Big Data Analytics

2.1 Syllabus topic - Big Data Analytics:

- Big Data analytics is the process of collecting, organizing and analysing large sets of data (*called* Big Data) to discover patterns and other useful information.
- Big Data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions.
- Analysts working with Big Data typically want the *knowledge* that comes from analysing the data.
- Data or information is in raw format. With increasing data size, it has become a need for inspecting, cleaning, transforming, and modelling data with the goal of finding useful information, making conclusions, and supporting decision making. This process is known as **data analysis**.
- Data mining is a particular data analysis technique where modelling and knowledge discovery for predictive rather than purely descriptive purposes is focused.
- Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information.
- In statistical applications, some people divide business analytics into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA focuses on confirming or falsifying existing hypotheses.
- Predictive analytics does forecasting or classification by focusing on statistical or structural models while in text analytics, statistical, linguistic and structural techniques are applied to extract and classify information from textual sources, a species of unstructured data.
- All are varieties of data analysis.
- So, the Data wave has changed the ways in which industries function. With Big Data has emerged the requirement to implement advanced analytics to it. Now experts can make more accurate and profitable decisions.

High performance Analytics required:

- To analyze such a large volume of data, Big Data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, and forecasting and data optimization.

- Collectively these processes are separate but highly integrated functions of high-performance analytics.
- Using Big Data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions in the future.

The challenges:

- For most organizations, Big Data analysis is a challenge.
- Consider the sheer volume of data and the different formats of the data (both structured and unstructured data) that is collected across the entire organization and the many different ways different types of data can be combined, contrasted and analyzed to find patterns and other useful business information.
- The first challenge is in breaking down data silos to access all data an organization stores in different places and often in different systems.
- A second challenge is in creating platforms that can pull in unstructured data as easily as structured data. This massive volume of data is typically so large that it's difficult to process using traditional database and software methods.
- **Benefits of Big data Analytics:**
- Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions.
- With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management.

2.2. Syllabus topic - Drivers of Big Data Analytics:

Drivers of Big Data Analytics are divided into two broad categories, they are

1. **Business**
2. **Technology**

1. Business:

Business entails market, sales and financial side of things, there are different drivers are involve in Business. Following are the drivers involve in the Big Data,

1. Data Driven Initiatives:
    1. Data driven Innovation
    2. Data driven decision making

3. Data driven discovery

2. Data Science is competitive advantage
3. Sustained Processes
4. Cost advantages of commodity hardware and Open Source Software
5. Quick turnaround and less bench time
6. Automation to backfill redundant/mundane task
7. Optimize workforce to leverage high talent cost

1. Data Driven initiatives:

There are primarily categorized into three types:

i. Data Driven Innovations:

- Ability to drive innovation through those uber targeted data indicators.

ii. Data driven decision making:

- Data driven decision-making is the inherent ability of analytics to sieve through globs of data and identify the best path forward.
- Whether in terms of finding the best route to validating the current route and estimating the success/failure in current strategy.
- It takes decision making away from gut and focus on data backed reasoning for higher chances of success.

iii. Data driven discovery:

- Your data know a whole lot about you than you image. Having a discovery mechanism will help you understand hidden insights that were not visible through traditional means.

2. Data Science as a competitive advantage:

- Big data as a capability to add to their competitive advantage. With a proper data driven framework, businesses could build sustainable capabilities and further leverage these capabilities as a competitive edge.
- If businesses were able to master big data driven capabilities, businesses could use these capabilities to establish secondary source of revenues by selling it to other businesses.

3. Sustained Processed:

- Data driven approach creates sustainable processes, which gives a huge endorsement to big data analytics strategy as a go for enterprise adoption.
- Randomness kills businesses and adds scary risks, while data driven strategy reduces the risk by bringing statistical models, which are measurable.

4. Cost advantages of commodity hardware & open source software:

- Cost advantage is music to CXO's ears.
- How about the savings your IT will enjoy from moving things to commodity hardware and leverage more open source platforms for cost effective ways to achieve enterprise level computations and beyond.
- No more overpaying of premium hardware when similar or better analytical processing could be done using commodity and open source systems.

5. Quick turnaround and less time bench:

- Complex processes and communication gives you hard time connecting with someone who could get the task done.
- Things take forever long and cost fortunes with substandard quality.
- A good big data and analytics strategy could reduce the proof of concept time smoothly and substantially.
- It reduces the burden on IT and gets more high quality, fast and cost effective solutions baked.
- So, you will waste less time waiting for analysis / insights and more time digging through complex data, and use it for better insights and analyses which was never heard of before.

6. Automation to backfill redundant/mundane tasks:

- How about doing something to the 80% of time that is wasted in data cleaning and pre-processing.
- There is great deal of automation that could be take part and sky rocket enterprise efficiency. Less manual time spent on data prep and more time is spent on doing analysis that would have substantial ROI compared to mundane data preps and monotonous tasks.

7. Optimize workforce to leverage high talent cost:

- Big data & analytics strategy ensures current workforce is leveraged to its core in handling enterprise big data and also ensures right number of data scientists is involved with clearer sight to their contribution and their ROI.

2.3 Syllabus topic - Big data Stack:

- There are thousands of big data architectures that can be used to define categorize and share but, in the case of LAMP stack, the ultimate decision comes down to cost and scalability.
- There are different layers involve in big data stack. Following layers are involve in big data stack
  1. Data Layer
  2. Integration Layer
  3. Analytical Layer
  4. Predictive Analytics Layer

1. Data Layer:

- The bottom layer of the stack is the foundation and is known as the data relational databases. Hadoop and NoSQL have emerged as the most popular open source technologies due to their cost-effectiveness and scalability, however they fall short in the lightning fast analytics proceeded by software such as Vertica

2. Integration Layer:

- The second layer is the integration layer which is responsible for pulling and dissecting data from a variety of different sources. Today it is not enough to pull data from one source so it is common to have "360 applications" which pull data from thousands of data points meshes it together and transforms it to allow data integration to happen.
- A number of vendors are producing features like Sqoop and Flume to perform this Action.

2.4 Syllabus topic – Typical Analytical Architecture:

- There is need of workspace to Data Science projects which are basically built for experimenting with data, with flexible as well as agile data architectures.
- Numbers of organizations will possess data warehouses which give excellent support for reporting in traditional way and signified data analysis activities but problems arise when there is need of more robust analysis.
- For the purpose of data sources to be loaded into the data warehouse there is need that the data should be well understood, in structured format, and normalized with the suitable data type definitions.
- Even if such type of centralization leads to security, backup, and failover of highly critical data.

- It also indicates that the data should carry out effective pre-processing as well as checkpoints prior to entering in this of controlled environment, which does not allow its use in data exploration and iterative analytics.
- As a result of such level of control on the enterprise data warehouse (EDW),it is possible that some more local systems emerge in the role of departmental warehoused and local data marts which are created by business users for the purpose of accommodating their requirements of flexible analysis.
- There may not be similar constraints regarding security and structure on their local data marts as of the main EDW and let users to implement some level of more in-depth analysis.
- Still such off system exists in isolation, usually are unsynchronized or not integrated with other types of data stores and also may not be backed up.
- For BI and reporting purposed data is acceded by more applications in the environment of enterprise in the data warehouse.
-

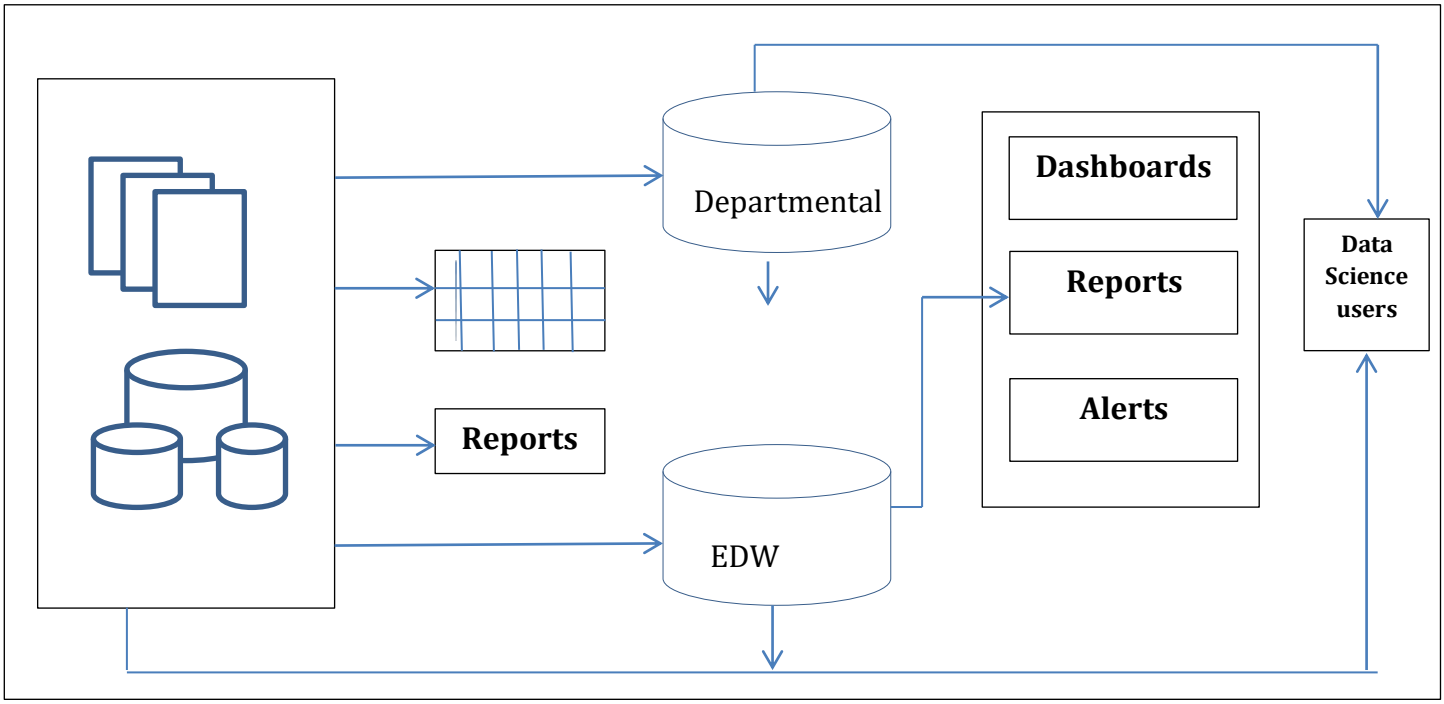

Figure: Typical analytic Architecture

- These are considered as high priority operational processed which retrieve critical data feeds from the data warehoused and repositories.
- When this workflow ends analysts obtain data which is basically provisioned for their downstream analytics.

- It is not allowed for users to run custom or intensive analytics on production databases, analysts have to generate data extracts from the enterprise data warehouse (EDW) for the purpose of analysing data offline in R different local analytical tools.
- Number of such tools is limited to in-memory analytics on desktops analysing samples of data instead of whole population of a dataset.
- Since the base of these analyses is data extracts, they are located in a separate location and the outcomes of the analysis and any insights on the quality of the data or anomalies-rarely are sent to the main data repository
- The moving speed of data is slow in EDW and also the process of chanting data schema takes longer became the process of accumulation of new data sources take more time in the EDW due to the through validation and structuring process.
- Mostly the departmental data warehouses are designed for a precise purpose and set of business requirements but when data is increased by time to time some them may be put into existing schemas to enable BI and generation of OLAP cubes for the process of analysis and reporting.
- Even if the EDW accomplish the objectives of reporting and rarely the generation of dashboards, EDWs normally restrict the capacity of analysis to iterate on the data in a unique nonproduction environment where they can carry out in-depth analytics or perform analysis on the data which is in unstructured form.
- The describe data architectures are developed for the purpose of storing as well as processing mission-critical data, most of the traditional data architectures restrain the data and more sophisticated analysis.

2.5 Syllabus topic - Virtualization & Big Data

- Solving big data challenges requires the management of large volumes of highly distributed data stores along with the use of compute- and data-intensive applications.
- Virtualization provides the added level of efficiency to make big data platforms a reality.
- Although virtualization is technically not a requirement for big data analysis, software frameworks are more efficient in a virtualized environment.
- Virtualization has three characteristics that support the scalability and operating efficiency required for big data environments:
  1. Partitioning
  2. Isolation
  3. Encapsulation

1. Partitioning:

– In virtualization, many applications and operating systems are supported in a single physical system by partitioning the available resources.

2. Isolation:

– Each virtual machine is isolated from its host physical system and other virtualized machines. Because of this isolation, if one virtual instance crashes, the other virtual machines and the host system aren't affected.

– In addition, data isn't shared between one virtual instance and another.

3. Encapsulation:

– A virtual machine can be represented as a single file, so you can identify it easily based on the services it provides.

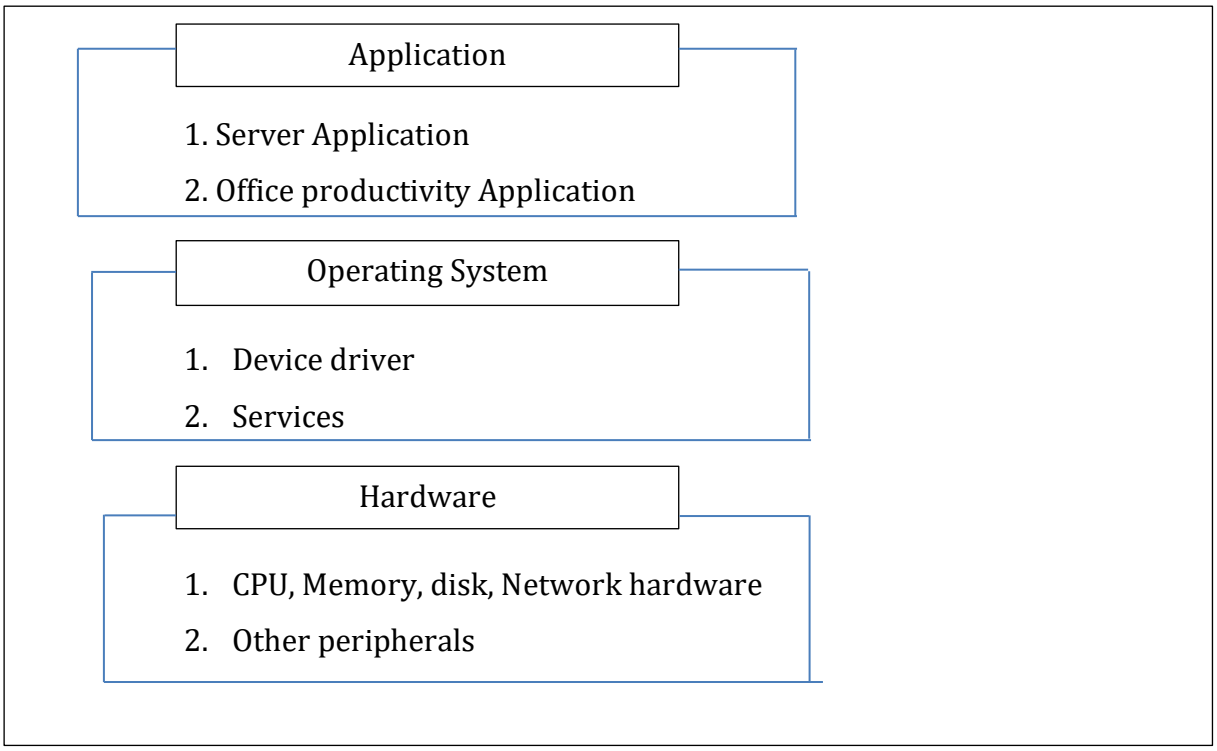

Figure: The various Virtualization Layers

2.5. Types of Virtualization

There are different types of virtualization; following are the some of the Virtualizations

1. Big Data Server Virtualization:
2. Big Data Application Virtualization:
3. Big Data Network Virtualization:
4. Big Data and storage Virtualization:
5. Big Data processor and Memory Virtualization:

1. Big Data Server Virtualization:

- In server virtualization, one physical server is partitioned into multiple virtual servers.
- The hardware and resources of a machine — including the random access memory (RAM), CPU, hard drive, and network controller — can be virtualized into a series of virtual machines that each runs its own applications and operating system.
- A virtual machine (VM) is a software representation of a physical machine that can execute or perform the same functions as the physical machine.
- A thin layer of software is actually inserted into the hardware that contains a virtual machine monitor, or hypervisor.
- Server virtualization uses the hypervisor to provide efficiency in the use of physical resources. Of course, installation, configuration, and administrative tasks are associated with setting up these virtual machines.
- Server virtualization helps to ensure that your platform can scale as needed to handle the large volumes and varied types of data included in your big data analysis. You may not know the extent of the volume needed before you begin your analysis.
- This uncertainty makes the need for server virtualization even greater, providing your environment with the capability to meet the unanticipated demand for processing very large data sets.
- In addition, server virtualization provides the foundation that enables many of the cloud services used as data sources in a big data analysis. Virtualization increases the efficiency of the cloud that makes many complex systems easier to

2. Big Data Application Virtualization:

- Application infrastructure virtualization provides an efficient way to manage applications in context with customer demand.
- The application is encapsulated in a way that removes its dependencies from the underlying physical computer system.
- This helps to improve the overall manageability and portability of the application.
- In addition, the application infrastructure virtualization software typically allows for codifying business and technical usage policies to make sure that each of your applications leverages virtual and physical resources in a predictable way.
- Efficiencies are gained because you can more easily distribute IT resources according to the relative business value of your applications.

- Application infrastructure virtualization used in combination with server virtualization can help to ensure that business service-level agreements are met.
- Server virtualization monitors CPU and memory usage, but does not account for variations in business priority when allocating resources.

3. Big Data Network Virtualization:

- Network virtualization provides an efficient way to use networking as a pool of connection resources. Instead of relying on the physical network for managing traffic, you can create multiple virtual networks all utilizing the same physical implementation.
- This can be useful if you need to define a network for data gathering with a certain set of performance characteristics and capacity and another network for applications with different performance and capacity.
- Virtualizing the network helps reduce these bottlenecks and improve the capability to manage the large distributed data required for big data analysis.

4. Big Data and Storage Virtualization:

- Data virtualization can be used to create a platform for dynamic linked data services. This allows data to be easily searched and linked through a unified reference source.
- As a result, data virtualization provides an abstract service that delivers data in a consistent form regardless of the underlying physical database.
- In addition, data virtualization exposes cached data to all applications to improve performance.
- Storage virtualization combines physical storage resources so that they are more effectively shared. This reduces the cost of storage and makes it easier to manage data stores required for big data analysis.

5. Big Data Processor and Memory Virtualization:

- Processor virtualization helps to optimize the processor and maximize performance. Memory virtualization decouples memory from the servers.
- In big data analysis, you may have repeated queries of large data sets and the creation of advanced analytic algorithms, all designed to look for patterns and trends that are not yet understood.
- These advanced analytics can require lots of processing power (CPU) and memory (RAM).
- For some of these computations, it can take a long time without sufficient CPU and memory resources.

2.6 Syllabus topic – Virtualization Approaches:

1. Elements of Big Data Solution:

- Big Data solution that must be considered if that project is to produce useful results and not just well processed, well reported data that appears with the guise of credibility but really is warmed-over garbage.

2. Types of Data:

- A Big Data project needs to be able to gobble up static objects and files such as documents, spread sheets and even presentations.
- It must be able to consume on-going streams of data coming from POS devices, smartphones, tablets and many other types of intelligent devices.
- Social media data must be harvested from sources such as LinkedIn and Twitter.
- Don't forget data that can be found on Web pages, manufacturing data from the enterprise's own systems, and data from stock reporting systems to weather reporting systems.

3. Transformation, Normalization and other magic

- Data items coming from different source are likely to be in formats designed to satisfy the needs of the original application; not the use intended by the Big Data application.
- You'd be amazed at how many different formats data items such as ID numbers, currency and other types of data are presented in.
- All data items must be transformed from their current format to one designed for Big Data analysis. Knowing what to transform and the proper final form can be a major challenge all by itself.

4. Modelling, Framework and simulation:

- There are an amazing number of modelling, data frameworks and simulation tools available, tools for data regression, neural networks, data clustering, decision trees and more. Selecting the proper tool for each type of data is critical.

5. Making Prediction:

- Once the proper data is selected, formatted, transformed and analyzed, systems can offer predictions about what will come next.
- Since these are heavily modelled items, they must be tested and measured against real-world data. Eventually, a model will emerge that provides useful guidance.

6. Reporting and Visualization:

- Most people are not very good at teasing out important insight from massive columns of numbers. They *are* pretty good at being able to look at a figure or graph and see something that sticks out as being different.
- Presenting the information in the proper form can make the difference between mounds of useless data and something offering quick and clear insight.

2.7. Business Intelligence vs. Data science:

- Nowadays to handle the various types of business problems, organization has to be more analytical and data driven
- Business drivers for advanced analytics

Sr.No	Business Driver	Example
01	Optimize business operations	Sales, Pricing, Profitability, efficiency
02	Identify Business risk	Customer churn, fraud, default
03	Predict new business opportunities	Upsell, cross-sell, best new customer prospects
04	Comply with laws or regulatory requirements	Anti-Money Laundering Fair Lending, Basel II-III, Sarbanes Oxley(SOX)

- We can observe that there are four generalized categories of common business problems necessary for them to leverage advanced analytics for the purpose of creating competitive advantage.
- Instead of just working on standard reporting on their areas, it is possible for organizations to apply some advanced analytical techniques for the purpose of optimizing processes and get more value from the usual tasks.
- The first three examples are not concerned with new problems.
- Organizations have been trying to provide good service increase sales for many years.
- What exactly new advantage is the chance to combine advanced analytical techniques with big data so as to generate more impactful analyse for the various traditional problems.
- The last example is concerned with various emerging regulatory requirements
- There are number of compliance as well as regulatory laws present for decades but new more requirements are added year by year which leads to increase in complexity and data requirements for organizations.
- Laws which represent the AML (Anti-Money Laundering) and fraud preventions need s some more advanced analytical techniques for the purpose of comply with and manage properly.

2.7.1 BI Vs. Data Science:

- The four business drivers which we have discussed in previous section need a variety of analytical techniques to address them properly.
- There are number of ways which helps to compare these groups of analytical techniques.
- One way for the evaluation of the type of analysis being carried out is to observe the time horizon and the type of analytical approaches being used.
- BI usually provides reports dashboards and queries on business questions for the current period or in the past
- BI systems helps to simplify to answer questions regarding quarter-to date revenue, progress towards quarterly targets and known quantity of given product was sold in a prior quarter or year
- These question considered as closed-ended and explain current or past behaviour normally by the process of aggregating historical data and grouping it in some way
- BI offers hindsight and little insights and usually answers questions regarding “when” and “where” events occurred.
- When compared with BI it is found that Data Science like to use disaggregated data with a more forward looking exploratory techniques concentrating on analysing the present and enabling informed decisions about the future.
- Instead of aggregating historical data to search for quantity of product sold in the previous quarter it is possible for a team to employ Data Science techniques like time series analysis.
- Such techniques help to guess future product sales and revenue more precisely as compared to extending a simple trend line.
- Also Data Science considered as ore exploratory in nature and may like to refer scenario optimization for the purpose of dealing with more open-ended questions.
- This approach helps to get insights into current activity and foresight into events while usually concentrating on questions regarding “how” and “why” event occur.
- Where BI problems need highly structured data which has been organized in rows and columns for accurate reporting. Data Science projects mostly refer various kinds of data sources including large or unconventional datasets
- Based on the future goals of organizations it may prefer to board on a PI project if there is reporting dashboards creation o simple visualizations or it may prefer to board on Data Science projects if it required to do a more sophisticated analysis with datasets which are in the form of disaggregated or distinct.

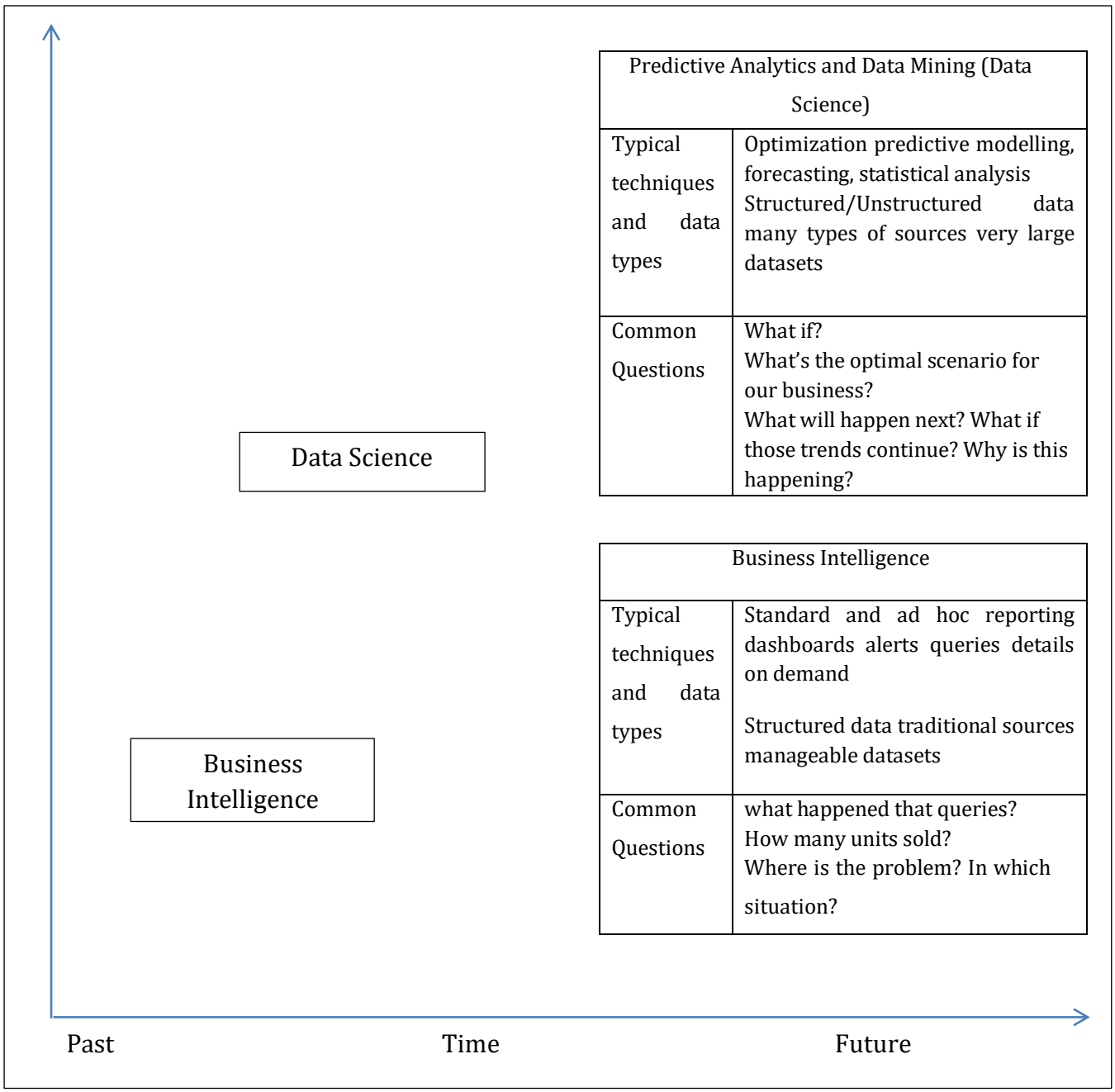

Figure: Comparing BI with Data Science

2.8 Syllabus topic - Applications of Big data analytics.

1. Transportation:

- Transportation system is one of the major areas where real-time data analytics is very much needed because of the required processing of data within a very short time for various purposes or services.
- For example, real-time data analytics of current traffic conditions could provide very useful information to the end user within a very short time for making a efficient decision, such as

- route selection for the destination
- estimate time to reach to the destination
- Changing route because of any kind of sudden incidents like accident, or roadblocks.
- Quick delivery of orders for any kind of goods, like pizza delivery, or emergency postal delivery.
- Dynamic time calculation for emergency vehicles like ambulance, fire service car, police van for the quick arrival to the destination.

2. Stock Market:

- A stock market is the aggregation of buyers and sellers where they buy or sale shares or stocks of listed companies.
- In stock market huge numbers of data are generated in every working day. These data is not only big in volume but also very dynamic. By analysing these data in real-time both buyers and sellers could be benefited and it also helps to detect fraud and illegal activities which certainly improves the performance of the stock market.
- Below is listed some points which can be achieved by real-time data analytics of stock market.
- Prediction of share prices before actual changes occurs in share prices. So that timely selling or buying of shares can be done for higher profit margin.
- Earlier decision making ability for buying or selling shares.
- Financial threads detection in quick time.
- Detection of illegal activities in market which helps to improve market performance.
- Automated trading of shares and threads detection system, which could increase number of buyer and seller in the market.
- All these merits of stock market can be achieved if real-time data analytics of stock market is possible. Otherwise it will take longer time if it is done manually.
- As a result it will neither help the buyer nor the seller to earn higher profit or not even the market itself to detect threads to improve market performance.