will improve conversion rates on an e-commerce site.<sup>12</sup> Big data once again fits into this model as it can test huge numbers, however, it can only be achieved if the groups are of a big enough size to gain meaningful differences.

- 2. Data fusion and data integration

- By combining a set of techniques that analyse and integrate data from multiple sources and solutions, the insights are more efficient and potentially more accurate than if developed through a single source of data.

- 3. Data mining

- A common tool used within big data analytics, data mining extracts patterns from large data sets by combining methods from statistics and machine learning, within database management. An example would be when customer data is mined to determine which segments are most likely to react to an offer.

- 4. Machine learning

- Well known within the field of artificial intelligence, machine learning is also used for data analysis. Emerging from computer science, it works with computer algorithms to produce assumptions based on data.<sup>14</sup> It provides predictions that would be impossible for human analysts.

- 5. Natural language processing (NLP).

- Known as a subspecialty of computer science, artificial intelligence, and linguistics, this data analysis tool uses algorithms to analyse human (natural) language.<sup>15</sup>

- Statistics.

- This technique works to collect, organise, and interpret data, within surveys and experiments.

- Other data analysis techniques include spatial analysis, predictive modelling, association rule learning, network analysis and many, many more. The technologies that process, manage, and analyze this data are of an entirely different and expansive field, that similarly evolves and develops over time. Techniques and technologies aside, any form or size of data is valuable. Managed accurately and effectively, it can reveal a host of business, product, and market insights. What does the future of data analysis look like? It's hard to say with the tremendous pace analytics and technology progresses, but undoubtedly data innovation is changing the face of business and society in its holistic entirety.

- Learn how to sort, analyse, and interpret data to inform business strategy with the UCT Data Analysis online short course.

- Share

- Big Data Analysis Techniques

- The global big data market revenues for software and services are expected to increase from \$42 billion to \$103 billion by year 2027.<sup>1</sup> Every day, 2.5 quintillion bytes of data

are created, and it's only in the last two years that 90% of the world's data has been generated. If that's any indication, there's likely much more to come.

- The world is driven by data, and it's being analyzed every second, whether it's through your phone's Google Maps, your Netflix habits, or what you've reserved in your online shopping cart – in many ways, data is unavoidable and it's disrupting almost every known market.<sup>3</sup> The business world is looking to data for market insights and ultimately, to generate growth and revenue. Although data is becoming a game changer within the business arena, it's important to note that data is also being utilized by small businesses, corporate and creative alike. A global survey from McKinsey revealed that when organizations use data, it benefits the customer and the business by generating new data-driven services, developing new business models and strategies, and selling data-based products and utilities.<sup>4</sup> The incentive for investing and implementing data analysis tools and techniques is huge, and businesses will need to adapt, innovate, and strategize for the evolving digital marketplace.

- Every day, 2.5 quintillion bytes of data are created, and it's only in the last two years that 90% of the world's data has been generated. Want To Know More About Our Programs?

India Afghanistan Aland

Islands Albania Algeria

Andorra Angola Anguilla.

## **Data Discovery**



Data discovery involves the collection and evaluation of data from various sources and is often used to understand trends and patterns in the data. It requires a progression of steps that organizations can use as a framework to understand their data. Data discovery, usually associated with business intelligence (BI), helps inform business decisions by bringing together disparate, siloed data sources to be analyzed. Having mounds of data is useless unless you find a way to extract insights from it. The data discovery process includes connecting multiple data sources, cleansing and preparing the data, sharing the data throughout the organization, and performing analysis to gain insights into business processes. Today, nearly all businesses collect huge amounts of data on their customers, markets, suppliers, production processes, and more. Data flows in from online and traditional transactions systems, sensors, social media, mobile devices, and other diverse sources. As a result, decision makers are drowning in data but starving for insights. Insights are hidden within that One approach that business data analysts use to uncover and investigate hidden but potentially Useful insights in data. It is a methodology for digging into data looking for interesting Relationships, trends, patterns, and anomalies requiring further exploration.

Exploration and Visual analytics enables the use of technology assisted analytical and pattern recognition software For visualization and drill-downs to turn data into knowledge and understanding. Data discovery offers businesses a way to make their data clean, easily

understandable, and user-Friendly. A comprehensive solution should be able to be used by all members of the business.

The main benefit of data discovery is the actionable insights that are uncovered in the data. These insights help users spot valuable opportunities before competitors without having to Consult the IT organization. Visual data discovery can enhance this value, allowing line of Business workers to find answers faster .Today, companies are finding that the use of artificial intelligence (AI) is greatly enhancing the Data discovery process. This process is also referred to as smart data discovery. In smart data Discovery, AI can automatically discover data relationships and accelerate a company's analyses With AI-powered recommendations. The underlying AI suggestion engine uses sophisticated AI Algorithms that run against any type of data without the user being aware that processing is Happening in the background. The AI engine identifies potential relationships such as Correlations by employing trained learning algorithms. Leading analytics platforms utilizing AI Offer recommended visualizations of related variables that users can choose to explore further. There are several exciting directions for innovation in the area of AI-powered analytics .

Including AI techniques can be used to suggest data preparation steps such as normalization, missing data handling, string pattern recognitions, and others.Algorithms can be used to identify and draw attention to particular patterns or outliers in the data for groups of related variables.Time series analysis has distinct needs and

techniques for pattern recognition, anomaly detection, and series relationships discovery. Behavioral data of expert users can be collected, analyzed, and used to influence recommended analysis actions.

AI suggestion engines and recommendations are increasingly used to augment analytics on an ever-expanding space of problems. This combination of human understanding with machine tirelessness enables business professionals to rapidly discover important relationships across vast amounts of data in time to take action. Solving Business Problems with Data Discovery Analysts are tasked with discovering insights in the massive amounts of data that businesses collect. Because it brings in data from so many different sources, data discovery enables businesses to use data in innovative ways.

It helps users explore data in new and different ways and to find insights that were not apparent prior to data discovery. And, once new trends or patterns are made, data discovery makes it easy for users to drill down into the variables and come up with new questions and insights. There are several exciting directions for innovation in the area of AI-powered analytics

Including AI techniques can be used to suggest data preparation steps such as normalization, missing data handling, string pattern recognitions, and others. Algorithms can be used to identify and draw attention to particular patterns or outliers in the data for groups of related variables. Time series analysis has distinct needs and

techniques for pattern recognition, anomaly detection, and series relationships discovery. Behavioral data of expert users can be collected, analyzed, and used to influence recommended analysis actions

AI suggestion engines and recommendations are increasingly used to augment analytics on an ever-expanding space of problems. This combination of human understanding with machine tirelessness enables business professionals to rapidly discover important relationships across vast amounts of data in time to take action. Solving Business Problems with Data Discovery Analysts are tasked with discovering insights in the massive amounts of data that businesses collect. Because it brings in data from so many different sources, data discovery enables businesses to use data in innovative ways. It helps users explore data in new and different ways and to find insights that were not apparent prior to data discovery. And, once new trends or patterns are made, data discovery makes it easy for users to drill down into the variables and come up with new questions and insights. These insights can include identifying customer problems such as the following

- Unexpected customer churn
- Customer relationship and management problems
- Subtle product issues such as returns and failures
- Price leakages due to excessive discounting
- Promotional failures

- Lost market share due to competitive actions such as aggressive pricing or a new product

Data discovery is enabling companies to capture a 360-degree view of their customers by compiling and assessing customer behavioral, transactional, and sentiment data across the many channels customers use to interact with companies. Data discovery is invaluable in helping decision makers detect early warning signs about customer dissatisfaction. Data discovery helps business leaders gain a more thorough understanding of how customers view the company. Text, sentiment, social, and speech analytics can be used to identify what customers are saying about your company across a variety of interactions, including social media comments and contact center interactions. Key word searches against customer sentiment can help business leaders identify where potential product or service problems may be coming to the fore with multiple customers.

Data discovery tools also offer banks myriad opportunities to learn more about their customers and act on these insights. For instance, data discovery tools can help bankers determine which products a particular customer is using (e.g., checking, savings) and then determine based on that customer's income, lifecycle status, and other factors whether she might be a good candidate for a cross-sell or upsell offer (e.g., certificate of deposit). With customer churn so high in financial services, bankers can also use data analysis and data discovery tools to determine the primary causes of customer defection among certain groups of

customers and also to spot the warning signs when a customer is about to jump ship.

Undetected and unaddressed, these problems can seriously undermine any business. Hence the urgency to find insights in the data and take action. With the right insights, companies can focus their efforts where they are needed to retain and delight customers rather than simply throwing customer-enticing tactics against the wall and see what sticks. Data discovery puts the power of big Data.

## **opensource technologies for Big data analytics**

The Primary Sources of Big Data: Machine Data. In-Demand Software Development Skills. Social Data. Transactional Data.

- GNU/Linux.
- Mozilla Firefox.
- VLC media player.
- SugarCRM.
- GIMP.
- VNC.
- Apache web server.
- LibreOffice.
- Open-source big data analytics refers to the use of open-source software and tools for Analyzing huge quantities of data in order to gather relevant and actionable information That an organization can use in order to further its business

goals. Open-source big data analytics refers to the use of open-source software and tools for analyzing huge quantities of data in order to gather relevant and actionable information that an organization can use in order to further its business goals. The biggest player in open-source big data analytics is Apache's Hadoop – it is the most widely used software library for processing enormous data sets across a cluster of computers using a distributed process for parallelism. Advertisement Techopedia Explains Open-Source Big Data Analytics Open-source big data analytics makes use of open-source software and tools in order to execute big data analytics by either using an entire software platform or various open-source tools for different tasks in the process of data analytics. Apache Hadoop is the most well-known system for big data analytics, but other components are required before a real analytics system can be put together.

Menu  
Toggler  
Techopedia Logo  
Search Icon  
Don't Miss an insight.  
Subscribe to Techopedia for free.  
Dictionary  
Data Management  
Open-Source Big Data Analytics  
Open-Source Big Data Analytics  
TABLE OF CONTENTS  
What Does Open-Source Big Data Analytics Mean?

Open-source big data analytics refers to the use of open-source software and tools for analyzing huge quantities of data in order to gather relevant and actionable information that an organization can use in order to further its business goals. The biggest player in open-source big data analytics is Apache's Hadoop – it is the most widely used software library for processing enormous data sets across a cluster of computers using a distributed process for parallelism. Advertisement Techopedia Explains Open-Source Big Data Analytics Open-source big data analytics makes use of open-source software and tools in order to execute big data analytics by either using an entire software platform or various open-source tools for different tasks in the process of data analytics. Apache Hadoop is the most well-known system for big data analytics, but other components are required before a real analytics system can be put together.

Hadoop is the open-source implementation of the MapReduce algorithm pioneered by Google and Yahoo, so it is the basis of most analytics systems today. Many big data analytics tools make use of open source, including robust database systems such as the open-source MongoDB, a sophisticated and scalable NoSQL database very suited for big data applications, as well as others.

Open-source big data analytics services encompass: Data collection system Control center for administering and monitoring clusters Machine learning and data mining library Application coordination service Compute engine Execution framework Parallelism.

## TABLE OF CONTENTS

### What Does Open-Source Big Data Analytics Mean ?

Open-source big data analytics refers to the use of open-source software and tools for analyzing huge quantities of data in order to gather relevant and actionable information that an organization can use in order to further its business goals. The biggest player in open-source big data analytics is Apache's Hadoop – it is the most widely used software library for processing enormous data sets across a cluster of computers using a distributed process for parallelism.

### Techopedia Explains Open-Source Big Data Analytics

Open-source big data analytics makes use of open-source software and tools in order to execute big data analytics by either using an entire software platform or various open-source tools for different tasks in the process of data analytics. Apache Hadoop is the most well-known system for big data analytics, but other components are required before a real analytics system can be put together.



Hadoop is the open-source implementation of the MapReduce algorithm pioneered by Google and Yahoo, so it is the basis of most analytics systems today. Many big data analytics tools make use of open source, including robust database systems such as the open-source MongoDB, a sophisticated and scalable NoSQL database very suited for big data applications, as well as others. Open-source big data analytics services encompass:

- Data collection system
- Control center for administering and monitoring clusters
- Machine learning and data mining library
- Application coordination service
- Compute engine
- Execution framework

Open source software (OSS) is software that is distributed with its source code, making it available for use, modification, and distribution with its original rights. Source code is the part of software that most computer users don't ever see; it's the code computer programmers manipulate to control how a program or application behaves. Programmers who have access to source code can change a program by adding to it, changing it, or fixing parts of it that aren't working properly. OSS typically includes a license that allows programmers to modify the software to best fit their needs and control how the software can be distributed.

## **clouds and big data**

Big Data refers to the large sets of data collected, while “Cloud Computing” refers to the mechanism that remotely takes this data in and performs any operations specified on that data. The public cloud has emerged as an ideal platform for big data. A cloud has the resources and services that a business can use on demand, and the business doesn't have to build, own or maintain the infrastructure. Thus, the cloud makes big data technologies accessible and affordable to almost any size of enterprise. Organizations to unify and connect to a single copy of all of their data with ease. The result is an ecosystem of thousands of businesses and organizations connecting to not only their own data, but also connecting to each other by effortlessly sharing and consuming shared data and data services. Big data refers to the data which is huge in size and also increasing rapidly with respect to time.

Big data includes structured data, unstructured data as well as semi-structured data. Big data can not be stored and processed in traditional data management tools it needs specialized big data management tools. It refers to complex and large data sets having 5 V's volume, velocity, Veracity, Value and variety information assets. It includes data storage, data analysis, data mining and data visualization. Examples of the sources where big data is generated includes social media data, e-commerce data, weather station data, IoT Sensor data etc. Characteristics of Big Data Variety of Big data – Structured, unstructured, and semi structured data Velocity of Big data – Speed of data generation Volume of Big data – Huge volumes of data that is

being generated Value of Big data – Extracting useful information and making it valuable Variability of Big data – Inconsistency which can be shown by the data at times.

Cloud computing refers to the on demand availability of computing resources over internet. These resources includes servers, storage, databases, software, analytics, networking and intelligence over the Internet and all these resources can be used as per requirement of the customer. In cloud computing customers have to pay as per use. It is very flexible and can be resources can be scaled easily depending upon the requirement. Instead of buying any IT resources physically, all resources can be availed depending on the requirement from the cloud vendors. Cloud computing has three service models i.e., Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

Examples of cloud computing vendors who provides cloud computing services are Amazon Web Service (AWS), Microsoft Azure, Google Cloud Platform, IBM Cloud Services etc .Characteristics of Cloud Computing On-Demand availability Accessible through a network Elastic Scalability Pay as you go model Multi-tenancy and resource pooling. Advantages of Cloud Computing Back-up and restore data Improved collaboration Excellent accessibility Low maintenance cost On-Demand Self-service. Disadvantages of Cloud Computing Vendor lock-in Limited Control Security Concern Downtime due to various reason Requires good Internet connectivity. Difference between Big Data and Cloud

Computing Big data refers to the data which is huge in size and also increasing rapidly with respect to time .Cloud computing refers to the on demand availability of computing resources over internet.

Big data includes structured data, unstructured data as well as semi-structured data .Cloud Computing Services includes Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).Volume of data, Velocity of data, Variety of data, Veracity of data, and Value of data are considered as the 5 most important characteristics of Big data.On-Demand availability of IT resources, broad network access, resource pooling, elasticity and measured service are considered as the main characteristics of cloud computing.

The purpose of big data is to organizing the large volume of data and extracting the useful information from it and using that information for the improvement of business.The purpose of cloud computing is to store and process data in cloud or availing remote IT services without physically installing any IT resources. Distributed computing is used for analyzing the data and extracting the useful information .Internet is used to get the cloud based services from different cloud vendors.Big data management allows centralized platform, provision for backup and recovery and low maintenance cost .Cloud computing services are cost effective, scalable and robust.Some of the challenges of big

data are variety of data, data storage and integration, data processing and resource management .Some of the challenges of cloud computing are availability, transformation, security

concern, charging model.Big data refers to huge volume of data, its management, and useful information extraction.Cloud computing refers to remote IT resources and different internet service models.

### **SHORT QUESTIONS**

- 1.Explain the data discovery?
- 2.Discuss open source technologies for big data analytics?
- 3.Explain clouds and big data?
- 4.Explain big data technology?
- 5.Write about new &old approaches?

### **LONG QUESTION**

- 1.What is big data technology?
- 2.Discuss the new&old approaches?
- 3.Explain data discovery in details?
- 4.Explain clouds ?

## **UNIT 3**

### **Business big data Consumption**

Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. These processes use familiar statistical analysis techniques—like clustering and



regression—and apply them to more extensive datasets with the help of newer tools. The four most popular types of business analytics are **descriptive, diagnostic, predictive, and prescriptive**.

The four most popular types of business analytics are descriptive, diagnostic, predictive, and prescriptive. The fifth—cognitive analytics is a new type that employs AI, ML, and [deep learning](#). Whilst each of these business analytics types is effective when used individually, they become extremely powerful when employed together.

## **Descriptive Analytics**

It analyses historical data to determine the response of a unit over a set of given variables. It tracks key performance indicators (KPIs) for a better understanding of the present state of a business. It involves the following five steps:

- Deciding which business metrics will effectively evaluate performance against objectives
- Identifying required data as per the current business state
- Collecting and preparing data using various processes like deduplication, transformation, and cleansing.
- Analyzing data for patterns to measure performance
- Presenting data in charts and graphs to make it understandable for non-analytics experts

### ***Examples of Descriptive Analytics***

- Summarizing past events, exchange of data, and social media usage
- Reporting general trends
- **Diagnostic Analytics**

Diagnostic Analytics is one of those business analytics types that help understand why things happened in the past. Using drill-downs, [data mining](#), data discovery, and correlations, you can comprehend the driving factors. This advanced analytics method is usually employed as a preceding step of Descriptive Analytics to

find the reasoning behind certain results in finance, marketing, cybersecurity, and more.

### ***Examples of Diagnostic Analytics***

- Examining market demand
- Identifying technical issues
- Explaining customer behavior
- Improving organization culture

### **Predictive Analytics**

It considers historical data trends for determining the probability of particular future outcomes. It uses several techniques like data mining, machine learning algorithms, and statistical modeling to forecast the likelihood of events.

Predictive analytics helps improve business areas, including customer service, efficiency, fraud detection and prevention, and risk management. It allows you to grow the most profitable customers, improve the operations of businesses, and determine customer responses and cross-sell opportunities.

### ***Examples of Predictive Analytics***

- Predicting customer preferences
- Detection of employee intentions
- Recommending products
- Predicting staff and resources



# Prescriptive Analytics

Prescriptive analytics generates recommendations to handle similar future situations relying on past performances. It employs several tools, statistics, and ML algorithms for the available internal data and external data. It gives you insights into what may happen, when, and why.

## ***Examples of Prescriptive Analytics***

- Tracking fluctuating manufacturing prices
- Improving equipment management
- Suggest the best course of action
- Price modeling
- Evaluating rates of readmission
- Identifying testing

## ***Business Analyst Master's Program Cognitive Analytics***

Combining [Artificial Intelligence](#) and [Data Analytics](#), Cognitive Analytics is one of the newest types of business analytics. It looks at the available data in the knowledge base and discovers the best solutions for the questions posed. Cognitive analytics covers multiple analytical techniques to analyze large data sets and monitor customer behavior patterns and emerging trends.

## ***Examples of Cognitive Analytics***

- Tapping unstructured data sources such as images, text documents, emails, and social posts.

## **Business Analytics**

Top companies choose different types of business analytics. Often they employ several types of business analytics in a step-like process, starting from Descriptive Analytics and concluding with Prescriptive Analytics.

- Amazon uses descriptive and predictive analytics of customers' historical shopping data for the prediction of the probability of a customer buying a product. It also uses these methods to personalize product recommendations.
- Microsoft uses prescriptive and predictive analytics to improve productivity and collaboration.
- Uber uses predictive modeling to estimate demand in real-time and has enhanced its customer support.
- Starbucks also benefits from predictive analytics to predict purchases and propose interesting offers.
- Apple's Siri, Microsoft's Cortana, and IBM's Watson use cognitive analysis.

## **Types of Business Analytics**

Every business analytics type plays a significant role depending on the requirements. However, prescriptive analytics is one of the most important types and is thus opted for by most companies. Descriptive analysis is the most suitable if you are aiming at analyzing the everyday reporting for your businesses.

When making assessments for future situations using ML and deep learning, use predictive analytics as it is a more advanced method. To estimate the best possible options, opt for prescriptive analytics to get actionable insights instead of data monitoring. It best suits healthcare decision-makers needs in optimizing and reducing production costs. When it comes to social media campaign analytics and other digital marketing analytics, diagnostic analytics helps view what works and what doesn't for your campaigns.

You can use these four techniques sequentially, or you can jump directly to prescriptive analytics if you have identified the key area that requires optimization to reach the desired outcome. Looking into the business strategies of top companies reveals that prescriptive and cognitive analytics are the front-runners in this spectrum.

# Business Analytics Tools

From a simple spreadsheet with statistical functions to complex predictive modeling applications and data mining, business analytics tools enable users to gain deeper insights with much-needed accuracy. Business analytics tools help analyze various business reports and data, generating the best possible outcome for users. For instance, OmniSci is a business analytics tool. It enables users to interactively query, visualize, and power Data Science workflows across massive data. Some other tools used by [Business Analysts](#) are as follows:

- Jira
- Confluence
- Trello
- Balsamiq
- Microsoft Visio
- Google Docs and Spreadsheets
- Rational Requisite Pro

Master the Fundamentals of Business Analysis [ENROLL NOW](#)

## Business Analytics Jobs

According to Glassdoor, a business analyst earns an average of \$68,346 annually. With thorough knowledge of business analytics, you can compete for the following job positions:

- [Business Analyst](#)
- Business Analyst Manager
- Data Analysis Scientist
- Data Business Analyst
- Information Security Analyst
- Quantitative Analyst
- IT Business Analyst

Learn best business analysis techniques by Purdue University, IB and EY experts. Sign-up for our [Professional Certificate Program in Business Analysis](#) TODAY!

### Conclusion

Despite the growing demand for business analytics, fewer candidates choose this career path. To thrive in today's business world, you must both understand and speak data. Enroll in our [Professional Certificate Program in Business Analysis](#) to acquire the in-demand skills and accelerate your journey to success.

## Data visualization and organization

Big data and organization of their visualization there: Introduction to Big Data Visualization For

analyzing big data efficiently, the most important thing is to choose perfect visualization tools. A perfect visualization tool will generate an efficient visual diagram which will lead to a correct decision. Insufficient visualization will lead to a loss for the organization. On Facebook, 4 petabytes of data are uploaded per day that contains different information like video, images, or textual information. Without visualizing those, it's hard to understand patterns and other relevant information.

- Facebook uses "[HiPlot](#)" to analyze and visualize it
- Another large organization IBM uses "[Big SQL](#)" integrated with other visualization tools like zeppelin notebooks, Data Science, [Tableau](#), and Cognos. Amazon uses AMZ Base, Amylaze, SellerApp, etc.

The selection of efficient big-data visualization tools will help change complex and extensive volume data into simple and human-readable visual diagrams. This visual diagram helps analysts predict more accurately that it will lead to business improvement.


Organization:

Big data is a term that describes large, hard-to-manage volumes of data – both structured and unstructured – that inundate businesses on a day-to-day basis. But it's not just the type or amount of data that's important, it's what organizations do with the data that matters.


**Here are 11 tips for making the most of your large data sets.**

1. Cherish your data. “Keep your raw data raw: don't manipulate it without having a copy,” says Teal. ...
2. Visualize the information.
3. Show your workflow. ...
4. Use version control. ...
5. Record metadata. ...
6. Automate, automate, automate. ...
7. Make computing time count. ...
8. Capture your environment.

What are its challenges?



# A Beginner's Guide to **Big Data** **Visualization** Tools and its Techniques



It is a large volume, complex dataset. So, such data can not visualize with the traditional method as the traditional method has many limitations.

- **Perceptual Scalability:** Human eyes cannot extract all relevant information from a large



volume of data. Even sometimes desktop screen has its limitations if the dataset is large. Too many visualizations are not always possible to fit on a single screen.

- **Real-time Scalability:** It is always expected that all information should be real-time information, but it is hardly possible as processing the dataset needs time.
- **Interactive scalability:** Interactive [data visualization](#) help to understand what is inside the datasets, but as its volume increases exponentially, visualizing the datasets take a long time. But the challenge is sometimes, and the system may freeze or crash while trying to visualize the datasets.

## Big Data

It contain a large volume of data with great variety, and this dataset increases its velocity exponentially. It could be structured, unstructured, or semi-structured. Managing it is a very tedious task. With time it popularity increases as we show interest in extracting information from that data. The volume of it increases exponentially with time. It cannot be stored in our traditional database. As per information, 720,000 hours of data have been uploaded on youtube. According to a survey, most of the data is unexplored. Every organization speeds up its analysis of it to find new opportunities for the development of the company. It will reduce costs and will increase company profit.

critical thinking

Gather complete information.

- Understand and define all terms.
- Question the methods by which the facts are derived.
- Question the conclusions.
- Look for hidden assumptions and biases.
- Question the source of facts.
- Don't expect all of the answers.
- Examine the big picture.



•  
•

- It describes this moment where we get so lost in little details that we fail to look at the overall picture and miss delivering what the business wants. Critical thinking implies that you interrogate data from a wider variety of perspectives to form a more objective analysis of the problem and a deeper understanding of what it is you are trying to solve before shaping an opinion. The must-have skill
- So, as a data professional, critical thinking can help you ask the right questions and focus solely on facts while keeping your gut feeling aside. It encourages you to broaden your perspective while collecting data and challenge the relevancy of it for the problem you are solving. Last but not least, critical thinking will help you constantly remain curious, which can lead to more innovative solutions.
- All right. So, how to work on it?
- Be willing to be wrong.
- The difficulty with critical thinking is that it requires accepting being wrong or, at least, a certain level of skepticism about your truth. It implies a willingness to fail, and let's be honest: nobody likes that. At best, it's uncomfortable. At worst, it's painful.
- It does NOT require from you any particular talent. Anybody can be wrong.
- What you need is to grow your acceptance of being vulnerable and a more profound sense of humility.

It's not fun or enjoyable, but it's essential for your personal growth. Take that path, and you will surely and shortly feel incredibly empowered!



. Critical Thinking Can Be Defined As...

The systemic evaluation or formulation of beliefs, or statements, by rational standards

A set of information and beliefs, generating and processing skills, and the habit of using those skills to guide behavior

## Critical thinkers:

- Ask questions
- Gather relevant information
- Think through solutions and conclusions
- Consider alternative systems of thought
- Communicate effectively

They're willing to admit when they're wrong or when they don't know the answer, rather than digging into a gut reaction or emotional point of view. This is why I have come up with the 90/10 rule – When working with data, 90% of your time should be spent on a structured strategic approach, while 10% of your time should be spent “exploring” the data. **Bernard Marr**

Expand search





## Big Data: The All-Important 90/10 Rule



**Bernard Marr**

Bernard Marr

 *Internationally Best-selling...*

Published May 6, 2015

[\*\*+ Follow\*\*](#)

In this post I outline my 90/10 rule for big data initiatives in businesses. The [post](#) was first published in my column for Data Science Central. The phenomenon of Big Data is giving us ever-growing volume and variety of data we which we can now store and analyze. Any regular reader of my posts knows that I personally prefer to focus on Smart Data, rather than Big Data - because the term places

too much importance on the size of the data. The real potential for revolutionary change comes from the ability to manipulate, analyze and interpret new data types in ever-more sophisticated ways.

## **The SMART Data Framework**

I've written previously about my SMART Data framework which outlines a step-by-step approach to delivering data-driven insights and improved business performance.

- 1. Start with strategy:** Formulate a plan – based on the needs of your business
- 2. Measure metrics and data:** Collect and store the information you need
- 3. Apply analytics:** Interrogate the data for insights and build models to test theories
- 4. Report results:** Present the findings of your analysis in a way that the people who will put them into effect will understand
- 5. Transform your business:** Understand your customers better, optimize business processes, improve staff wellbeing or increase revenues and profits. My work involves helping businesses use data to drive business value. Because of this I get to see a lot of half-finished data projects, mothballed when it was decided that external help was needed. The biggest mistake by far is putting insufficient thought – or neglecting to put any thought – into a structured strategic approach to big data projects. Instead of starting with strategy, too many companies

start with the data. They start frantically measuring and recording everything they can in the belief that big data is all about size. Then they get lost in the colossal mishmash of everything they've collected, with little idea of how to go about mining the all-important insights.

This is why I have come up with the 90/10 rule – When working with data, 90% of your time should be spent on a structured strategic approach, while 10% of your time should be spent “exploring” the data.

## **The 90/10 Rule**

The 90% structured time should be used putting the steps outlined in the SMART Data framework into operation. Making a logical progression through an ordered set of steps with a defined beginning (a problem you need to solve), middle (a process) and an ending (answers or results).

This is after all why we call it Data Science. Business data projects are very much like scientific experiments, where we run simulations testing the validity of theories and hypothesis, to produce quantifiable results. The other 10% of your time can be spent freely playing with your data – mining for patterns and insights which, while they may be valuable in other ways, are not an integral part of your SMART Data strategy.

Yes, you can be really lucky and your data exploration can deliver valuable insights – and who knows what you might find, or what inspiration may come to you? But it should always play second-fiddle to following



the structure of your data project in a methodical and comprehensive way.

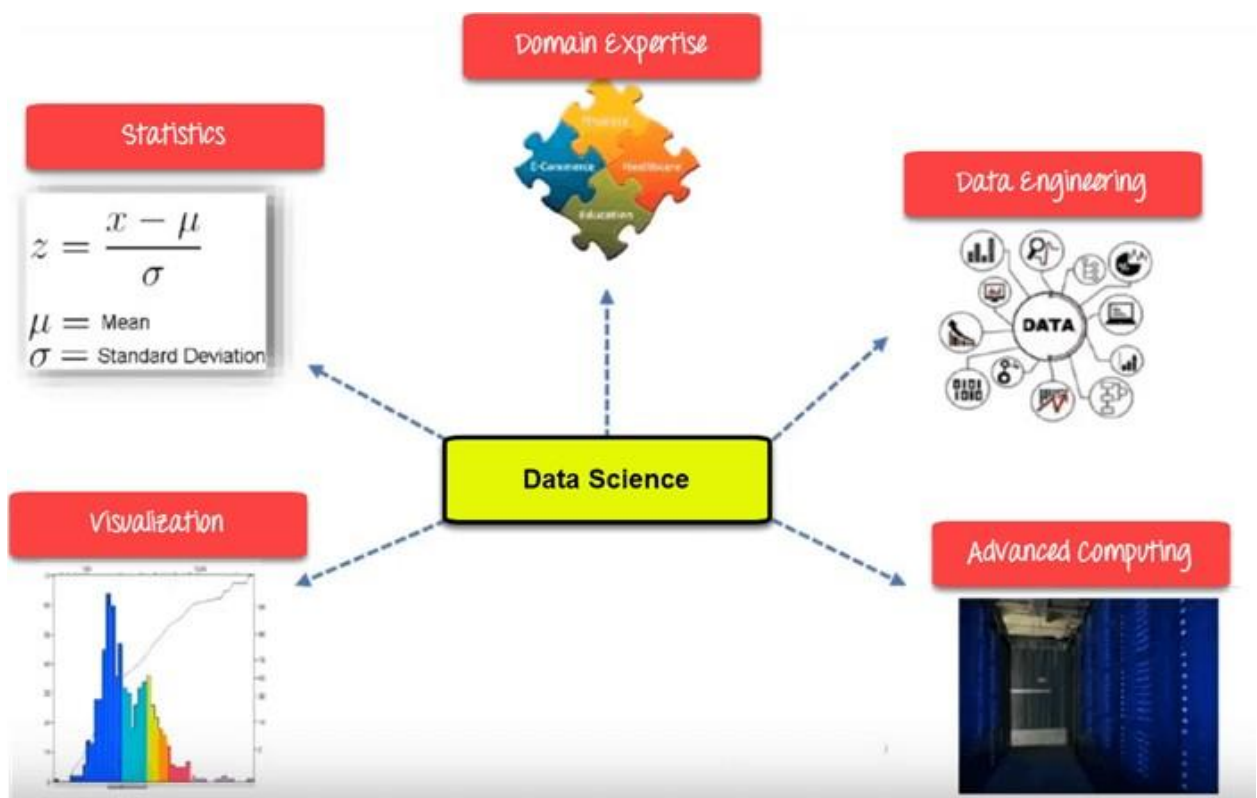
## **Always start with strategy**

I think this is a very important point to make, because it's something I often see companies get the wrong way round. Too often, the data is taken as the starting point, rather than the strategy. Businesses that do this run the very real risk of becoming "data rich and insight poor". They are in danger of missing out on the hugely exciting benefits that a properly implemented and structured data-driven initiative can bring.

Working in a structured way means "Starting with strategy", which means identifying a clear business need and what data you will need to solve it. Businesses that do this, and follow it through in a methodical way will win the race to unearth the most valuable and game-changing

## **sciences and learning knowledge**

Big Data sciences and learning knowledge are Data science is an interdisciplinary field focused on extracting knowledge from typically large data sets and applying the knowledge and insights from that data to solve problems in a wide range of application domains.



## Data Analyst:

**Role:** A data analyst is responsible for mining vast amounts of data. They will look for relationships, patterns, trends in data. Later he or she will deliver compelling reporting and visualization for analyzing the data to take the most viable business decisions.

**Languages:** R, Python, HTML, JS, C, C++ , SQL

Statistician:

**Role:** The statistician collects, analyses, and understands qualitative and quantitative data using statistical theories and methods.

**Languages:** SQL, R, Matlab, Tableau, Python, Perl, Spark, and Hive

Data Administrator:

**Role:** Data admin should ensure that the [database](#) is accessible to all relevant users. He also ensures that it is performing correctly and keeps it safe from [hacking](#).

**Languages:** Ruby on Rails, SQL, Java, C#, and Python

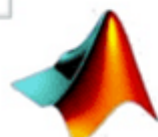
Business Analyst:

**Role:** This professional needs to improve business processes. He/she is an intermediary between the business executive team and the IT department.

**Languages:** SQL, Tableau, Power BI and, Python

Also, read Data Science Interview Questions and Answers: [Click Here](#)

Tools for Data Science



MATLAB

## privacy and security of big data

Big data privacy involves **properly managing big data to minimize risk and protect sensitive data**. Because big data comprises large and complex data sets, many traditional privacy processes cannot handle the scale and velocity required.

Big data privacy entails effectively handling big data to minimize risk and secure sensitive data. **Because big data consists of enormous and complex data sets, many standard privacy mechanisms cannot keep up with the requisite scale and velocity.**

Big data analytics have made it **impossible to anonymize data**. Big data analytics provide a wide range of data sets, so there is a chance that consumers might lose their identification factors. It's impossible to achieve anonymity because every SME relies on accounting and finance software hosted by a third party. Principles of **Transparency, Legitimate Purpose and Proportionality**. The processing of personal data shall be allowed subject to adherence to the principles of transparency, legitimate purpose, and proportionality.

### Security and Privacy Issues of Big Data

José Moura<sup>1,2</sup>, Carlos Serrão<sup>1</sup>

<sup>1</sup> ISCTE-IUL, Instituto Universitário de Lisboa, Portugal

<sup>2</sup> IT, Instituto de Telecomunicações, Lisboa, Portugal

{jose.moura, carlos.serrao}@iscte.pt

## ABSTRACT

This chapter revises the most important aspects in how computing infrastructures should be configured and intelligently managed to fulfill the most notably security aspects required by Big Data applications. One of them is privacy. It is a pertinent aspect to be addressed because users share more and more personal data and content through their devices and computers to social networks and public clouds. So, a secure framework to social networks is a very hot topic research. This last topic is addressed in one of the two sections of the current chapter with case studies. In addition, the traditional mechanisms to support security such as firewalls and demilitarized zones are not suitable to be applied in computing systems to support Big Data. SDN is an emergent management solution that could become a convenient mechanism to implement security in Big Data systems, as we show through a second case study at the end of the chapter. This also discusses current relevant work and identifies open issues.

Keywords: Big Data, Security, Privacy, Data Ownership, Cloud, Social Applications, Intrusion Detection, Intrusion Prevention.

## INTRODUCTION

The Big Data is an emerging area applied to manage datasets whose size is beyond the ability of commonly used software tools to capture, manage, and timely analyze that amount of data. The quantity of data to be analyzed is expected to

double every two years (IDC, 2012). All these data are very often unstructured and from various sources such as social media, sensors, scientific applications, surveillance, video and image archives, Internet search indexing, medical records, business transactions and system logs. Big data is gaining more and more attention since the number of devices connected to the so-called “Internet of Things” (IoT) is still increasing to unforeseen levels, producing large amounts of data which needs to be transformed into valuable information. Additionally, it is very popular to buy on-demand additional computing power and storage from public cloud providers to perform intensive data-parallel processing.

In this way, security and privacy issues can be potentially boosted by the volume, variety, and wide area deployment of the system infrastructure to support Big Data applications. As Big Data expands with the help of public clouds, traditional security solutions tailored to private computing infrastructures, confined to a well-defined security perimeter, such as firewalls and demilitarized zones (DMZs) are no more effective. Using Big Data, security functions are required to work over the heterogeneous composition of diverse hardware, operating systems, and network domains.

In this puzzle-type computing environment, the abstraction capability of Software-Defined Networking (SDN) seems a very important characteristic that can enable the efficient deployment of Big Data secure services on-top of the heterogeneous infrastructure. SDN introduces abstraction because it separates the control (higher) plane

from the underlying system infrastructure being supervised and controlled. Separating a network's control logic from the underlying physical routers and switches that forward traffic allows system administrators to write high-level control programs that specify the behavior of an entire network, in contrast to conventional networks, whereby administrators (if allowed to do it by the device manufacturers) must codify functionality in terms of low-level device configuration. Using SDN, the intelligent management of secure functions can be implemented in a logically centralized controller, simplifying the following aspects: implementation of security rules; system (re)configuration; and system evolution.

The robustness drawback of a centralized SDN solution can be mitigated using a hierarchy of controllers and/or through the usage of redundant controllers at least for the most important system functions to be controlled. The National Institute of Standards and Technology (NIST) launched very recently a framework with a set of voluntary guidelines to help organizations make their communications and computing operations safer (NIST, 2014). This could be achieved through a systematic verification of the system infrastructure in terms of risk assessment, protection against threats, and capabilities to respond and recover from attacks. Following the last verification principles, Defense Advanced Research Projects Agency (DARPA) is creating a program called Mining and Understanding Software Enclaves (MUSE) to enhance the quality of the US military's software. This program is designed to produce more robust software that can work with big datasets without causing errors or crashing

under the sheer volume of information (DARPA, 2014). In addition, security and privacy are becoming very urgent Big Data aspects that need to be tackled (Agrawal, Das, & El Abbadi, 2011). To illustrate this, the social networks have enabled people to share and distribute valuable copyrighted digital contents in a very easy way. Consequently, the copyright infringement behaviors, such as illicit copying, malicious distribution, unauthorized access and usage, and free sharing of copyright-protected digital contents, will become a much more common phenomenon.

To mitigate these problems, Big Data should have solid solutions to support author's privacy and author's copyrights (Marques & Serrão, 2013a). Also, users share more and more personal data and user generated content through their mobile devices and computers to social networks and cloud services, losing data and content control with a serious impact on their own privacy. Finally, one potentially promising approach is to create additional uncertainty for attackers by dynamically changing system properties in what is called a cyber moving target (MT) (Okhravi, Hobson, Bigelow, & Streilein, 2014). They present a summary of several types of MT techniques, consider the advantages and weaknesses of each, and make recommendations for future research in this area.

The current chapter endorses the most important aspects of Big Data security and privacy and is structured as follows. The first section discusses the most important challenges to the aspects of information security and privacy imposed by the novel requirements of Big Data



applications. The second section presents and explains some interesting solutions to the problems found in the previous section.

The third and fourth sections are related with two case studies in this exciting emergent area.

## **BIG DATA CHALLENGES TO INFORMATION SECURITY AND PRIVACY**

With the proliferation of devices connected to the Internet and connected to each other, the volume of data collected, stored, and processed is increasing everyday, which also brings new challenges in terms of the information security. In fact, the currently used security mechanisms such as firewalls and DMZs cannot be used in the Big Data infrastructure because the security mechanisms should be stretched out of the perimeter of the organization's network to fulfill the user/data mobility requirements and the policies of BYOD (Bring Your Own Device).

Considering these new scenarios, the pertinent question is what security and privacy policies and technologies are more adequate to fulfill the current top Big Data privacy and security demands (Cloud Security Alliance, 2013). These challenges may be organized into four Big Data aspects such as infrastructure security (e.g. secure distributed computations using MapReduce), data privacy (e.g. data mining that preserves privacy/granular access), data management (e.g. secure data provenance and storage) and, integrity and reactive security (e.g. real time monitoring of

anomalies and attacks).

Considering Big Data there is a set of risk areas that need to be considered. These include the information lifecycle (provenance, ownership and classification of data), the data creation and collection process, and the lack of security procedures. Ultimately, the Big Data security objectives are no different from any other data types – to preserve its confidentiality, integrity and availability.

## **SHORT QUESTION**

- 1.What is business big data, consumption data?
- 2.What Business Analytics type do different companies prefer?
- 3.What types of Business Analytics are right for you?
4. Decide sciences and learning knowledge of data?
- 5.What is data Visualization ?
- 6.What is Organization?

## **LONG QUESTIONS**

- 1.What is data visualization and organization?
- 2.Explain the rules of critical thinking?
- 3.Describe the privacy and security of big data?
- 4.What is business big data in details?

## **UNIT 4**



## **predictive analytics and Target**

Predictive

analytics is a branch of advanced analytics that makes predictions about future outcomes using

**historical data combined with statistical modeling, data mining techniques and machine learning.** Companies employ predictive analytics to find patterns

in this data to identify risks and opportunities.



# Bigdata Analytics

Data is emerging as the world's newest resource for competitive advantage among nations, organizations and business.



**1** Big data has few key characteristics such as volume, sources, velocity, variety and veracity.

**2** Along with the volume, the number of sources, from where the data is extracted are also growing.

**3** Data is increasingly accelerating the velocity at which it is created, as the process are moved from batch to a real time business.

**4** The demands of the business from these data also has increased, from an answer next week to an answer in a minute.

requires text mining to analyze the data.

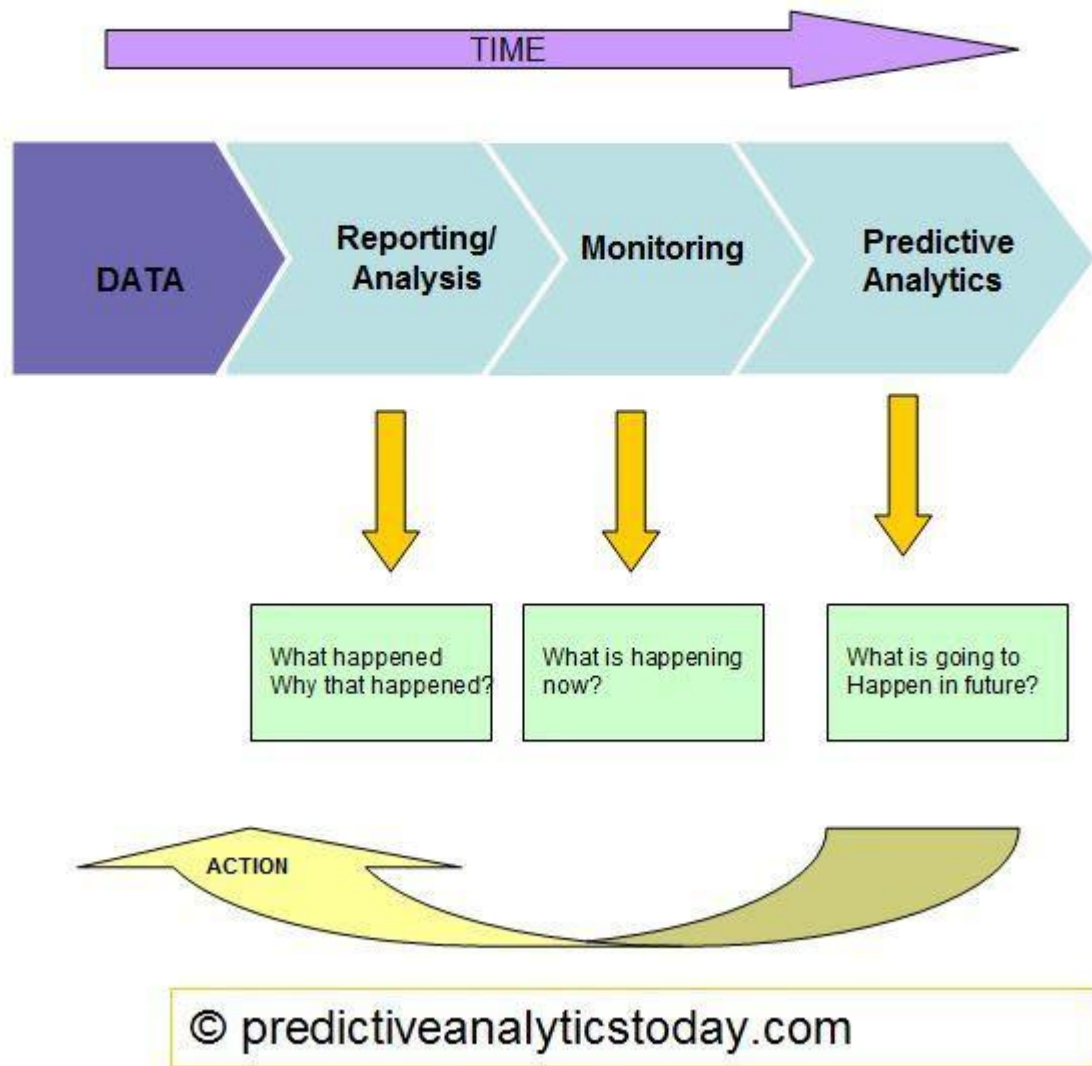
The business data is also growing at these same exponential rate too. Along with the volume, the number of sources, from where the data is extracted are also growing. Data is increasingly accelerating the velocity at which it is created, as the process are moved from batch to a real time business. The demands of the business from these data also has increased, from an answer next week to an answer in a minute.

### Bigdata Platforms and Bigdata Analytics Software

#### Big data Analytics

Business intelligence (BI) provides OLAP based, standard business reports, ad hoc reports on past data. These ad hoc analysis looks at the static past of data. This has its purpose and business uses, but does not meet the needs of a forward looking business. Forward looking big data analytics requires statistical analysis, statistical forecasting, casual analysis, optimization, predictive modeling and text mining on the large chunk of data available. There are performance issues, when these high volume past data are used in the relational data model, for a forward looking big data analytics, for future in the current system landscape in many organizations.

# Predictive Analytics



## Predictive Analytics Value Chain

Big Data Analytics will help organizations in providing an overview of the drivers of their business by introducing big data technology into the organization. This is the application of advanced analytic techniques to a very large data sets. These can not be achieved by standard data warehousing applications. These technologies are hadoop, mapreduce, massively parallel



processing databases, in memory database, search based applications, data-mining grids, distributed file systems, distributed databases, cloud etc.

The Technology drivers for Big data Analytics

- Multi core processors
- Lower power consumption
- Low cost storage
- High speed local networking

**The information subject to a given process, typically including most or all information on a piece of storage media.**

## Target Definition

Machine Learning methods can be classified into two broad categories: supervised and un-supervised. Supervised learning learns from labelled set of observations, where observations are known to belong to certain classes (for classification problems) or have certain values (regression problem). Un-supervised learning learns from unlabelled set of observations, where nothing else is known apart from observations themselves.

## logistics Regression

Logistic regression is a **statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.** A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. There are three

main types of logistic regression: **binary, multinomial and ordinal**. They differ in execution and theory.

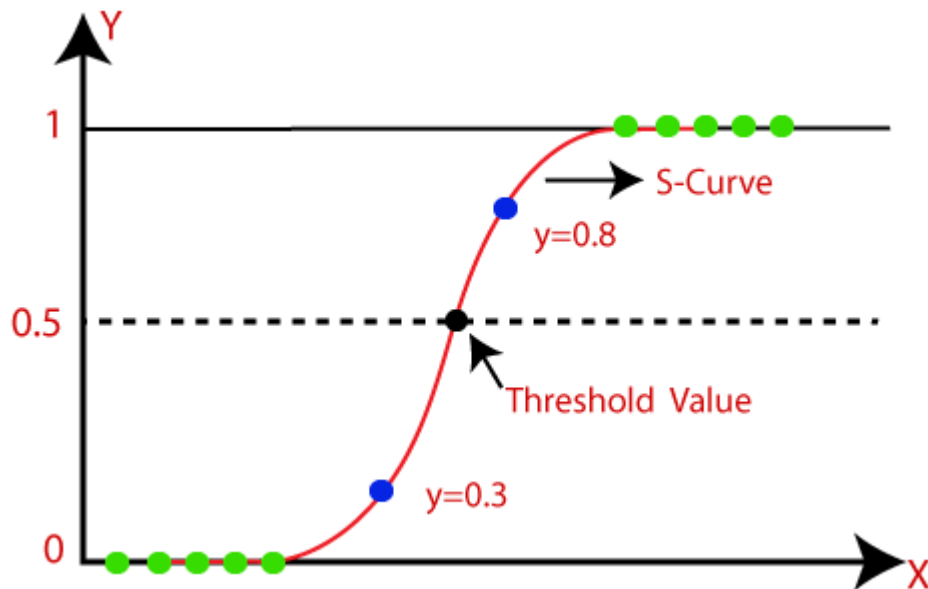
Logistic Regression is one of the basic and popular algorithms to solve a classification problem. It is named 'Logistic Regression' because its underlying technique is quite the same as Linear Regression. The term "Logistic" is taken from the Logit function that is used in this method of classification.

## Logistic Regression in Machine Learning

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1.**
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems.**
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability

to provide probabilities and classify new data using continuous and discrete datasets.

- Logistic Regression can be used to classify the observations using different types of data and can easily



determine the most effective variables used for the classification. The below image is showing the logistic function:

## Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

- In Logistic Regression  $y$  can be between 0 and 1 only, so for this let's divide the above equation by  $(1-y)$ :

$$\frac{y}{1-y}; 0 \text{ for } y=0, \text{ and infinity for } y=1$$

- But we need range between  $-\infty$  to  $+\infty$ , then take logarithm of the equation it will become:

$$\log \left[ \frac{y}{1-y} \right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

## Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

- **Binomial:** In binomial Logistic regression, there can be only two possible types of the

dependent variables, such as 0 or 1, Pass or Fail, etc.

- **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

**Python Implementation of Logistic Regression (Binomial)** To understand the implementation of Logistic Regression in Python, we will use the below example:

**Example:** There is a dataset given which contains the information of various users obtained from the social networking sites. There is a car making company that has recently launched a new SUV car. So the company wanted to check how many users from the dataset, wants to purchase the car.

For this problem, we will build a Machine Learning model using the Logistic regression algorithm. The dataset is shown in the below image. In this problem, we will predict the **purchased variable (Dependent Variable)** by using **age and salary (Independent variables)**.

**Steps in Logistic Regression:** To implement the Logistic Regression using Python, we will use the same steps as we have done in previous topics of Regression. Below are the steps:

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

**1. Data Pre-processing step:** In this step, we will pre-process/prepare the data so that we can use it in our code efficiently. It will be the same as we have done in Data pre-processing topic. The code for this is given below:

```
1. #Data Pre-processing Step
2. # importing libraries
3. import numpy as nm
4. import matplotlib.pyplot as mtp
5. import pandas as pd
6.
7. #importing datasets
8. data_set= pd.read_csv('user_data.csv')
```

By executing the above lines of code, we will get the dataset as the output. Consider the given image:

Index	User ID	Gender	Age	EstimatedSalary	Purchased
92	15809823	Male	26	15000	0
150	15679651	Female	26	15000	0
43	15792008	Male	30	15000	0
155	15610140	Female	31	15000	0
32	15573452	Female	21	16000	0
180	15685576	Male	26	16000	0
79	15655123	Female	26	17000	0
40	15764419	Female	27	17000	0
128	15722758	Male	30	17000	0
58	15642885	Male	22	18000	0
29	15669656	Male	31	18000	0
13	15704987	Male	32	18000	0
74	15592877	Male	32	18000	0
0	15624510	Male	19	19000	0

Format    Resize    ☒ Background color    ☒ Column min/max    Save and Close    Close

Now, we will extract the dependent and independent variables from the given dataset. Below is the code for it:

1. #Extracting Independent and dependent Variable
2. `x= data_set.iloc[:, [2,3]].values`
3. `y= data_set.iloc[:, 4].values`

In the above code, we have taken [2, 3] for x because our independent variables are age and salary, which are at index 2, 3. And we have taken 4 for y variable because our dependent



variable is at index 4. The output will be:

	0	1
0	19	19000
1	35	20000
2	26	43000
3	27	57000
4	19	76000
5	27	58000
6	27	84000
7	32	150000
8	25	33000
9	35	65000
10	26	80000
11	26	52000
12	20	86000

	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	0
9	0
10	0
11	0
12	0

Now we will split the dataset into a training set and test set. Below is the code for it:

1. # Splitting the dataset into training and test set.
2. from sklearn.model\_selection **import** train\_test\_split
3. x\_train, x\_test, y\_train, y\_test= train\_test\_split(x, y, test\_size= 0.25, random\_state=0)

The output for this is given below:

**For  
set:**

**test**

	0	1
0	-0.804802	0.504964
1	-0.0125441	-0.567782
2	-0.309641	0.157046
3	-0.804802	0.273019
4	-0.309641	-0.567782
5	-1.1019	-1.43758
6	-0.70577	-1.58254
7	-0.210609	2.15757
8	-1.99319	-0.0459058
9	0.878746	-0.770734
10	-0.804802	-0.596776
11	-1.00287	-0.422817
12	-0.111576	-0.422817

	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	0
9	0
10	0
11	0
12	0

**For training set:**

	0	1
0	-0.804802	0.504964
1	-0.0125441	-0.567782
2	-0.309641	0.157046
3	-0.804802	0.273019
4	-0.309641	-0.567782
5	-1.1019	-1.43758
6	-0.70577	-1.58254
7	-0.210609	2.15757
8	-1.99319	-0.0459058
9	0.878746	-0.770734
10	-0.804802	-0.596776
11	-1.00287	-0.422817
12	-0.111576	-0.422817

	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	0
9	0
10	0
11	0
12	0

In logistic regression, we will do feature scaling because we want accurate result of predictions. Here we will only scale the independent variable because dependent variable have only 0 and 1 values. Below is the code for it:

1. #feature Scaling
2. from sklearn.preprocessing **import** Standard Scaler
3. st\_x= StandardScaler()
4. x\_train= st\_x.fit\_transform(x\_train)
5. x\_test= st\_x.transform(x\_test)

The scaled output is given below:

	0	1
0	-0.804802	0.504964
1	-0.0125441	-0.567782
2	-0.309641	0.157046
3	-0.804802	0.273019
4	-0.309641	-0.567782
5	-1.1019	-1.43758
6	-0.70577	-1.58254
7	-0.210609	2.15757
8	-1.99319	-0.0459058
9	0.878746	-0.770734
10	-0.804802	-0.596776
11	-1.00287	-0.422817
12	-0.111576	-0.422817

	0	1
0	0.581649	-0.88
1	-0.606738	1.46
2	-0.0125441	-0.56
3	-0.606738	1.89
4	1.37391	-1.46
5	1.47294	0.997
6	0.0864882	-0.79
7	-0.0125441	-0.24
8	-0.210609	-0.56
9	-0.210609	-0.19
10	-0.309641	-1.29
11	-0.309641	-0.56
12	0.383585	0.099

## 2. Fitting Logistic Regression to the Training set:

We have well prepared our dataset, and now we will train the dataset using the training set. For providing training or fitting the model to the training set, we will import the **LogisticRegression** class of the **sklearn** library.

After importing the class, we will create a classifier object and use it to fit the model to the

logistic regression. Below is the code for it:

```
1. #Fitting Logistic Regression to the training set
2. from sklearn.linear_model import LogisticRegression
3. classifier= LogisticRegression(random_state=0)
4. classifier.fit(x_train, y_train)
```

**Output:** By executing the above code, we will get the below output:

**Out[5]:**

```
1. LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
2.                    intercept_scaling=1, l1_ratio=None, max_iter=100,
3.                    multi_class='warn', n_jobs=None, penalty='l2',
4.                    random_state=0, solver='warn',
5.                    tol=0.0001, verbose=0, warm_start=False)
```

Hence our model is well fitted to the training set.

### **3. Predicting the Test Result**

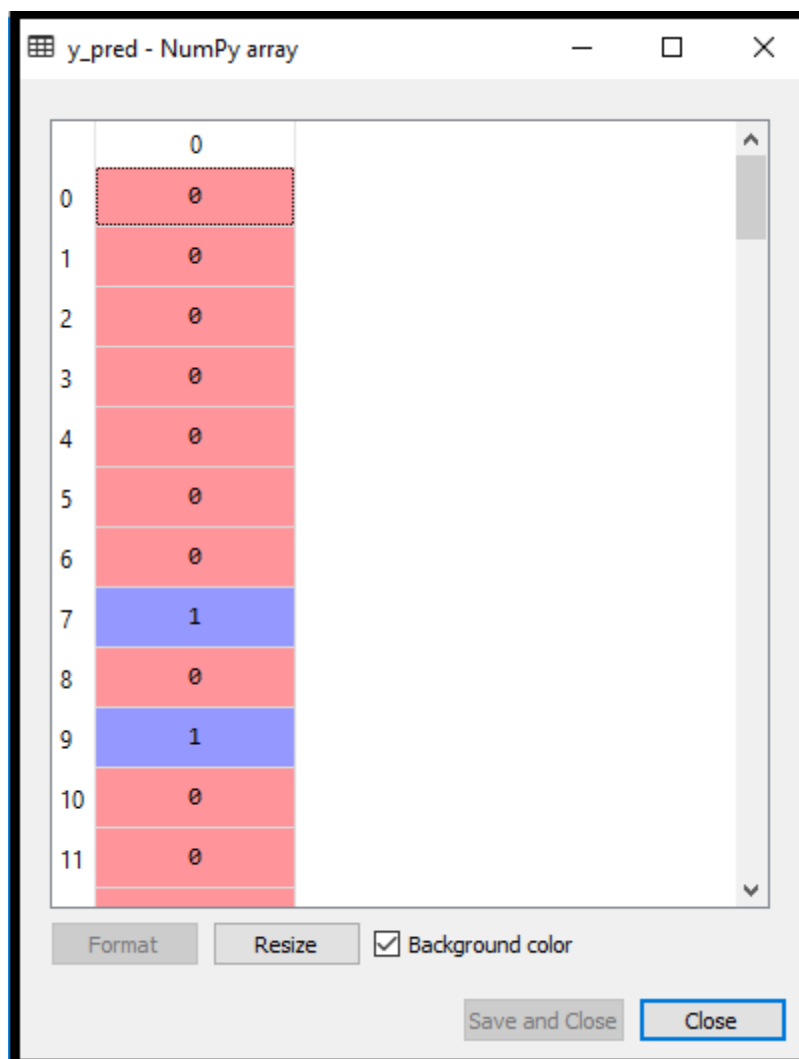
Our model is well trained on the training set, so we will now predict the result by using test set data. Below is the code for it:

```
1. #Predicting the test set result
```

```
2.y_pred= classifier.predict(x_test)
```

In the above code, we have created a `y_pred` vector to predict the test set result.

**Output:** By executing the above code, a new vector (`y_pred`) will be created under the variable explorer option. It can be seen as:



The above output image shows the corresponding predicted users who want to purchase or not purchase the car.

#### 4. Test Accuracy of the result

Now we will create the confusion matrix here to check the accuracy of the classification. To create it, we need to import the **confusion\_matrix** function of the sklearn library. After importing the function, we will call it using a new variable **cm**. The function takes two parameters, mainly **y\_true**( the actual values) and **y\_pred** (the targeted value return by the classifier). Below is the code for it:

```
1. #Creating the Confusion matrix
2. from sklearn.metrics import confusion_matrix
3. cm= confusion_matrix()
```

#### Output:

By executing the above code, a new confusion matrix will be created. Consider the below image:



We can find the accuracy of the predicted result by interpreting the confusion matrix. By above output, we can interpret that  $65+24= 89$  (Correct Output) and  $8+3= 11$  (Incorrect Output).

## 5. Visualizing the training set result

Finally, we will visualize the training set result. To visualize the result, we will use **ListedColormap** class of matplotlib library. Below is the code for it:

```
1. #Visualizing the training set result
2. from matplotlib.colors import ListedColormap
   p
3. x_set, y_set = x_train, y_train
4. x1, x2 = nm.meshgrid(nm.arange(start = x_
   set[:, 0].min() -
   1, stop = x_set[:, 0].max() + 1, step = 0.01
```



```

    ),
5.nm.arange(start = x_set[:, 1].min() -
    1, stop = x_set[:, 1].max() + 1, step = 0.01
    ))
6.mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
    alpha = 0.75, cmap = ListedColormap(('purple', 'green' )))
8.mtp.xlim(x1.min(), x1.max())
9.mtp.ylim(x2.min(), x2.max())
10.    for i, j in enumerate(nm.unique(y_set)):

11.        mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
12.            c = ListedColormap(('purple', 'green'))(i), label = j)
13.    mtp.title('Logistic Regression (Training set)')
14.    mtp.xlabel('Age')
15.    mtp.ylabel('Estimated Salary')
16.    mtp.legend()
17.    mtp.show()

```

In the above code, we have imported the **Listed Colormap** class of Matplotlib library to create the colormap for visualizing the result. We have created two new variables **x\_set** and **y\_set** to replace **x\_train** and **y\_train**. After that, we have used the **nm.meshgrid** command to create a rectangular grid, which has a range of -

1(minimum) to 1 (maximum). The pixel points we have taken are of 0.01 resolution.

To create a filled contour, we have used **mtp.contourf** command, it will create regions of provided colors (purple and green). In this function, we have passed the **classifier.predict** to show the predicted data points predicted by the classifier.

**Output:** By executing the above code, we will get the below output:



The graph can be explained in the below points:

- In the above graph, we can see that there are some **Green points** within the green region and **Purple points** within the purple region.
- All these data points are the observation points from the training set, which shows the result for purchased variables.

- This graph is made by using two independent variables i.e., **Age on the x-axis** and **Estimated salary on the y-axis**.
- The **purple point observations** are for which purchased (dependent variable) is probably 0, i.e., users who did not purchase the SUV car.
- The **green point observations** are for which purchased (dependent variable) is probably 1 means user who purchased the SUV car.
- We can also estimate from the graph that the users who are younger with low salary, did not purchase the car, whereas older users with high estimated salary purchased the car.
- But there are some purple points in the green region (Buying the car) and some green points in the purple region (Not buying the car). So we can say that younger users with a high estimated salary purchased the car, whereas an older user with a low estimated salary did not purchase the car.

### **The goal of the classifier:**

We have successfully visualized the training set result for the logistic regression, and our goal for this classification is to divide the users who purchased the SUV car and who did not purchase the car. So from the output graph, we can clearly see the two regions (Purple and Green) with the

observation points. The Purple region is for those users who didn't buy the car, and Green Region is for those users who purchased the car.

### **Linear Classifier:**

As we can see from the graph, the classifier is a Straight line or linear in nature as we have used the Linear model for Logistic Regression. In further topics, we will learn for non-linear Classifiers.

### **Visualizing the test set result:**

Our model is well trained using the training dataset. Now, we will visualize the result for new observations (Test set). The code for the test set will remain same as above except that here we will use **x\_test** and **y\_test** instead of **x\_train** and **y\_train**. Below is the code for it:

```
1. #Visualizing the test set result
2. from matplotlib.colors import ListedColormap
   p
3. x_set, y_set = x_test, y_test
4. x1, x2 = nm.meshgrid(nm.arange(start = x_
   set[:, 0].min() -
   1, stop = x_set[:, 0].max() + 1, step = 0.01
   ),
5. nm.arange(start = x_set[:, 1].min() -
   1, stop = x_set[:, 1].max() + 1, step = 0.01
   ))
6. mtp.contourf(x1, x2, classifier.predict(nm.arr
```

```

    ay([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
7.alpha = 0.75, cmap = ListedColormap(['purple', 'green' ]))
8.mtp.xlim(x1.min(), x1.max())
9.mtp.ylim(x2.min(), x2.max())
10.    for i, j in enumerate(nm.unique(y_set)):

11.        mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
12.                    c = ListedColormap(['purple', 'green'])(i), label = j)
13.    mtp.title('Logistic Regression (Test set)')

14.    mtp.xlabel('Age')
15.    mtp.ylabel('Estimated Salary')
16.    mtp.legend()
17.    mtp.show()

```

## Output:



The above graph shows the test set result. As we

can see, the graph is divided into two regions (Purple and Green). And Green observations are in the green region, and Purple observations are in the purple region. So we can say it is a good prediction and model. Some of the green and purple data points are in different regions, which can be ignored as we have already calculated this error using the confusion matrix (11 Incorrect output).

Hence our model is pretty good and ready to make new predictions for this classification problem.

---

## Trees and neural network

A Decision Tree is an algorithm used for supervised learning problems such as classification or regression. A decision tree or a classification tree is a tree in which each internal (nonleaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called **recursive partitioning**. The recursion is completed

when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees is an example of a greedy algorithm, and it is the most common strategy for learning decision trees.

Decision trees used in data mining are of two main types –

- **Classification tree** – when the response is a nominal variable, for example if an email is spam or not.
- **Regression tree** – when the predicted outcome can be considered a real number (e.g. the salary of a worker).

Decision trees are a simple method, and as such has some problems. One of this issues is the high variance in the resulting models that decision trees produce. In order to alleviate this problem, ensemble methods of decision trees were developed. There are two groups of ensemble methods currently used extensively –

- **Bagging decision trees** – These trees are used to build multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for a consensus prediction. This algorithm has been called random forest.
- **Boosting decision trees** – Gradient boosting combines weak learners; in this case, decision trees into a single strong learner, in an iterative fashion. It fits a weak tree to the data and iteratively keeps fitting weak learners in order to correct the error of the previous model.

- in which you select the attributes and conditions that will produce the tree. Then, the tree is pruned to remove irrelevant branches that could inhibit accuracy. Pruning involves spotting outliers, data points far outside the norm, that could throw off the calculations by giving too much weight to rare occurrences in the data.
- Maybe temperature is not important when it comes to your golf score or there was a day when you scored really poorly that's throwing off your decision tree. As you're exploring the data for your decision tree, you can prune specific outliers like your one bad day on the course. You can also prune entire decision nodes, like temperature, that may be irrelevant to classifying your data.

A neural network is **made up of densely connected processing nodes, similar to neurons in the brain.** Each node may be connected to different nodes in multiple layers above and below it. These nodes move data through the network in a feed-forward fashion, meaning the data moves in only one direction. Neural network stems from the way simple neurons are linked to form a complex system greater than the sum of its parts.

Each neuron can make simple decisions based on mathematical calculations. Together, many neurons can analyze complex problems and provide accurate answers. A shallow network is composed of an



input, hidden layer and output layer. A deep neural network has more than one hidden layer, which increases the complexity of the problems it can analyze. A neural network learns to complete a task by examining labeled training examples. The samples must be labeled so the network can learn to distinguish between items using visual patterns correlated with the labels.

A neural network has three functions:

Scoring input

Calculating loss

Updating the model, which begins the process over again A neural network is a corrective feedback loop, giving more weight to data that supports correct guesses and less weight to data that leads to mistakes. A feature known as backpropagation trains the network to identify correct responses and ignore incorrect responses.

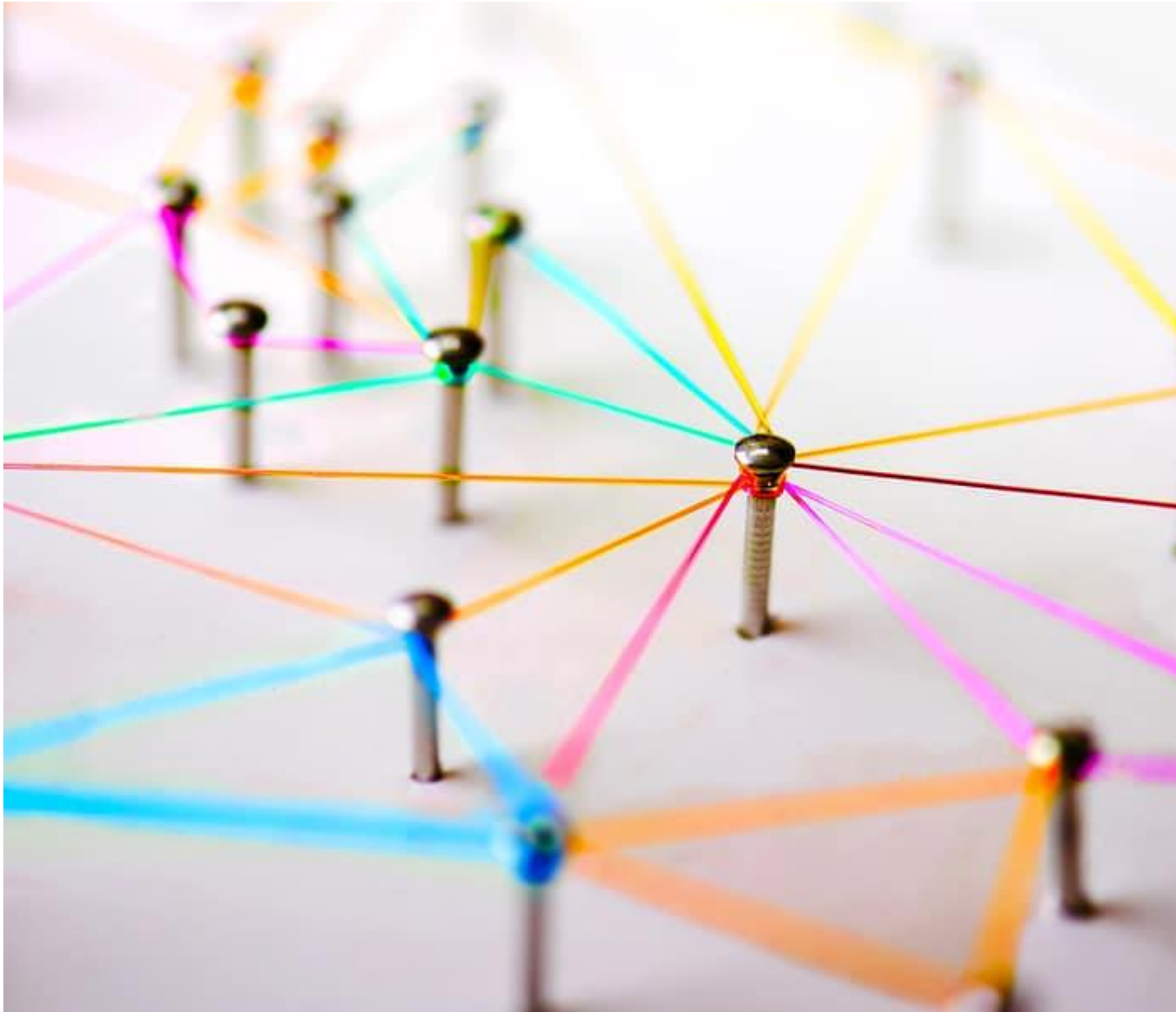
## Big data network analytics

Abstract. Social networking big data is a collection of extremely big data sets with great diversity in social networks. Social networking big data is also a core component for the social influence analysis and the security

## How Social Networking Big Data Provides Opportunities to Capture More Engagement

people and with non-human artifacts have significantly improved data scientists' productivity in organizations, big data analytics can hoard the wisdom of crowds, disclose patterns and harvest best practices. With the aim to make big data work on social media to yield more traffic and interaction to a business

Disclaimer: The information provided in this article is solely the author's opinion and not investment advice – it is provided for educational purposes only. By using this, you agree that the information does not constitute any investment or financial instructions. Do conduct your own research and reach out to financial advisors before making any investment decisions.



In recent times, the volume of information generated by businesses or individuals has increased immensely. In parallel to this, data from sensors, social media, mobile and location are also mounting at an unprecedented rate. Social network site Facebook, for instance, is nearly fully connected, with 2.45 billion monthly active users as of the third quarter of 2019, making it a single, large connected component for data generation. Thus, to garner more engagement and bolster a company

website traffic easily through social networks, there is a need to convey planning and leverage big data. However, to make use of big data successfully, businesses will need to utilize certain technology and analytical tools that can transform the value of the data for businesses' marketing purposes. And for this, they will have to glean high-quality data from their social networks; know about their social lives in general; and make a strategy to move their data. Also, ensuring the characteristic of data such as volume, velocity and variety, organizations can integrate with distinct internet-based social networks and media tools. And once they practice data from social media channels like Facebook, Twitter, Instagram and others, they will get more and better social interactions.

## Social Networking Big Data

Social networking big data is a collection of enormously big data sets with excessive diversity in social networks. It is a core component of social influence analysis and security. Currently, the work on social networking big data focuses on information processing, including data mining and analysis. Developments and advancements in social networks and analytics majorly revolve around internet-based computing paradigms such as cloud and services computing. Most social networks at the present time

connect groups or individuals who divulge similar interests or features. However, in years to come, it is expected that such networks will connect other entities as well, including software components, Web-based services, data resources and workflows.

## Rules of big data

### Big Data Analytics - Association Rules

**A rule is defined as an implication of the form  $X \Rightarrow Y$  where  $X, Y \subseteq I$  and  $X \cap Y = \emptyset$ .** The sets of items (for short item-sets)  $X$  and  $Y$  are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule.

**Association Rule** Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items .It allows retailers to identify relationships between the items that people buy together frequently .Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

### **TID Items**

1 Bread, Milk

### **TID Items**

- 2 Bread, Diaper, Beer, Eggs
- 3 Milk, Diaper, Beer, Coke
- 4 Bread, Milk, Diaper, Beer
- 5 Bread, Milk, Diaper, Coke

[GEEKSFORGEEKS](https://www.geeksforgeeks.org/association-rule-mining-interesting-associations-and-relationships-among-large-sets-of-data-items/)

### **Association Rule**

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis .Market Based Analysis is one of the key techniques used by large relations to show associations between items .It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

### **TID Items**

- 1 Bread, Milk
- 2 Bread, Diaper, Beer, Eggs
- 3 Milk, Diaper, Beer, Coke

## TID Items

4 Bread, Milk, Diaper, Beer

5 Bread, Milk, Diaper, Coke

Before we start defining the rule, let us first see the basic definitions. **Support Count**(  $\sigma$ ) – Frequency of occurrence of a itemset.

Here  $\sigma(\{\text{Milk, Bread, Diaper}\})=2$

**Frequent Itemset** – An itemset whose support is greater than or equal to mins up threshold.

**Association Rule** – An implication expression of the form  $X \rightarrow Y$ , where X and Y are any 2 item sets.

Example:  $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

## Rule Evaluation Metrics –

### **.Support(s)**

–  
The number of transactions that include items in the  $\{X\}$  and  $\{Y\}$  parts of the rule as a percentage of the total number of transaction .It is a measure of how frequently the collection of items occur together as a percentage of all transactions.

$$\text{.Support} = \frac{\sigma(X+Y)}{\text{total}}$$

–  
It is interpreted as fraction of transactions that contain both X and Y.

### **.Confidence(c)**

–  
It is the ratio of the no of transactions that includes all items in  $\{B\}$  as well as the no of transactions that includes all items in  $\{A\}$  to the no of transactions that includes all items in  $\{A\}$ .

• **Conf(X=>Y)** =  $\frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}$  –

It measures how often each item in Y appears in transactions that contains items in X also.

• **Lift(I)** –

The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the item sets X and Y are independent of each other .The expected confidence is the confidence divided by the frequency of {Y}.

• **Lift(X=>Y)** =  $\frac{\text{Conf}(X=>Y)}{\text{Supp}(Y)}$  –

Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected .Greater lift values indicate stronger association.

## Association Rules Applications

Understanding the customer purchasing behaviour by using association rule mining enables different applications.

## SHORT QUESTIONS

- 1.What is decision trees and neural network?
- 2.What is big data network analytics?
- 3.Explain the Rules of big data?
- 4.Explain about predictive analytics?
- 5.Explain about target definition?



## **LONG QUESTION**

1. Write about predictive analytics and target definition?
2. Explain logistics Regression?
3. What is decision tree and neural network?
4. What is big data network analytics in details?

## **UNIT -5**

### **RDBMS**

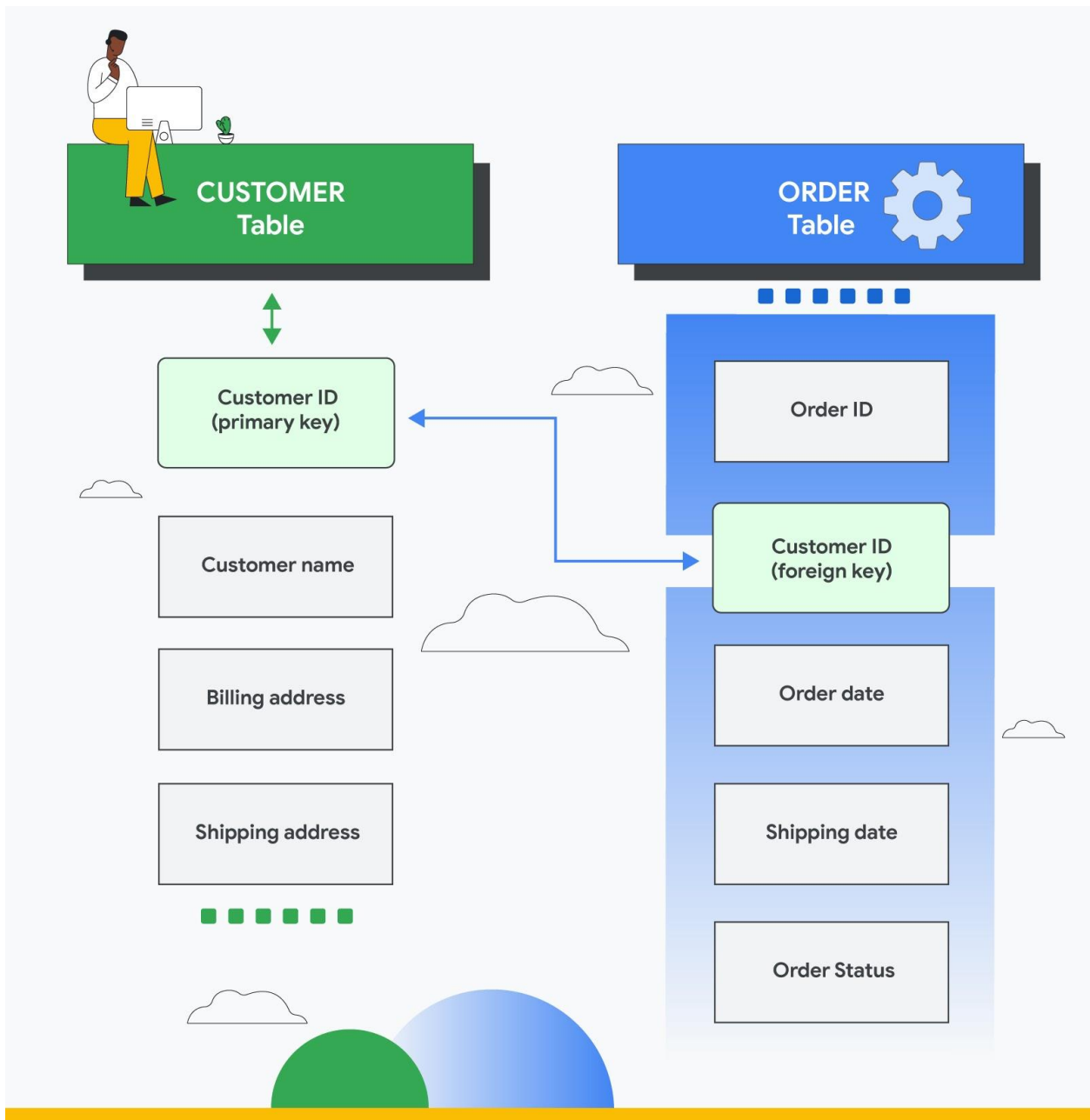
RDMS (Relational Database Management System): RDBMS is an information management system, which is based on a data model .In RDBMS tables are used for information storage. Each row of the table represents a record and column represents an attribute of data. Organization of data and their manipulation processes are different in RDBMS from other databases. RDBMS ensures ACID (atomicity, consistency, integrity, durability) properties required for designing a database. The purpose of RDBMS is to store, manage, and retrieve data as quickly and reliably as possible. Hadoop: It is an open-source software framework used for storing data and running applications on a group of commodity hardware. It has large storage capacity and high

processing power. It can manage multiple concurrent processes at the same time. It is used in predictive analysis, data mining and machine learning. It can handle both structured and unstructured form of data. It is more flexible in storing, processing, and managing data than traditional RDBMS. Unlike traditional systems, Hadoop enables multiple analytical processes on the same data at the same time. It supports scalability very flexibly. Below is a table of differences between RDBMS and Hadoop:

An RDBMS is a **type of database management system (DBMS) that stores data in a row-based table structure which connects related data elements**. An RDBMS includes functions that maintain the security, accuracy, integrity and consistency of the data. This is different than the file storage used in a DBMS.

### RDBMSs Importance in Big Data Environment

Most key operational information is likely kept in the RDBMS for both small and large businesses. Many businesses use multiple RDBMS for different aspects of their operations. Customer information may be saved in one database while transactional data is kept in another.



A relational database is a collection of information that organizes data in predefined relationships where data is stored in one or more tables (or "relations") of columns and rows, making it easy to see and understand how different data structures relate to each other. Relationships are a logical connection between different

tables, established on the basis of interaction among these tables.

Learn how Google Cloud's relational databases [Cloud SQL](#), [Cloud Spanner](#) and [AlloyDB for PostgreSQL](#) can help you reduce operational costs and help you build transformative applications.

Ready to get started? Create a 90-day [Cloud Spanner free trial instance](#) with 10 GB of storage at no cost.

## **Relational database defined**

A relational database (RDB) is a way of structuring information in tables, rows, and columns. An RDB has the ability to establish links—or relationships—between information by joining tables, which makes it easy to understand and gain insights about the relationship between various data points.

## **The relational database model**

Developed by E.F. Codd from IBM in the 1970s, the relational database model allows any table to be related to another table using a common attribute. Instead of using hierarchical structures to organize data, Codd proposed a shift to using a data model where data is stored, accessed, and related in tables without reorganizing the tables that contain them.

Think of the relational database as a collection of spreadsheet files that help businesses organize, manage, and relate data. In the relational database model, each “spreadsheet” is a table that stores information, represented as columns (attributes) and rows .

**Most key operational information is likely kept in the RDBMS for both small and large businesses.** Many businesses use multiple RDBMS for different aspects of their operations. Customer information may be saved in one database while transactional data is kept in another.

## **Components of Hadoop and file system**

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly. Hadoop consists of four main modules:

Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large data sets. Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.

MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input

data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result. Hadoop Common – Provides common Java libraries that can be used across all modules. Hadoop makes it easier to use all the storage and processing capacity in cluster servers, and to execute distributed processes against huge amounts of data. Hadoop provides the building blocks on which other services and applications can be built.

Applications that collect data in various formats can place data into the Hadoop cluster by using an API operation to connect to the Name Node. The Name Node tracks the file directory structure and placement of “chunks” for each file, replicated across Data Nodes. To run a job to query the data, provide a [MapReduce](#) job made up of many map and reduce tasks that run against the data in HDFS spread across the Data Nodes. Map tasks run on each node against the input files supplied, and reducers run to aggregate and organize the final output.

The Hadoop ecosystem has grown significantly over the years due to its extensibility. Today, the Hadoop ecosystem includes many tools and applications to help collect, store, process, analyze, and manage big data. Some of the most popular applications are:

- [Spark](#) – An open source, distributed processing system commonly used for big data workloads. Apache Spark uses in-memory caching and optimized execution for fast performance, and it supports general batch processing, streaming

analytics, machine learning, graph databases, and ad hoc queries.

- [Presto](#) – An open source, distributed SQL query engine optimized for low-latency, ad-hoc analysis of data. It supports the ANSI SQL standard, including complex queries, aggregations, joins, and window functions. Presto can process data from multiple data sources including the Hadoop Distributed File System (HDFS) and Amazon S3.
- [Hive](#) – Allows users to leverage Hadoop MapReduce using a SQL interface, enabling analytics at a massive scale, in addition to distributed and fault-tolerant data warehousing.
- [HBase](#) – An open source, non-relational, versioned database that runs on top of Amazon S3 (using EMRFS) or the Hadoop Distributed File System (HDFS). HBase is a massively scalable, distributed big data store built for random, strictly consistent, real-time access for tables with billions of rows and millions of columns.
- Zeppelin – An interactive notebook that enables interactive data exploration.

## Running Hadoop on AWS

Amazon EMR is a managed service that lets you process and analyze large datasets using the latest versions of [big data](#) processing frameworks such as

Apache Hadoop, Spark, HBase, and Presto on fully customizable clusters.

- Easy to use: You can launch an Amazon EMR cluster in minutes. You don't need to worry about node provisioning, cluster setup, Hadoop configuration, or cluster tuning.
- Low cost: Amazon EMR pricing is simple and predictable: You pay an hourly rate for every instance hour you use and you can leverage Spot Instances for greater savings.
- Elastic: With Amazon EMR, you can provision one, hundreds, or thousands of compute instances to process data at any scale.
- Transient: You can use EMRFS to run clusters on-demand based on HDFS data stored persistently in Amazon S3. As jobs finish, you can shut down a cluster and have the data saved in [Amazon S3](#). You pay only for the compute time that the cluster is running.
- Secure: Amazon EMR uses all common security characteristics of AWS services:
  - Identity and Access Management (IAM) roles and policies to manage permissions.
  - Encryption in-transit and at-rest to help you protect your data and meet compliance standards, such as HIPAA.
  - Security groups to control inbound and outbound network traffic to your cluster nodes.



## Hadoop technologies stack

Hadoop stack holders and components of technologies There are four major elements of Hadoop i.e., HDFS, MapReduce, YARN, and Hadoop Common. Most of the tools or solutions are used to supplement or support these major element There are four major elements of Hadoop i.e. HDFS, MapReduce, YARN, and Hadoop Common. Most of the tools or solutions are used to supplement or support these major elements.

The advantage of Hadoop is that **it offers both a distributed storage engine as well as a possibility to use a Hadoop cluster for a distributed analytical engine for big data analytics.** Dubbed the three Vs; **volume, velocity, and variety**, these are key to understanding how we can measure big data and just how very different 'big data' is to old fashioned data. Find out more about the 3vs of Big Data at Big Data LDN, the UK's leading data conference & exhibition for your entire data team

### Hadoop Big Data Analytics – How Big is Big

Hadoop big data analytics has the power to change the world. Hadoop, the de facto platform for the distributed big data, also plays an important role in big data analytics. Organizations now realize the inherent value of transforming these big data into actionable insights. Data science is the highest form of big data analytics that produce the most accurate actionable insights, identifying what will happen next and what to do about it.

The RapidMiner platform is an excellent solution for handling unstructured data like text files, web traffic logs, and even images. But we will discuss how the volume of big data can be easily handled – without writing a single line of code (unless you want to, of course).

**Analytical Engines in RapidMiner** Rapid Miner offers flexible approaches to remove any limitations in data set size. The most often used is the in-memory engine, where data is loaded completely into memory and is analyzed there. This and other engines are outlined below.

**In-Memory:** The natural storage mechanism of RapidMiner is in-memory data storage, highly optimized for data access usually performed for analytical tasks .In-memory analytics is always the fastest way to build analytical models Data set size is restricted by hardware (memory): The more memory is available the larger the data sets which can be analyzed Data set size: On decent hardware, up to ca. 100 million data points

**In-Hadoop:** The advantage of Hadoop is that it offers both a distributed storage engine as well as a possibility to use a Hadoop cluster for a distributed analytical engine for big data analytics.

Not applicable for quick, interactive analysis Runtime depends on the power of the Hadoop cluster, but it has virtually infinite scalability Due to overhead introduced by Hadoop, its usage is not recommended for smaller data set sizes Data set size: Unlimited (limit is the external storage capacity)Loop-based workflow design: All the storage types

above can be combined with loop-based workflows where data processing / modeling is performed on partitions of the data and the results are combined afterwards. It depends on the data set as well as the analysis if, for example, a loop-based approach using the in-memory approach is faster.

Not applicable for all analysis tasks Memory is no longer the limitation but runtime becomes more important Data set size: Unlimited (limit is the external storage capacity) Runtime Comparison of Analytical Engines within RapidMiner Below, you can find a runtime comparison for the creation of a Naïve Bayes model with:

In-Memory engine

In-Hadoop engine

Hadoop Big Data-Analytics-graph

It can easily be seen that the default in-Memory engine of RapidMiner is the fastest approach in general but will of course fail as soon as the data set size hits the memory limit of the machine. However, training a model on millions of data points is a matter of seconds or minutes only with the In-Memory approach. On decent hardware, RapidMiner recommends organizations use this fast engine as the default for data set sizes up to 100 million data points.

With the in-Hadoop engine, the size of the Hadoop cluster was only three nodes in the experiments above, and is prohibitively slow on small data sets but scales up nicely to very large data sets. This can be further improved by adding more computation nodes to the Hadoop

cluster. Since the overhead for most data sets of common sizes is so large, it is only recommended to use Hadoop as the underlying engine for data sets sizes of 500 million data points and more and where runtime is an important issue at the same time.

## **Conclusion**

Hadoop is not just an effective distributed storage system for large amounts of data, but also, importantly, a distributed computing environment that can execute analyses where the data is. RapidMiner makes use of all the possibilities offered by Hadoop by allowing users to do a distributed advanced analysis on data on Hadoop.

Looking at the runtimes for analytical algorithms, it can be easily seen that limitations in terms of data set sizes have vanished today – but at the price of larger runtimes. This is in all cases prohibitive for interactive reports, but likely also for predictive analytics if the model creation has to be done fast or in real-time. In those cases, an in-memory engine is still the fastest option. The in-Hadoop engine is slow for smaller data sets but is the fastest and sometimes the only option when data sets are really big in terms of volume.

Depending on the application at hand, a certain engine will always be superior to the others and therefore RapidMiner supports both in-memory and in-Hadoop engines in order to give users the flexibility to solve all their analytical problems. The company's goal is that RapidMiner users can

always select the best engine for their specific application and get the optimal results in minimal times.

## Managing Resources and applications

Data management is the practice of collecting, organizing, protecting, and storing an organization's data so it can be analyzed for business decisions. As organizations create and consume data at unprecedented rates, data management solutions become essential for making sense of the vast quantities of data.

Data management is the practice of collecting, organizing, protecting, and storing an organization's data so it can be analyzed for business decisions. As organizations create and consume data at unprecedented rates, data management solutions become essential for making sense of the vast quantities of data. Today's leading data management software ensures that reliable, up-to-date data is always used to drive decisions. The software helps with everything from data preparation to cataloging, search, and governance, allowing people to quickly find the information they need for analysis.

## Types of Data Management

Data management plays several roles in an organization's data environment, making essential functions easier and less time-intensive. These data management techniques include the following:

Data preparation is used to clean and transform raw data into the right shape and format for analysis, including making corrections and combining data sets.

Data pipelines enable the automated transfer of data from one system to another.

ETLs (Extract, Transform, Load) are built to take the data from one system, transform it, and load it into the organization's data warehouse.

Data catalogs help manage metadata to create a complete picture of the data, providing a summary of its changes, locations, and quality while also making the data easy to find.

Data warehouses are places to consolidate various data sources, contend with the many data types businesses store, and provide a clear route for data analysis .Data governance defines standards, processes, and policies to maintain data security and integrity .Data architecture provides a formal approach for creating and managing data flow .Data security protects data from unauthorized access and corruption .Data modeling documents the flow of data through an application or organization. **types of data management systems**

- Customer Relationship Management System or CRM. ...
- Marketing technology systems. ...
- Data Warehouse systems. ...
- Analytics tools.

Real-world big data examples:

- Learning about consumer shopping habits
- Customized marketing
- Discovering new customer leads
- Fuel optimization devices for the industry of transportation
- Prediction of user demand for ridesharing businesses
- Observing health conditions via data from wearables
- Live road mapping for independent vehicles
- Streamlined media streaming
- Predictive inventory ordering
- Customized health plans for cancer patients
- Real-time monitoring of data and cybersecurity protocols

## **Data ware application and Data ware housing Hadoop concepts**

**Data warehouse is the collection of historical data from different operations in an enterprise.** 2. Big data is a technology to store and manage large amount of data. Data warehouse is an architecture used to organize the data. A data warehouse is **a central repository of information that can be analyzed to make more informed decisions.** Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence.

The difference between Hadoop and data warehouse is like a hammer and a nail- Hadoop is a big data

technology for storing and managing big data, whereas data warehouse is an architecture for organizing data to ensure integrity.

## **Table of Contents**

- Structured data.
- Unstructured data.
- Semi-structured data

A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence

## **Data Warehouse Architecture**

The single tier Data Warehouse architecture is composed of a single hardware layer. This hardware layer is composed of a single hardware layer. There are three approaches to creating a data warehouse layer: Single tier, two-tier, and three-tier.

**Single-tier architecture:** A single-layer structure aimed at keeping data space minimal. This structure is rarely used in real life.

**Two-tier architecture:** Data warehouse is the aggregation of data in a format that is easy to transform and load into a database. Data warehouses can be implemented in a number of different ways, and it is important to pick the right one for your business needs. The most important thing to consider is scalability. If you want to store large amounts of data in a



small amount of space, then you should consider using a data warehouse.

**Three-Tier Data Warehouse Architecture:** The Top, Middle, and Bottom Tiers of this Architecture of Data Warehouse are collectively referred to as the Top Tier.

1. The bottom tier of the Datawarehouse is a relational database system. This database system typically contains a relational database system. Back-end tools clean, transform, and load data into this layer.
2. A middle tier OLAP server is either ROLAP or MOLAP-based. It abstracts OLAP from the end user by serving as a middle tier OLAP server. Data warehouses that facilitate end-user interaction with the database and middle tier OLAP servers that abstract OLAP from the end user are known as middle tier OLAP servers.
3. The front-end client layer of the top-tier is important because it is the first point of interaction with the data. It is where data is presented to the end user, and decisions are made with the data. The front-end client layer of top-tier must work with real-time data and must be able to process data quickly. It is also important to work with data that is in a format that top-tier can understand and use. Typically, top-tier data is in a relational database format, but it could be a file or a stream. Top-tier data must be well-structured, must be validated, and must be structured in a way that allows for easier data profiling and analytics.

A data warehouse system must meet the following architecture features:

## **Data Warehouse Architecture Properties**

We sometimes wish to keep analytical and transactional processing as far away as possible.

The scalability of the solution should be demonstrated by the ability to process a huge volume of data and stream it to different destinations, at high speed, in various formats. The data stream should be processed and presented in the required format, at the right time and location, with the minimum impact to the existing infrastructure. The data stream must be protected and managed with the highest level of confidentiality and integrity. The size of the data stream and the rate at which the data is being generated must be determined by the business requirements, and the available hardware and software resources must be utilized to the fullest extent possible.

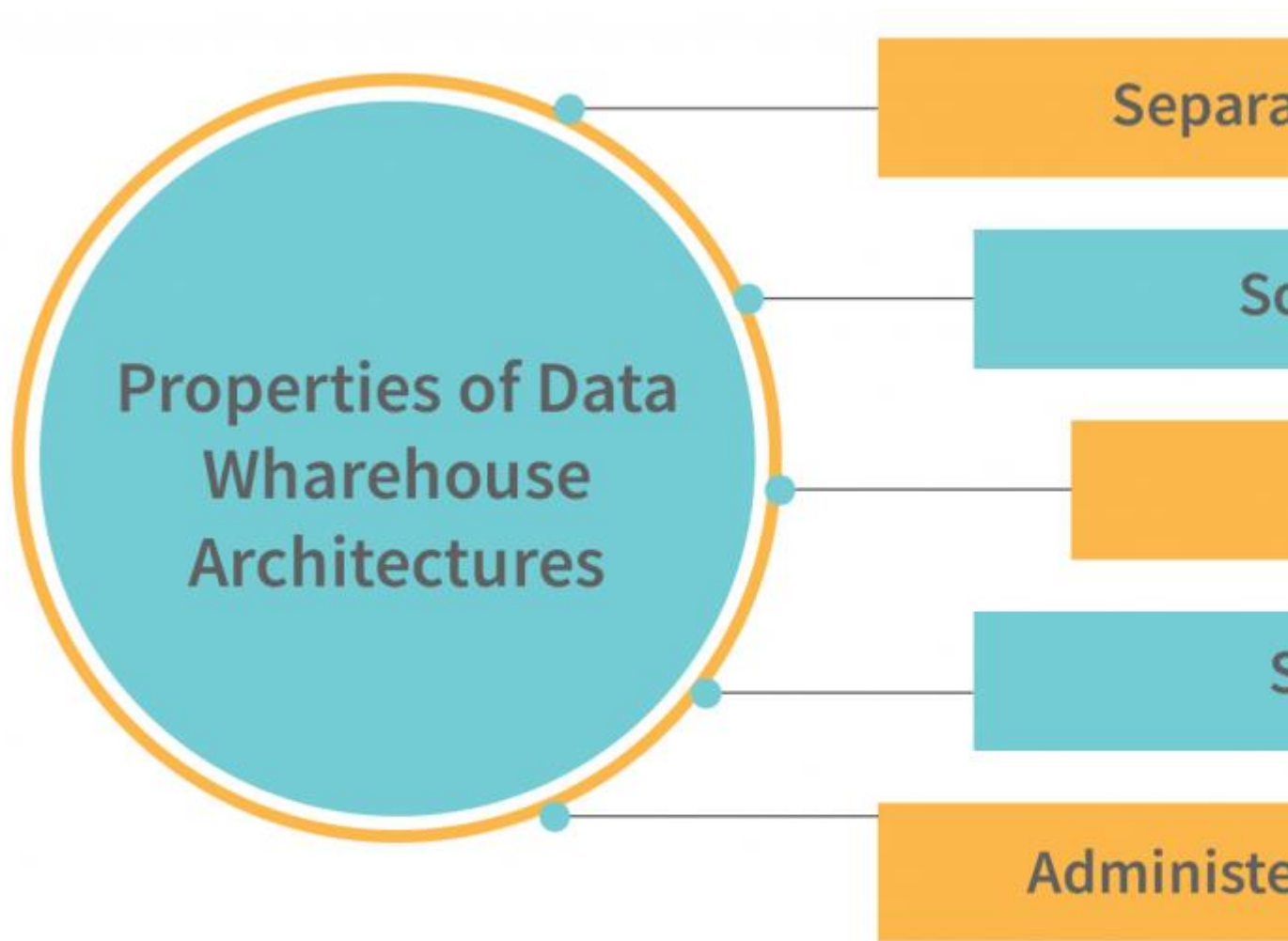
The architecture should be extensible; new functionality can be implemented in an existing service by extending the service's APIs. For example, an insurance company could extend their customer service platform to provide a new feature that allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, can be implemented in an existing service by extending the service's APIs. For example, an insurance company could extend their customer service platform to provide a new feature that

allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, should be implemented in the core services; the core services can be extended for new business functions, such as customer relationship management.

Data security is a critical aspect of the data governance strategy. Data security controls at the source include establishing data access controls and data encryption. Data security controls at the perimeter include data security policies and monitoring access to the data.

It should be simple and straightforward, and users should be able to work with the data in an efficient and effective manner. Data Warehouse management should be easy to understand and implement. Data Warehouse management should not be complicated and difficult for beginners should not find their way into data warehouse management. It should be simple to use and easy to understand<sup>2</sup> Data Warehouse Architecture Properties

A data warehouse system must meet the following architecture features:



- We sometimes wish to keep analytical and transactional processing as far away as possible.
- The scalability of the solution should be demonstrated by the ability to process a huge volume of data and stream it to different destinations, at high speed, in various formats. The data stream should be processed and presented in the required format, at the right time and location, with the minimum impact to the existing infrastructure. The data stream must be protected and managed with the highest level of confidentiality and

integrity. The size of the data stream and the rate at which the data is being generated must be determined by the business requirements, and the available hardware and software resources must be utilized to the fullest extent possible.

- The architecture should be extensible; new functionality can be implemented in an existing service by extending the service's APIs. For example, an insurance company could extend their customer service platform to provide a new feature that allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, can be implemented in an existing service by extending the service's APIs. For example, an insurance company could extend their customer service platform to provide a new feature that allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, should be implemented in the core services; the core services can be extended for new business functions, such as customer relationship management.
- Data security is a critical aspect of the data governance strategy. Data security controls at the source include establishing data access controls and data encryption. Data security controls at the perimeter include data security policies and monitoring access to the data.
- It should be simple and straightforward, and users should be able to work with the data in an efficient and effective manner. Data Warehouse management should

be easy to understand and implement. Data Warehouse management should not be complicated and difficult for beginners should not find their way into data warehouse management. It should be simple to use and easy to understand.

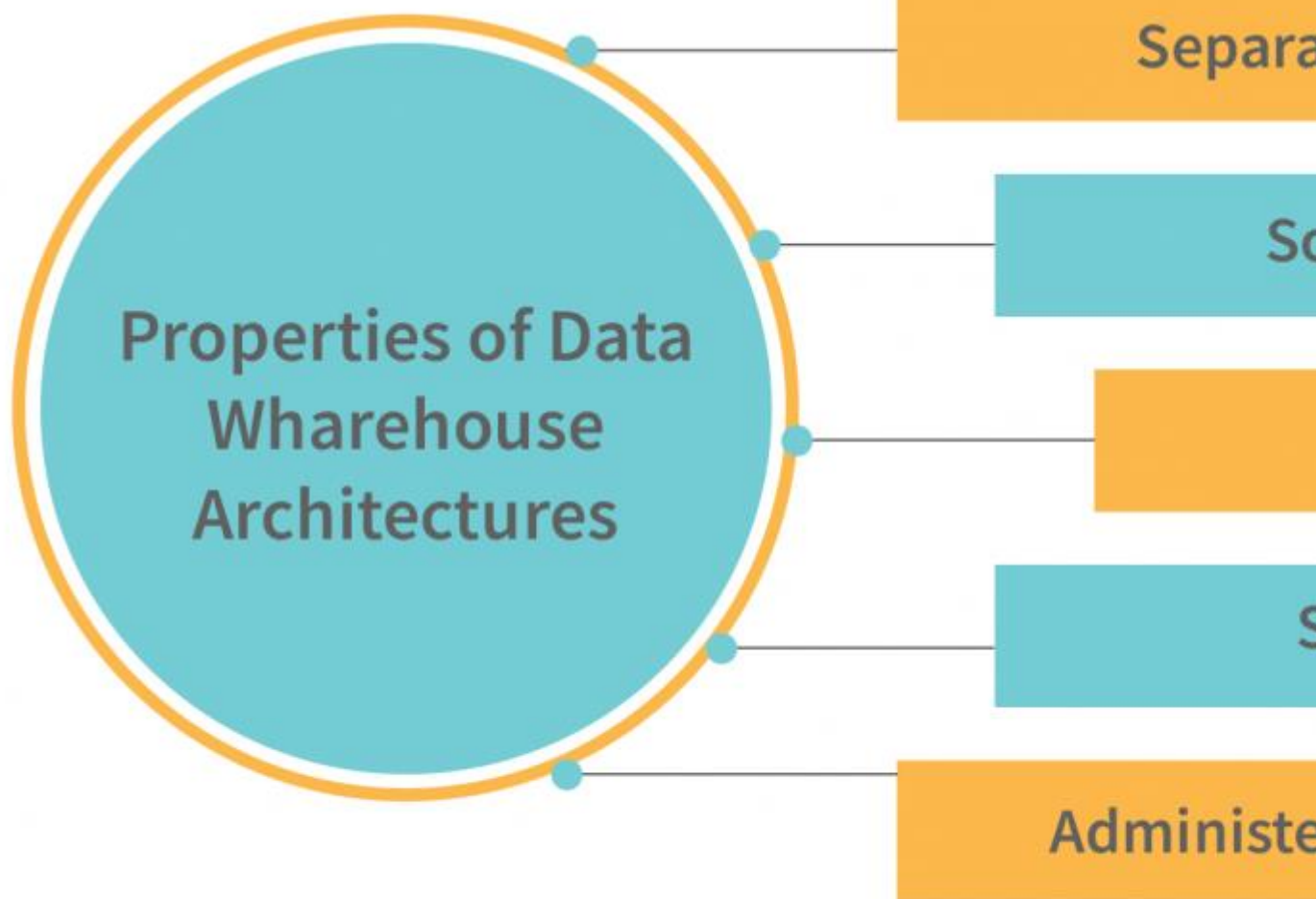
- We sometimes wish to keep analytical and transactional processing as far away as possible.
- The scalability of the solution should be demonstrated by the ability to process a huge volume of data and stream it to different destinations, at high speed, in various formats. The data stream should be processed and presented in the required format, at the right time and location, with the minimum impact to the existing infrastructure. The data stream must be protected and managed with the highest level of confidentiality and integrity. The size of the data stream and the rate at which the data is being generated must be determined by the business requirements, and the available hardware and software resources must be utilized to the fullest extent possible.
- The architecture should be extensible; new functionality can be implemented in an existing service by extending the service's APIs. For example, an insurance company could extend their customer service platform to provide a new feature that allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, can be implemented in an existing service by extending the service's APIs. For example, an insurance company could

extend their customer service platform to provide a new feature that allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, should be implemented in the core services; the core services can be extended for new business functions, such as customer relationship management.

- Data security is a critical aspect of the data governance strategy. Data security controls at the source include establishing data access controls and data encryption. Data security controls at the perimeter include data security policies and monitoring access to the data.

- 

To Data Warehouse Architecture Properties



g architecture features:

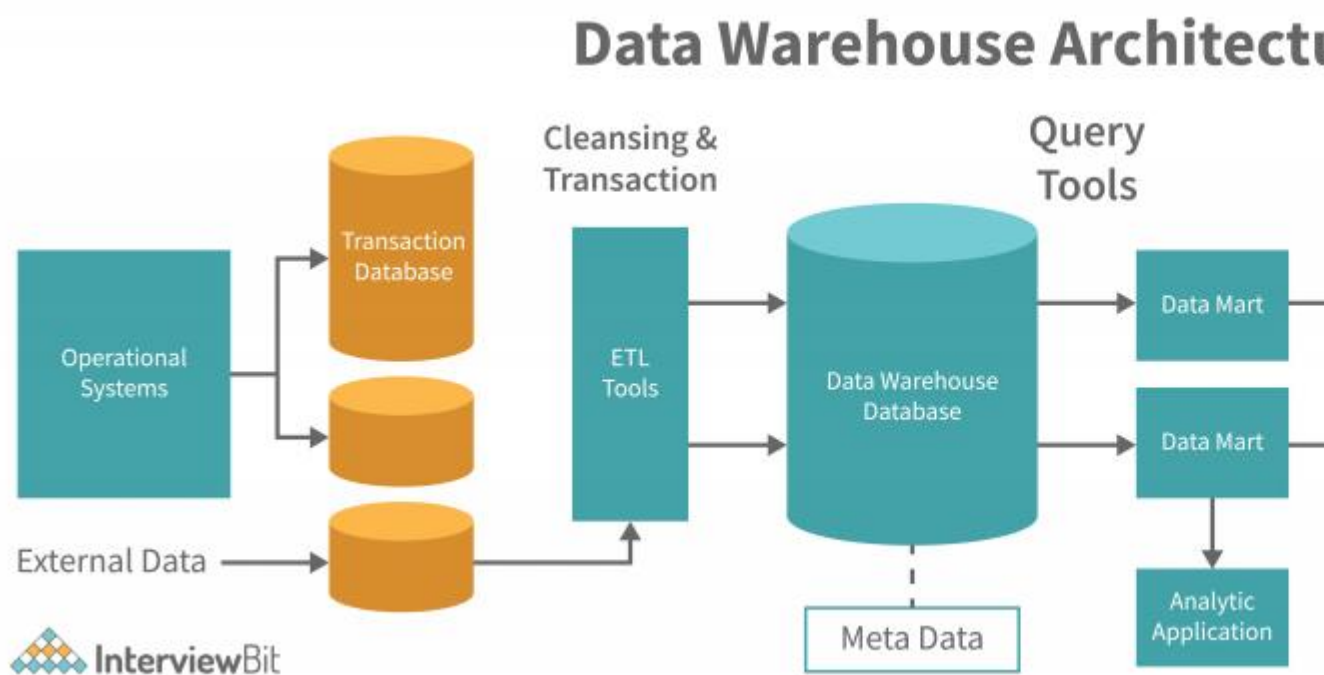
- We sometimes wish to keep analytical and transactional processing as far away as possible
- The scalability of the solution should be demonstrated by the ability to process a huge volume of data and stream it to different destinations, at high speed, in various formats. The data stream should be processed and presented in the required format, at the right time and location, with the minimum impact to the existing



infrastructure. The data stream must be protected and managed with the highest level of confidentiality and integrity. The size of the data stream and the rate at which the data is being generated must be determined by the business requirements, and the available hardware and software resources must be utilized to the fullest extent possible.

- The architecture should be extensible; new functionality can be implemented in an existing service by extending the service's APIs. For example, an insurance company could extend their customer service platform to provide a new feature that allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, can be implemented in an existing service by extending the service's APIs. For example, an insurance company could extend their customer service platform to provide a new feature that allows customers to obtain a personalized quote based on their preferences. Newer technologies, such as artificial intelligence, should be implemented in the core services; the core services can be extended for new business functions, such as customer relationship management.
- Data security is a critical aspect of the data governance strategy. Data security controls at the source include establishing data access controls and data encryption. Data security controls at the perimeter include data security policies and monitoring access to the data.

- It should be simple and straightforward, and users should be able to work with the data in an efficient and effective manner. Data Warehouse management should be easy to understand and implement. Data Warehouse management should not be complicated and difficult for beginners should not find their way into data warehouse management. It should be simple to use and easy to understand.



These are the operating tools

## SHORT QUESTIONS

1. Define RDBMS, and explain it?
2. Components of hadoop and file system?
3. Explain Managing Resources and applications?
4. What is hadoop technologies?

## **LONG QUESTIONS**

- 1.What is Hadoop technologies stack?
2. Data ware applications and data ware housing Hadoop concepts?
- 3.Explain components of hadoop in details?
- 4.What is hadoop technologies and hadoop technologies stack?
- 5.Write about RDBMS in details?

### **Internal Examination-1**

#### **Section-A**

**Answer any FIVE of the following Questions .Each question carries 2 marks. (5×2=10)**

- 1 (a) What is big data?
- (b) Explain Industry of Big data and examples?
- (C) Describe unstructured data analytics?
- (E) Write about cloud and big Data ?
- (F) Big data Emerging Technologies?
- (G) Opensource technologies for big data analytics?
- (H) Data discovery and limitations explain?
- (I) Terminologies and approaches of data?

(J) What is BASE ?

### **Section-B**

**Answer any ONE full question from each unit.**

**Question carries 10 marks. (2×10=20)**

#### **Unit -1**

(2) How many branches are there in big data? Explained in detail?

(OR)

(3) What are the differences between the bigdata and business data analytics?

#### **UNIT-2**

(4) Describe the objectives of Big data technologies and approaches?

(OR)

(5) What are the elements of opensource data? Explain their strategies?

### **Internal Examination-2**

## **Section-A**

**Answer any FIVE of the following Questions.**

**Each Question carries 2 marks (5\*2=10)**

1 (a) State the importance of data visualization and organization?

(b) What is privacy data?

(c) Define the concept of scale and convergences?

(D) Define data and cloud analyze?

(e) explain neural network?

(f) classification of trees?

(g) sequence of network analysis?

(h) Hadoop files system?

(i) Define stack technologies?

(j) Managing resources and applications with YARN?

## **PART-B**

**Answer any ONE full question from each Unit.**

**Each Question carries 12marks. (5\*12=60).**

## **UNIT-1**

(1) Explain the data security and privacy in big data?

(OR)

(2) Hadoop concepts and stack?

## **UNIT -2**

(3) Big data and RDBMS versus Hadoop data forms?

(OR)

(4) Logistics Regression and decision trees?

## **MASTER OF COMPUTER APPLICATIONS DEGREE EXAMINATION**

**Paper MCA 304B:**

### **BIG DATA ANALYTICS**

**(Under C. B. S. E revised regulation w.e.f.)**

**(Common paper to university and all affiliated colleges)**

**Time: 3 hours Max marks:70**

**Answer any FIVE of the following Questions.**

**Each Question carries 2 marks (5\*2=10)**

### **Unit -1**

(1) what is big data and security and privacy?

(OR)

(2) Limitations of data big analytics? Explain BASE and industry data?

## **UNIT-2**

(3) Describe the new and old approaches of data?

(OR)

(4) Define cloud and big data analytics?

## **UNIT -3**

(5) Big data 90/10 rules of critical thinking?

(OR)

(6) Decision and analytics of knowledge?

## **UNIT -4**

(7) Explain the Logistics Regression and decision trees?

(OR)

(8) Define Rules of big Data? and the Association Rule and methods?

## **UNIT -5**

(9) Explain RDBMS? RDBMS and versus Hadoop?

(OR)

(10) Hadoop and Data ware housing concepts, application of Hadoop using PIG,YARN,HIVE?

