UNIT I

CHAPTER 1

Introduction and Life Cycle

Syllabus Topics

Introduction : Big data overview, state of the practice in Analytics - BI Vs Data Science, Current Analytical Architecture, drivers of Big Data, Emerging Big Data Ecosystem and new approach.

Data Analytic Life Cycle : Overview, phase 1 - Discovery, Phase 2 - Data preparation, Phase 3 - Model Planning, Phase 4 - Model Building, Phase 5 - Communicate Results, Phase 6 - Opearationalize. Case Study : GINA.

Syllabus Topic : Big Data Overview

1.1 Big Data Overview

Q. 1.1.1 Write a short note on Big Data. (Refer section 1.1) **(4 Marks)**

- Now a day the amount of data created by various advanced technologies like Social networking sites, E-commerce etc. is very large. It is really difficult to store such huge data by using the traditional data storage facilities.
- Until 2003, the size of data produced was 5 billion gigabytes. If this data is stored in the form of disks it may fill an entire football field. In 2011, the same amount of data was created in every two days and in 2013 it was created in every ten minutes. This is really tremendous rate.
- In this topic, we will discuss about big data on a fundamental level and define common concepts related to big data. We will also see in deep about some of the processes and technologies currently being used in this field.

Big Data

DEFINITION

***Big data** means huge amount of data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big Data is complex and difficult to store, maintain or access in regular file system. Big Data becomes a complete subject, which involves different techniques, tools, and frameworks.*

Sources of big data

There are various sources of big data. Now days in number of fields such huge data get created. Following are some of the fields.

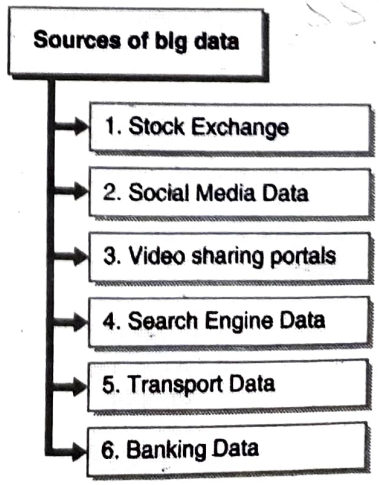

Fig. 1.1.1 : Sources of big data

➜ **1. Stock Exchange**

The data in the share market regarding information about prices and status details of shares of thousands of companies is very huge.

➜ **2. Social Media Data**

The data of social networking sites contains information about all the account holders, their posts, chat history, advertisements etc. On topmost sites like facebook and whatsapp, there are literally billions of users.

➜ **3. Video sharing portals**

Video sharing portals like YouTube, Vimeo etc. contains millions of videos each of which requires lots of memory to store.

➜ **4. Search Engine Data**

The search engines like Google and Yahoo holds lot much of metadata regarding various sites.

➜ **5. Transport Data**

Transport data contains information about model, capacity, distance and availability of various vehicles.

➜ **6. Banking Data**

The big giants in banking domain like SBI or ICICI hold large amount of data regarding huge transactions of account holders.

☞ Categories of Data

– The data can be categorized in three types.

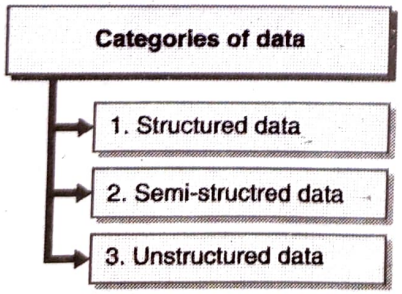

Fig. 1.1.2 : Categories of data

➜ **1. Structured Data**

This type of data is stored in relations (tables) in Relational Database Management System.

➜ **2. Semi-structured Data**

This type of data is neither raw data nor typed data in a conventional database system. A lot of data found on the web can be described as semi-structured data. This type of data does not have any standard formal model. This data is stored using various formats like XML and JSON.

➜ **3. Unstructured Data**

This data do not have any pre-defined data model. The data of video, audio, Image, text, web logs, system logs etc. comes under this category.

☞ Important issues regarding data in traditional file

– In general there are some important issues regarding data in traditional file storage system.

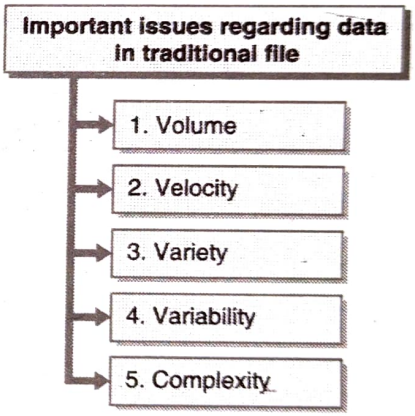

Fig. 1.1.3 : Important issues regarding data in regarding data in traditional file

➜ **1. Volume**

Now a day the volume of data regarding different fields is high and potentially increasing day by day. Organizations collect data from a variety of sources, including business transactions, social media and information etc.

➜ **2. Velocity**

The configuration of system with single processor, limited RAM and limited storage capacity cannot store and manage high volume of data.

➜ **3. Variety**

The form of data from different sources is different.

➜ **4. Variability**

The flow of data coming from sources like social media is inconsistent because of daily emerging new trends. It can show sudden increase in size of data which is difficult to manage.

➜ **5. Complexity** As the data is coming from various sources, it is difficult to link, match and transform such data across systems. It is necessary to connect and correlate relationships, hierarchies and multiple data linkages of the data. All these issues are solved by the new advanced **Big Data Technology.**

1.1.1 Defining Data Science and Big Data

Q. 1.1.2	Define the term data science. (Refer section 1.1.1)	(2 Marks)
Q. 1.1.3	Define Big Data. (Refer section 1.1.1)	(2 Marks)

☞ **Defining Data science**

DEFINITION

***Data science** is a field of Big Data which searches for providing meaningful information from huge amounts of complex data. Data science is a system used for retrieving the information in different forms, either in structured or unstructured.*

- Data Science unites different fields of work in statistics and computation in order to understand the data for the purpose of decision making.

☞ **Defining Big Data**

DEFINITION

***Big Data** is described as volumes of data available in changing level of complexity, produced at different velocities and changing level of ambiguity, that cannot be processed using conventional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions.*

- Data that can be defined as Big Data comes from variety of fields such as machine-generated data from sensor networks, nuclear plants, airplane engines, and consumer-driven data from social media.
- The producers of the Big Data that resides within organizations include legal, sales, marketing, procurement, finance, and human resources departments.

1.1.2 Examples of Big Data Applications

Q. 1.1.4	List the examples of big data. (Refer section 1.1.2)	(2 Marks)
Q. 1.1.5	Explain the examples of big data. (Refer section 1.1.2)	(6 Marks)

There are various big data applications as follows :

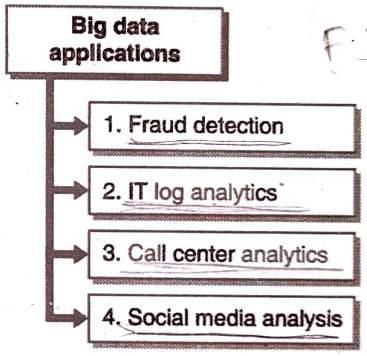

Fig. 1.1.4 : Big data applications

➜ **1. Fraud detection**

- Fraud detection is a Big Data application example for businesses which has operations like any type of claims or transaction processing.
- Number of times the detection of fraud is concluded long after the fact. At this point the damage has been already done, all that's left is to decrease the harm and revise policies to prevent it in future.
- The Big Data platforms can analyze claims and transactions of businesses. They identify large-scale patterns across many transactions or detect anomalous behaviour of some users. This helps to avoid the fraud.

➜ **2. IT log analytics**

- An enormous quantity of logs and trace data is generated in IT solutions and IT departments. Many times such data go unexamined: organizations simply don't have the manpower or resource to go through all such information.
- Big data has the ability to quickly identify large-scale patterns to help in diagnosing and preventing problems. It helps the organization with a large IT department.

➜ **3. Call center analytics**

- Now we turn to the customer-facing Big Data application examples, of which call center analytics are particularly powerful.
- Without a Big Data solution, much of the insight that a call center can provide will be ignored or exposed later.
- By making sense of time/quality resolution metrics, the Big Data solutions are able to identify recurring problems or customer and staff behaviour patterns. Big data can also capture and process call content itself.

➜ **4. Social media analysis**

- With the help of Social media we can observe the real-time insights into how the market is responding to products and campaigns.
- With the help of these insights, it is possible for companies to adjust their pricing, promotion, and campaign placement to get optimal results.

1.1.3 Data Explosion

Q. 1.1.6 Explain Data explosion in detail. (Refer section 1.1.3) **(4 Marks)**

DEFINITION

***The data explosion** is nothing but the rapid growth of the data. One reason to this explosive growth of data is innovation.*

- The Innovation has changed the way we engage in business, provide services, and the associated measurement of value and profitability.
- There are three basic trends available which becomes very essential to build up the data in the last few years. They are :

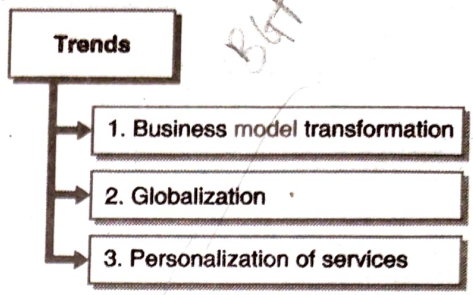

Fig. 1.1.5 : Basic Trends

➜ **1. Business model transformation**

- The primary business models have been transformed through globalization and connectivity.
- Modern companies can be moved towards the service oriented technologies rather than product oriented.
- In the service oriented the value of the organization from customers point of view is measured by how much the service is effective instead of how much product is useful.
- What this transformation commands means to every business is that you want to produce more data in terms of products and services to provide to each segment and channel of customers, and to hold lot of data from every customer, consisting of social media, surveys, forums, feedback from customers, call center, competitive market research, and many more.
- The amount of data created and stored by each organization today beats what the same organization produced prior to the business transformation.
- Now the data which is primary or having the higher priority are kept in center and the supporting data which is required but not available or accessible previously now can be available and also accessible with the help of multiple channels.
- Here the volume equation of data exploding to Big Data comes in the picture.

➜ **2. Globalization**

- Globalization is a key trend that has radically changed the commerce of the world, starting from manufacturing to customer service.
- It has also changed the variations and formats of data.

3. Personalization of services

- Business transformation's maturity index is measured by the degree of personalization of services and the value perceived by their customers from such transformation.
- Business transformation's maturity index model is one of the main causes for the velocity of data that is generated.

New sources of data

- As the technologies are growing, the data can be generated from various sources such as social media, mobile devices, sensor media and many more which were not present before.
- The appearances of newer business models and the aggressive expansion of technology capabilities over the last decade or more has covered the way for incorporating all of the data across the enterprise into one holistic platform to create a significant and appropriate business decision support platform.

1.1.4 Data Volume

Q. 1.1.7	Explain Data volume. (Refer section 1.1.4)	(6 Marks)
Q. 1.1.8	Write a short note on data volume. (Refer section 1.1.4)	(4 Marks)

DEFINITION

***Data volume** in the big data can be defined as the amount of data that is generated in continuous manner.*

- There are different data types available with different sizes.
- For example, to store a blog text data can requires a few kilobytes, whereas a voice calls or video files can requires a few megabytes, and sensor data, machine logs, and clickstream data can requires in gigabytes.

Different sources of data

- The following sub points describe some of the examples of data generated by different sources.

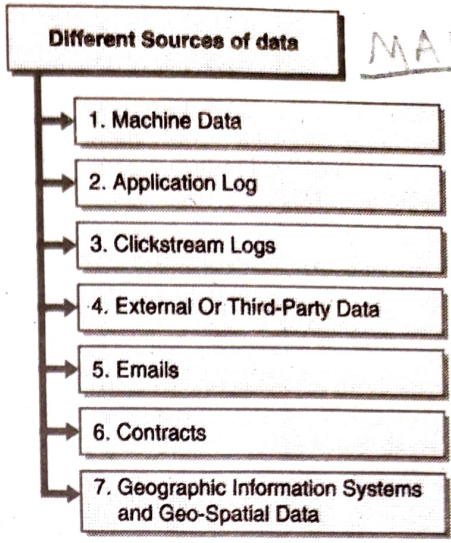

Fig. 1.1.6 : Different sources of data

1. Machine Data

- Every device which we are using can able to generate the large amount of the data. Such data can contain the usage and the behavior of the user who uses these devices or machines.
- Machine-generated data is often differentiated by a fixed pattern of numbers as well as text, which occurs in a rapid-fire fashion.
- There are a number of examples of machine-generated data; consider an example, a robot can send signal when it perform any movement and that movement pattern is fixed.
- Opposite to this example a dredging machine which is used to work on road can send the signals only once at the end of the day that how much movement it had during the whole day, how much payload it had moved and also machine status.
- Another example is that sensors which are placed at the top of building controls; the heating and cooling can send the signal throughout the day. Opposite to this the sensors at the automobiles send various signals based on type of road, speed of driving, weight, and more to support centers.
- All types of the transmitting devices can vary from small volume explode to the huge volume explode.

2. Application Log

- Application log is one of the types of machine-generated data.
- Different devices can generate logs at different velocity and formats.
- Now days tablets, cellular phones can generates the logs for every activity from the device at any time of the day, including geographic information, data type, access type, activity period, and so on.

3. Clickstream Logs

- The clickstream log can be useful to capture the usage information of the web page.
- The clickstream log gives the insight into what a user is doing on the web page, and can provide data that is extremely useful for behavior and usability analysis, marketing, and general research.
- One of the leading problems with the data is the volume, since the actions can be logged in several places, such as web servers, routers, proxy servers, ad servers, etc.
- Current tools can only collect partial information from this data since data formats differ across the system and the pure volume makes it tough for current technologies to process this data.

4. External or Third-Party Data

- There are many data sets which can be acquired by the organizations from the external sources.
- Although some of the data is structured, most of the data has different formats and often comes in heavy volumes.
- **Example :** Weather data.

5. Emails

- The customers or employees of an organization can generate the huge amount of emails on regular basis.
- These emails can contains all the benefits required by the organization and so they need to be managed.
- To justify the power and management of emails, there are various examples present such as intellectual property, competitive analysis and so on.
- If organizations do not realize the associated risks, problems, and penalties associated with email management, it is time to set policies in place.
- Practical management of this content is desirable for every enterprise.

6. Contracts

- Another type of data which can be generated on regular basis or we can say as daily basis by enterprises is contracts.
- The contracts types can be categorized as supplier, human resource, and customer and so on. According to the type of the contract the content of it also gets changed and results in large volume of data.
- According to the language and the semantics contained inside every contract gives directions for different types of legal suggestions.
- The Legal teams have to waste their time to find out the values which may be useful for the enterprise or not based on the nature of the data exposed along with the significance of the time of discovery.

7. Geographic Information Systems and Geo-Spatial Data

- One of the popular device and smart-phone application is the global positioning system (GPS), which uses geo-spatial data to direct anyone from point A to point B.
- In addition, the GPS is built with features like guided voice navigation, points of interest (POI), and so on.
- Another rising feature is the addition of GPS to cameras and camcorders. With the help of this feature, one can set locations where the picture was taken along with dates and other information. This feature can be very useful for the journalists.
- But this kind of real-time data interaction needs huge amount of data to transmit back and forth every second for millions of customers across the world.

- Today the data generated by the social media is huge for processing because at every minute the customers generate new data.
- Social media sites offer not only customer's viewpoint, but also competitive positioning, trends, and access to communities created by common interest.
- Now the organizations make the use of social media pages to personalize marketing of products and services to each customer.

1.1.5 Data Velocity

Q. 1.1.9 Explain data velocity. (Refer section 1.1.5) **(6 Marks)**

DEFINITION

***Data velocity** in big data can be defined as the flow of the data from various sources such as networks, human resources, social media etc. The data can be huge and flows in continuous manner.*

- With the initiation of the big data, it becomes very important to understand the velocity of data.
- The fundamental reason for this arises from the fact that in the early days of data processing, we used to analyze data in batches, obtained over time.
- Traditionally the data is divided into chunks having the fixed size and can be processed via the different layers starting from source to targets, and whatever the result is produced at the end is stored in a data warehouse for future use in reporting and analysis.
- When the flow of input data is at constant rate and the results are used for examining with considering all the process delay then the above data processing technique in batches or micro batches works very well.
- The size of the batches is fixed; due to this reason it is very easy to maintain the scalability and throughput of the data processing architecture.
- In the case of Big Data, the data streams in a constant manner and the result sets are useful when the acquisition and processing delays are small.

Consider the following some examples of data velocity.

☞ **Examples of data velocity** AS MS

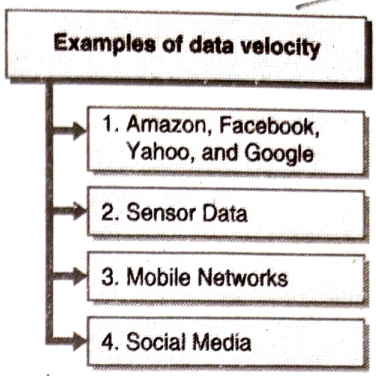

Fig. 1.1.7 : Examples of data velocity

➜ **1. Amazon, Facebook, Yahoo, and Google**

- The business models accepted by the companies such as Amazon, Facebook, Yahoo, and Google now became the de-facto business models for nearly all web-based companies, works on the fact that by capturing the customer clicks and navigations on the website, you can deliver personalized browsing and shopping experiences.
- In this process of clickstreams there are lot of clicks (thousands or millions of clicks) collected from users at every second, amounting to huge volumes of data.
- This collected data can be processed, divided, and represented to study population behaviors based on time of day, geography, efficiency of advertisement, click behavior, and guided navigation response.
- The result sets of these models can be stored to create a superior experience for the next set of clicks demonstrating similar behaviors.
- The most important example for the big data velocity is the velocity of data generated by user clicks on any website.

➜ **2. Sensor Data**

- One more major example of data velocity comes from a range of sensors like Global Positioning System (GPS), tire-pressure systems, On-Star-vehicle, based on geo-spatial and location based intelligence associated with the sensor on the automobile, heating and cooling

systems on buildings, smart-meters, mobile devices, biometric systems, airplane sensors and engines.

- The data that is produced from sensor networks can vary from gigabytes to terabytes per second.
- Consider an example of airplane in which each flight creates a huge amount of data from the airplane engine sensor. Here lot of data needs to be read because this data is useful for the data statistical modeling purpose.

➜ **3. Mobile Networks**

- Today we can share the pictures, music and also data with the help of mobile devices.
- The entire data that can be transmitted through the mobile networks provides the intuition to the providers depending on the performance of the their network, capacity of each tower i.e. how much data can be processed at every tower, time , related geographies, latency and so on.
- The network can be crashed as the movement of the velocity of data is huge every time. To improve the quality of service we have to study the data movement.

➜ **4. Social Media**

- As we know that there are several social media sites which generates the data having the different velocity and also having the different formats.
- **Examples :** Facebook, YouTube can contain the post of any size whereas the Twitter contains the post of fixed size.
- Only the size of the post is not important, also we have to understand how many time the post can be shared or it can be forwarded and how much resultant data it collected.
- Sometimes a post can be viral creating a millions of reposts which will result in the huge amount of data needs to process.

1.1.6 Big Data Infrastructure and Challenges

Q. 1.1.10 Explain the challenges of big data infrastructure.

(Refer section 1.1.6) **(6 Marks)**

The following sub-points illustrates the four different areas that have developed and yet prove challenging. They are :

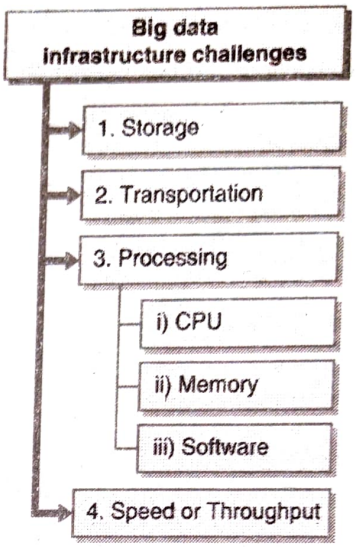

Fig. 1.1.8 : Big data infrastructure challenges

➜ **1. Storage**

- The first and major problem to big data is obvious a storage. As the big data is increased very rapidly, there is need to process this huge data as well as to store it.
- Also you require the extra 0.5 times storage to process and store the intermediate result set.
- Consider an example : The U.S. Census Bureau's has described the first storage problem in the 1800s that they are not able to store and report the statistics.
- Collecting data and performing the calculation to trace out the desired output were time consuming and error-prone having no ability to do correction in the errors and enhance the processing efficiency.
- Herman Hollerith developed a card with holes indicating numerous data formats that have the capability that when it can be punched it would generate responses.
- Every collection of holes which are punched, produce a pattern of answers to asked questions, having the capability to gather patterns into groups.
- With the help of a processing machine he had developed, he was capable to process data by reading the cards and caches the input for future use.

- This type of processing can be considered as one of the primitive type for offline storage.
- When the population is increased and the extra changes are made to the questions then it becomes very challenging to the storage problem.
- To handle this problem, Hollerith categorized data as either static or dynamic, providing the basic master data and transaction data type of classification.
- As technology advances we will use the magnetic tape or disk-based storage rather than punch cards.
- When we are dealing with the transaction processing and data warehousing, the storage becomes the problem.
- Because of the design of the underlying software we cannot use the storage which can be available on a disk.
- One more problem with storage is the cost per byte. Usually, fast-performing storage has been costly and the cost takes up to the overall total cost of ownership.

➜ **2. Transportation**

- One of the major issues that always tackle the data world is transporting data between different systems and then storing it or loading it into memory for performing the operations.
- This constant movement of data has been one of the reasons that structured data processing developed to be limiting in nature, where the data had to be transported between the compute and storage layers.
- As the networking technologies are improving still it could not resolve this problem, although it facilitated the bandwidth of the transport layers to be much larger and highly scalable.

➜ **3. Processing**

- Processing data requisites the capability to merge some type of logical and mathematical calculations together in one cycle of operation.
- Processing data can be classified into three main areas, they are as follows :

➜ **(i) CPU or processor**

- The Computer processing units have been developing from the early 1970s to today.
- In the every generation the speed of the computation and processing power has been improved with having the access to the large memory and also developed the architecture evolution inside the software layers.

➜ **(ii) Memory**

- The storage of data offline proved that it needs the storage evaluation as well as the management of the data within a disk.
- As the processor evaluations improving the capability of the processor, the memory has become cheaper and faster in terms of speed.
- According to the allocated memory to system, the processes resides in the system can change significantly.

➜ **(iii) Software**

- One of the main components of data processing is the software which is used to develop the programs to transform and process the data.
- Software's can be generated according to the generations considering the different layers starting from the operating system to different programming languages.
- In its lowest form the software converts sequenced instruction sets into machine language that is used to process data with the infrastructure layers CPU + memory + storage.
- Operating systems such as Linux which is open source is made available to everyone for innovation to develop additional capabilities to the base software platform which may influence the complete infrastructure and processing architecture improvements.

➜ **4. Speed or Throughput**

- The leading ongoing challenge is the speed or throughput of data processing. The term speed can be referred as a mixture of various layers of architecture such as hardware, software, networking, and storage.

- Every architecture layer has its limitations and when the limitations of all these layers are grouped, it becomes a challenge for the throughput of the data processing.
- When we are dealing with the data processing and data management, the speed or througtput can be a major problem from the financial point of view.

Syllabus Topic : State of the Practice in Analytics BI Vs Data Science

1.2 State of Practice in Analytics BI Vs Data Science

Q. 1.2.1 Write a short note on "State of practice in Analytics BI Vs Data Science". (Refer section 1.2) **(8 Marks)**

- Nowadays to handle the various types of business problems, organizations has to be more analytical and data driven.

Table 1.2.1 : Business Drivers for Advanced Analytics

Business Driver	Example
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, SarbanesOxley(SOX)

- In the given Table 1.2.1 we can observe that there are four generalized categories of common business problems which organizations usually faces where it becomes necessary for them to leverage advanced analytics for the purpose of creating competitive advantage.
- Instead of just working on standard reporting on these areas, it is possible for organizations to apply some advanced analytical techniques for the purpose of optimizing processes and get more value from ıe 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 analyses 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 prevention needs some more advanced analytical techniques for the purpose of comply with and manage properly.

Syllabus Topic : BI Vs Data Science

1.2.1 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 toward quarterly targets, and know quantity of given product was sold in a prior quarter or year.

- These questions considered as closed-ended and explain current or past behavior, normally by the process of aggregating historical data and grouping it in some way.
- BI offers hindsight and little insight 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 technique, concentrating on analyzing 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 more 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 insight into current activity and foresight into upcoming events, while usually concentrating on questions regarding "how" and "why" events occur.
- Where BI problems needs 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 an future goals of organization, it may prefer to board on a PI project if there is reporting, dashboards creation, or 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.

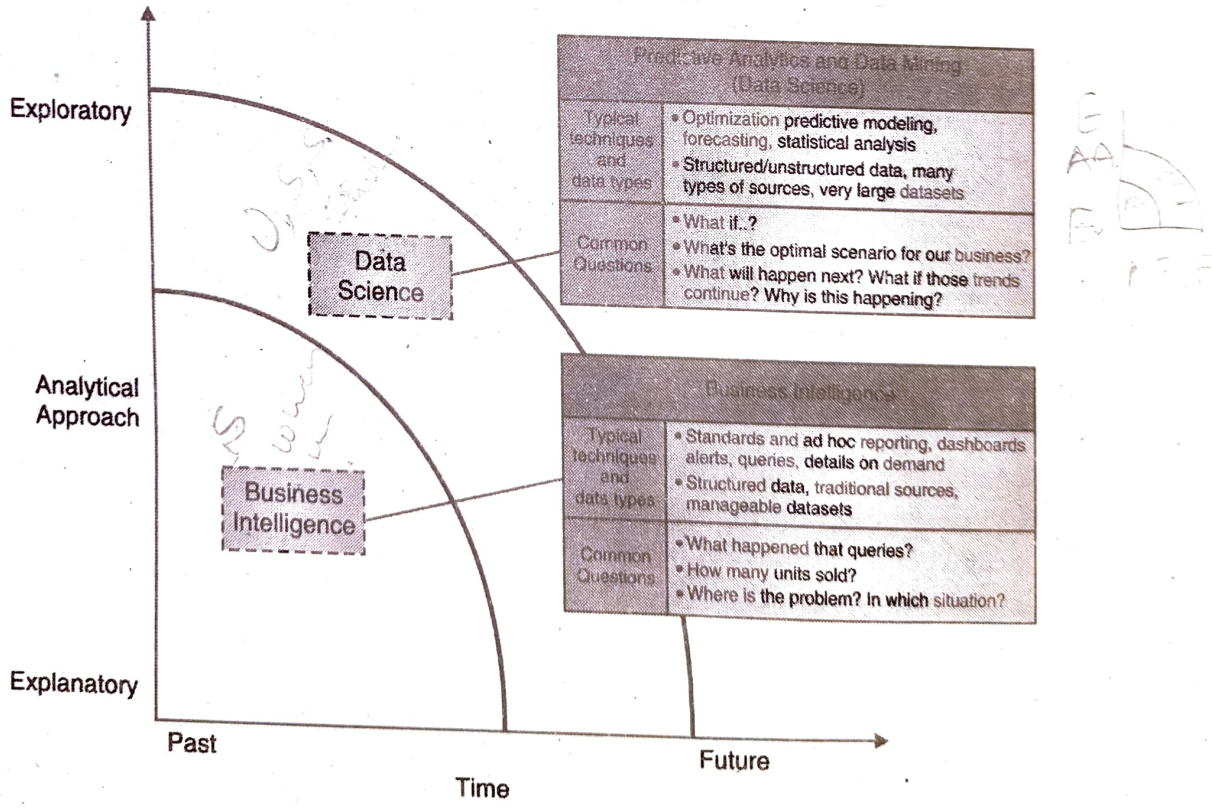

Fig. 1.2.1 : Comparing BI with Data Science

Syllabus Topic : Current Analytical Architecture

1.2.2 Current Analytical Architecture

Q. 1.2.2 Explain Current Analytical Architecture with suitable diagram. (Refer section 1.2.2) **(8 Marks)**

- 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.
- Number of organizations still posses data warehouses which give excellent support for reporting in traditional way and simplified data analysis activities but problems arise when there is need of more robust analysis.
- Fig. 1.2.2 illustrates typical data architecture as well as various challenges it presents to data scientists and other users who are trying to implement advanced analytics.

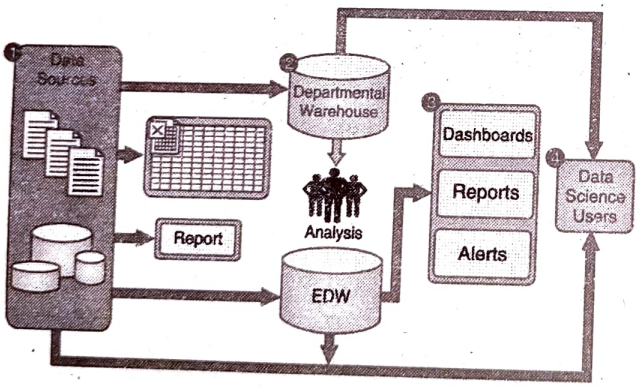

Fig. 1.2.2 : Typical analytic architecture

1. 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 preprocessing as well as checkpoints prior to entering in this sort of controlled environment, which does not allow its use in data exploration and iterative analytics.
2. 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 warehouses and local data marts which are created by business users for the purpose of accommodating their requirement of flexible analysis.

   There may not be similar constraints regarding security and structure on these local data marts as of the main EDW and let users to implement some level of more in-depth analysis.

   Still such one-off systems exist in isolation, usually are unsynchronized or non-integrated with other types of data stores, and also may not be backed up.
3. For BI and reporting purposes, data is accessed by more applications in the environment of enterprise in the data warehouse.

   These are considered as high-priority operational processes which retrieve critical data feeds from the data warehouses and repositories.
4. 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 analyzing data offline in R or different local analytical tools.
   - Number of such tools are limited to in-memory analytics on desktops analyzing 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 changing data schema takes longer because the process of accumulation of new data sources take more time in the EDW due to the thorough validation and data 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 of 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 objective of reporting and rarely the generation of dashboards, EDWs normally restrict the capacity of analysts 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 described data architectures are developed for the purpose of storing as well as processing mission-critical data, providing support to enterprise applications, and enabling corporate reporting activities.
- Even though reports and dashboards keeps their importance in organizations, most of the traditional data architectures restrain data exploration and more sophisticated analysis.
- Also there are various more implications for data scientists in traditional data architectures.
- It is difficult to reach and leverage high-value data, and activities such as predictive analytics and data mining last in line for data. Since the enterprise data warehouses are designed for central data management as well as reporting, the required data for analysis are usually preferred after operational processes.
- Data moves in the form of batches from enterprise data warehouse to local analytical tools. This workflow indicates that data scientists are restricted to carrying out in-memory analytics which will limit the size of the data sets which are allowed for them to use.
- Data Science projects will leave isolated as well as ad hoc, instead of centrally managed. The result of this isolation leads to the concept that organization will not be able to harness the power of advanced analytics in a scalable way, and the projects regarding Data Science will remain as nonstandard initiatives, which are usually not aligned with corporate business goals or strategy.
- The result of all such symptoms of the traditional data architecture leads to slow "time-to-insight" and decreases business impact which can be solved by an environment which supports advanced analytics in which data were more readily accessible and supported.
- One answer to this problem is to accept analytic sandboxes to enable data scientists to carry out advanced analytics in a controlled as well as sanctioned way.

Syllabus Topic : Drivers of Big Data

1.2.3 Drivers of Big Data

Q. 1.2.3 Write a short note on Drivers of Big Data. (Refer section 1.2.3) **(4 Marks)**

- Now before proceeding to Drivers of Big Data, we will see some previous history regarding data stores and the types of repositories as well as tools to manage these data stores.
- As illustrated in Fig. 1.2.3, in the 1990s the measurement unit for volume of information was terabytes.
- Most of the organizations like to analyze structured data in the form of rows and columns and refer relational databases and data warehouses for the purpose of managing large stores of enterprise information.
- In the next decade there was emergence of different types of data sources mostly productivity and publishing tools like content management repositories and network attached storage systems for the purpose of managing such type of information, and the size of data began to increase and the measurement unit changes to petabyte.
- In the 2010s, the information which organizations need to handle has broadened because of inclusion of several other kinds of data.
- In this decade, each and every element is leaving a digital footprint. Fig. 1.2.3 illustrates a summary perspective regarding the sources of Big Data created

by new applications and the size and growth rate of the data.

- The new applications which are able to generate data volumes in huge amount (exabyte scale), enables various opportunities for new analytics and driving new value for organizations.
- Nowadays there are various sources of data such as :
  - Medical regarding data like genomic sequencing and diagnostic imaging.
  - Several Photos as well as video uploaded to the WWW(World Wide Web)
  - Video surveillance, for example several video cameras spread across a city.
  - Mobile devices, which helps to search geospatial location of the users and also metadata about SMS, calls, and usage of various apps on smart phones
  - Smart devices which helps to get sensor-based collection of data from smart electric grids, smart buildings, and various kinds of public as well as industry infrastructures
  - Advanced IT devices which uses advanced technologies such as radio-frequency identification (RFID) readers, GPS navigation systems, etc.

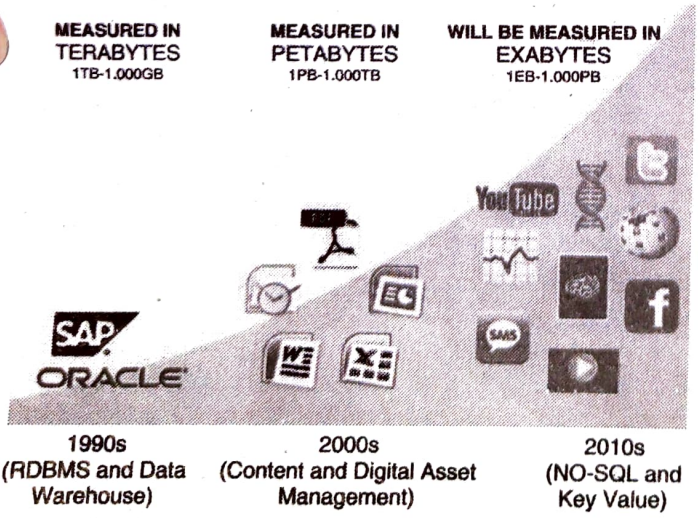

Fig. 1.2.3 : Data evolution and the rise of Big Data sources

  - In the Big data environment, huge amount of data is generated from many new sources. This overflow of data needs advanced techniques of analytics and new market players to grab benefit of new opportunities and new market dynamics.

Syllabus Topic : Emerging Big Data Ecosystem and New Approach

1.2.4 Emerging Big Data Ecosystem and New Approach

Q. 1.2.4 Explain Big Data Ecosystem. (Refer section 1.2.4) **(8 Marks)**

- Nowadays several organizations as well as different types of data collectors are realizing that the data gathered by them from individuals has intrinsic value and which leads to emergence of a new economy.
- As there is continuous evolution of this new digital economy, the market experiences the involvement of data vendors and data cleaners which use crowd sourcing for the purpose of testing the outcomes of machine learning techniques.
- Other vendors add value to open source tools and do the repackaging of them in a simpler way and present the tools to market.
- For the open source framework Hadoop, such value add is provided by Vendors like Cloudera, Hortonworks, and Pivotal.
- There are four main elements in the interconnected web under the roof of ecosystem. These are as follows :
- **Data devices** (shown in section (1) of Fig. 1.2.4) and the "Sensornet" collect information from various locations and constantly create new data about this data.
- For each and every gigabyte of new data generated, an extra petabyte of data is generated regarding that data.

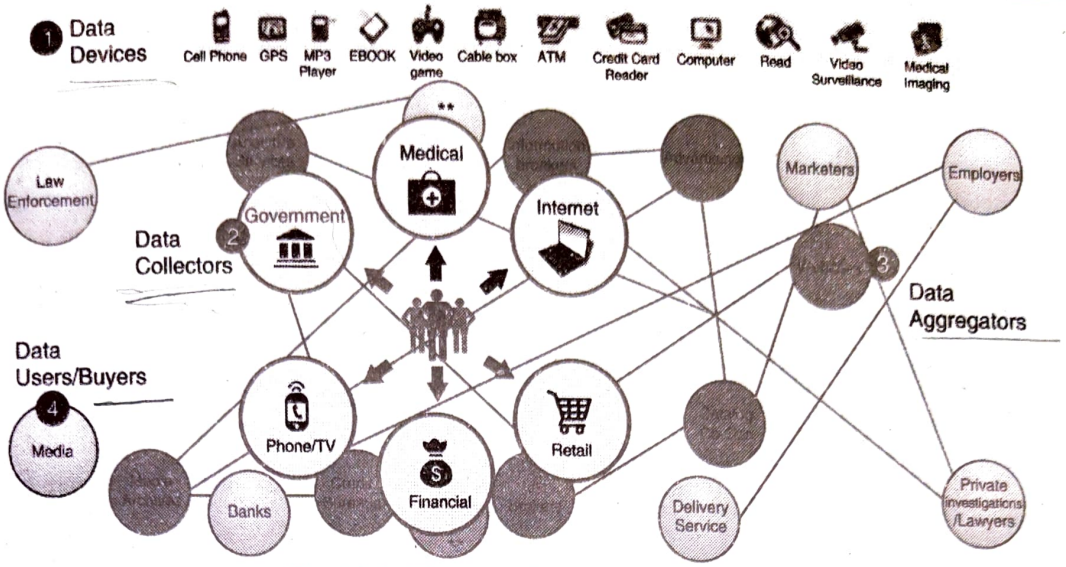

Fig. 1.2.4 : Emerging Big Data Ecosystem

- For example, consider a person playing an online video game with the help of desktop, game console, or smartphone. In such situation, the provider of game gather information regarding the skill as well as levels attained by the player. Intelligent systems helps to monitor and record the way and time the user plays the game. As a result, it is possible for the game provider to fine-tune the difficulty of the game, suggest other similar games which may most likely interest the user, and give offer of some extra equipment and enhancements for the game depending upon user's age, gender, and interests. This data may be stored locally or even uploaded to the cloud environment of game provider for the purpose of analyzing the gaming habits and opportunities for additional sell, and collect general profiles of particular kinds of users.
- Smart-phones are considered as one important rich source of data. Along with messaging and basic phone usage, they collect and transfer data regarding Internet usage, SMS usage, and real-time location. This kind of metadata is then used for the purpose of analyzing traffic patterns by the way of scanning the density regarding the smart-phones in various locations so as to track the speed of cars or the traffic congestion on busy roads. In this manner the GPS devices in vehicles provides real-time updates to drivers and also offer substitute routes to skip traffic jams.
- Retail shopping loyalty cards is another source to gather user data. These cards are used to record the spending amount of an individual as well as locations of stores the user visits, the types of products purchased, the stores in which frequent purchasing is done and the combinations of products which are mostly purchased together. Such information provides insights regarding shopping and travel habits of users which help the organization to promote their certain products.
- **Data collectors** (the gray ovals in section(2) in the Fig. 1.2.4) contains sample entities which gather data from the device and users.
- Data gathered from a cable TV provider helps to track the shows an individual likes, TV channels the person may and may not pay to watch on demand, and the amount which person may pay for premium TV content.
- Retail stores gather the information about the path which a customer prefer in the store at the time of using a shopping cart with an RFID chip consequently they can determine the products which get the most traffic.
- **Data aggregators** (the dark gray ovals in section(3) in the Fig. 1.2.4)are concerned with the data gathered from the several entities from the "SensorNet" or the IoT (Internet of Things). These organizations do the compilation of data from the devices and usage patterns

gathered by government agencies, retail stores, and websites. In return, they get benefit to transform and package the data just like product to sell to the list brokers, who needs to create marketing lists of people who may be considered as best targets for particular ad campaigns.

- **Data users and buyers** (the dark ovals in section(4) in the Fig. 1.2.4) get direct advantage from the data gathered and aggregated by others in the data value chain.
- In baking field, various banks which buy user data may like to know the information about the customers having good income to apply for a another mortgage or a home equity line of credit. To give input for such analysis, banks may purchase data from a data aggregator. This type of information may contain demographic data about people staying in particular locations; persons who seem to have a particular level of debt, yet still posses good credit scores (or some other good things like paying bills on time and having savings accounts) which can be referred to decide credit worthiness. Collecting information from such several sources and aggregators will facilitate a more targeted marketing campaign that would be more challenging before Big Data because of lack of data or high-performing technologies.
- Use of technologies like Hadoop to carry out natural language processing on data in the form of unstructured, textual from social media websites, enables users to determine the reaction to events like presidential campaigns.
- Sometimes people like to find out public sentiments toward a candidate by the way of analyzing associated blogs and online comments. Likewise, data users may need to track and be ready for natural disasters by the process of identification of areas a hurricane affects first and the way it moves, depending upon which geographic locations are tweeting regarding it or discussing it on social media.
- We can observe that the type of data and the related market dynamics vary to a great extent as per the emerging Big Data ecosystem,
- These data sets can contain sensor related data, text format data, structured datasets, and data from social media.
- It is confirmed that it is difficult for these data sets to work well within traditional EDWs, architecture of which is devoted to streamline reporting and dashboards and managed centrally.
- Hence Big Data problems and projects needs new approaches to succeed.

Syllabus Topic : Data Analytic Life Cycle - Overview

1.3 Data Analytic Life Cycle : Overview

Q. 1.3.1 Explain Data Analytical Life Cycle with all the six phases. (Refer section 1.3) **(8 Marks)**

- At this level we need to know more deep knowledge of specific roles and responsibilities of the data scientist.
- The data scientist lifecycle is illustrated in Fig. 1.3.1 which gives the high-level overview of the data scientist discovery and analysis process.
- It depicts the iterative behaviour of work performed by the data scientist's with several stages being repetitive in order to make sure that the data scientist is utilizing the "right" analytic model to locate the "right" insights.

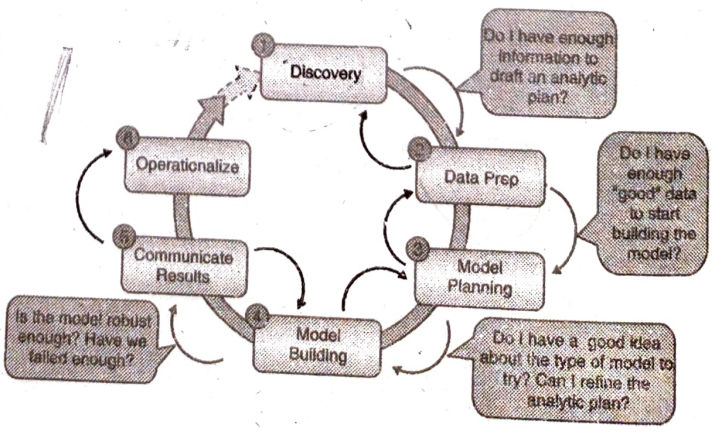

Fig. 1.3.1 : Data Scientist Lifecycle

Syllabus Topic : Phase 1 - Discovery Phase

1.3.1 Phase 1 - Discovery Phase

The following activities of data scientists can be focused by the Discovery :

- Acquisition of a complete understanding of the business process and the business domain. This consists of recognizing the key metrics and KPIs against which the business users will measure success.
- Recognizing the most vital business questions and business decisions that the business users are attempting to answer in support of the targeted business process. This also should contain the occurrence and optimal timeliness of those answers and decisions.
- Evaluating available resources and going through the process of framing the business problem as an analytic hypothesis. At this stage data scientist constructs the initial analytics development plan that will be used to direct and document the resulting analytic models and insights.
- It should be noticed that understanding into which production or operational environments the analytic insights requires to be published is somewhat that should be recognized in the analytics development plan.
- Such information will be essential as the data scientist recognizes in the plan where to "operationalize" the analytic insights and models.
- This is a best opportunity for tight association with the BI analyst who likely has already defined the metrics and processes required to support the business proposal.
- Requirements and the decision making environment of the business users can be well understand by the BI analyst to starts the data scientist's analytics development plan.

Syllabus Topic : Phase 2 - Data Preparation

1.3.2 Phase 2 - Data Preparation

The following activities of data scientists can be focused by the data preparation :

- Provisioning an analytic workspace, or an analytic sandbox, where the data scientist can work free of the constraints of a production data warehouse environment. Preferably, the analytic environment is set up such that the data scientist can self-provision as much data space and analytic horsepower as required and can fine-tune those requirements throughout the analysis process.
- Obtaining, cleaning, aligning, and examining the data. This contains use of data visualization techniques and tools to get an understanding of the data, recognizing outliers in the data and calculating the gaps in the data to decide the overall data quality; determine if the data is "good enough."
- Transforming and enhancing the data. The data scientist will look to use analytic techniques, such as logarithmic and wavelet transformations, to sort out the potential skewing in the data. The data scientist will also look to use data enhancement techniques to create new composite metrics such as frequency, recentness, and order. The data scientist will make use of standard tools like SQL and Java, as well as both commercial and open source extract, transform, load (ETL) tools to transform the data.
- After this stage is completed, the data scientist wants to feel comfortable enough with the quality and prosperity of the data to move ahead to the next stage of the analytics development process.

Syllabus Topic : Phase 3 - Model Planning

1.3.3 Phase 3 - Model Planning

The following activities of data scientists can be focused by the model planning :

- Determining the numerous analytical models, methods, techniques and workflows to discover as part of the analytic model development. The data scientists knows in advance that which of the analytic models and methods are suitable but it is good thing to plan to check at least one to make sure that the opportunity to build a more predictive model is not missed.

- Determine association and co-linearity between variables in order to select key variables to be used in the model development. The data scientist desires to estimate the cause-and-effect variables as early as possible. Keep in mind, association does not provides assurance causation, so care must be taken in choosing variables that can be calculated while going forward.

Syllabus Topic : Phase 4 - Model Building

1.3.4 Phase 4 - Model Building

The following activities of data scientists can be focused by the model building :

- Manipulating the data sets for testing, training, and production. Whatever new transformation techniques are developed can be tested to observe if the quality, reliability, and predictive capabilities of the data can be enhanced or not.
- Calculating the feasibility and reliability of data to use in the predictive models. Decision calls are depends on quality and reliability of the data to check; is the data "good enough" to be used in developing the analytic models.
- At the end, developing, testing, and filtering the analytic models is done. Testing is carrying out to notice which variables and analytic models deliver the maximum quality, most predictive and actionable analytic insights.
- The model building stage is highly iterative step where manipulation of the data, calculating the reliability of the data, and determining the quality and predictive powers of the analytic model will be modified number of times.
- In this stage the data scientist may be unsuccessful many times in testing different variables and modelling techniques before resolved into the "right" one.

Syllabus Topic : Phase 5 - Communicate Results

1.3.5 Phase 5 - Communicate Results

The following activities of data scientists can be focused by the communicate results :

- Determining the quality and reliability of the analytic model and statistical implication, ability of measuring and taking the action of the resulting analytic insights. The data scientist wants to make sure that the analytic process and model was successful and accomplishes the required analytic goal of the project.
- To communicate with the insights of analytic model, results and the suggestions requires the use of graphics and charts. It is significant that the business stakeholders such as business users, business analysts, and the BI analysts should realize and obtain the resulting analytic insights.
- The BI analysts are partner in this stage of the data science lifecycle. The BI analysts have the strong understanding of what to present to their business users and how to present it.

Syllabus Topic : Phase 6 - Operationalize

1.3.6 Phase 6 - Operationalize

The following activities of data scientists can be focused by the operationalize :

- Providing the final suggestions, reports, meetings, code, and technical documents.
- Optionally, running a pilot or analytic lab to validate the business case, and the financial return on investment (ROI) and the analytic lift.
- Carrying out the analytic models in the production and operational environments. This engross working with the application and production teams to decide how best to surface the analytic results and insights.
- Combining the analytic scores into management dashboards and operational reporting systems, like sales systems, procurement systems, and financial systems etc.

- The operationalization stage is another area where association between the data scientist and the BI analysts should be very useful.
- Numerous BI analysts have the experience of combining reports and dashboards into the operational systems, as well as establishing centers of excellence to spread analytic learning and skills across the organization.

Syllabus Topic : Case Study - GINA

1.4 Case Study - GINA : Global Innovation Network and Analysis

Q. 1.4.1 Write a short note on Case of GINA. (Refer section 1.4) **(8 Marks)**

- EMC's GINA (Global Innovation Network and Analytics) team is a group of senior technologists placed in centers of excellence (COEs) all over the world.
- The main goal of team is to connect employees all over the world to drive innovation, research as well as university partnerships.
- The basic consideration of GINA team was that its approach would offer an interface to share ideas globally and enhance sharing of knowledge between GINA members who are not at one place geographically.
- A data repository has been created to store both structured and unstructured data to achieve three important goals :
  1. Store formal as well as informal data.
  2. Keep track of research from technologists all over the world.
  3. To enhance the operations and strategy, extract data for patterns and insights.
- The case study of GINA illustrates an example of the way by which a team applied the Data Analytics Lifecycle for the purpose of analyzing innovation data at EMC.
- Innovation is generally considered as a hard concept to measure, and this team is going to use advanced analytical methods so as to identify key innovators within the company.

1.4.1 Phase 1 - Discovery

- In this phase, identification of data sources is started by the team.
- Even though GINA has technologists which are skilled in several different aspects of engineering, it had few data and ideas regarding what it needs to explore but do not have a formal team which could perform these analytics.
- They consults with various experts and decided to outsource the work to the volunteers within EMC.
- The list of roles is as follows on the working team which were fulfilled :
  - **User of Business, Sponsor of Project, Manager of Project :** Vice President
  - **Business Intelligence Analyst :** Representatives from IT Field
  - **DBA (Data Engineer and Database Administrator) :** Representatives from IT
  - **Data Scientist :** Distinguished Engineer who are able to develop social graphs.
- The approach of project sponsor is to influence social media and blogging for the purpose of accelerating the set of innovation as well as research data across the world and to inspire teams of data scientists who can work as "volunteer" globally.
- The data scientists should show passion about data, and the project sponsor should have ability to tap into this passion of greatly talented people to achieve challenging work in a creative way.
- The data regarding the project is divided into two important categories. The first category regards with the idea submissions of near about five years from EMC's internal innovation contests, called as the Innovation Roadmap or Innovation Showcase.

- The Innovation Roadmap is nothing but an organic innovation process in which ideas are submitted by employees globally which are then judged.
- For further incubation, rest out of these ideas are selected.
- Consequently the data is combination of structured data, like idea counts, submission dates, inventor names, and unstructured content, like the textual descriptions regarding the ideas themselves.
- The second category of data consists of encompassed minutes as well as notes which represents innovation and research activity globally
- Additionally it represents combination of structured and unstructured data. The structured data consists of attributes like dates, names as well as geographic locations.
- In the unstructured documents data is regarding "who, what, when, and where" which represents rich data regarding knowledge growth and transfer inside the company.
- There are 10 important IHs which are developed by GINA team :

1. **IH1 :** It is possible to map innovation activity in dissimilar geographic locations to corporate strategic directions.
2. **IH2 :** The delivery time of ideas minimizes by the transfer of global knowledge as part of the idea delivery process.
3. **IH3 :** Innovators participating in global knowledge are able to deliver ideas fast as compared to those who do not.
4. **IH4 :** It is possible to analyze and evaluate an idea submission for the likelihood of receiving funding.
5. **IH5 :** Knowledge invention and increase for a specific topic can be measured as well as compared across geographic locations.
6. **IH6 :** Research-specific boundary can be identified by the knowledge transfer activity spanners in different regions.
7. **IH7 :** It is possible to map strategic corporate themes to geographic locations.
8. **IH8 :** Continuous knowledge growth and transfer events minimize the time required to create a corporate asset from an idea.
9. **IH9 :** Lineage maps get revealed when corporate asset is not generated by the knowledge expansion and transfer.
10. **IH10 :** It is possible to classify and map emerging research topics to particular ideators, innovators, boundary spanners, and assets.

1.4.2 Phase 2 - Data Preparation

- A new analytics sandbox is set up by the team with its IT department for the purpose of storing and experimenting on the data.
- In the process of data exploration exercise, the data scientists and data engineers come to know that specific data require conditioning and normalization.
- Also they come to know that various missing datasets were difficult to testing some of the analytic hypotheses.
- As data is explored by the team, it promptly realized that without good quality data, it would not be able to carry out the subsequent steps in the lifecycle process.
- Consequently it was essential to conclude for project what level of data quality and cleanliness was necessary.
- In the case of the GINA, the team realizes that several of the names of the researchers and people who are communicating with the universities were misspelled or had spaces at leading and trailing side in the data-store.
- Such little problems must be addressed in this phase to enable better analysis as well as data aggregation in subsequent phases.

1.4.3 Phase 3 - Model Planning

- In the GINA project, for large amount of dataset, it looks viable to use social network analysis techniques to observe the networks regarding innovators.

- In other cases, it was hard to provide appropriate methods to test hypotheses because of the lack of data.
- In one case (IH9), a decision is made by the team to begin a longitudinal study to start tracking data points over time about people who are developing new intellectual property.
- This data collection support the team to test the next two ideas later :
- **IH8 :** Continuous knowledge growth and transfer events minimize the time required to create a corporate asset from an idea.
- **IH9 :** Lineage maps get revealed when corporate asset is not generated by the knowledge expansion and transfer.
- For the longitudinal study being proposed, there is need to team to establish goal criteria for the purpose of study.
- Particularly, it required to decide the end goal of a successful idea which had traversed the entire journey. The parameters regarding the scope of the study consist of the following considerations:
  - Identify the correct milestones for the purpose of accomplishing this goal.
  - Trace the way by which people shift ideas from each and every milestone towards the goal.
  - After this, trace ideas which unable to reach the goals, and trace others which are able to reach the goal. Compare the journeys of both types of ideas.
- Make comparison regarding the times and the outcomes with the help of a few different methods based on the way by which data is collected and assembled.

1.4.4 Phase 4 - Model Building

- In this phase several analytical methods are employed by GINA team.
- It contains the work by the data scientist through NLP (Natural Language Processing) techniques on the descriptions in textual format of the Innovation Roadmap ideas.
- Also social network analysis is conducted using R and RStudio, and then developed social graphs and visualizations of the network of communications regarding improvement through R's ggplot2 package.
- Examples of this work are shown in Fig. 1.4.1.

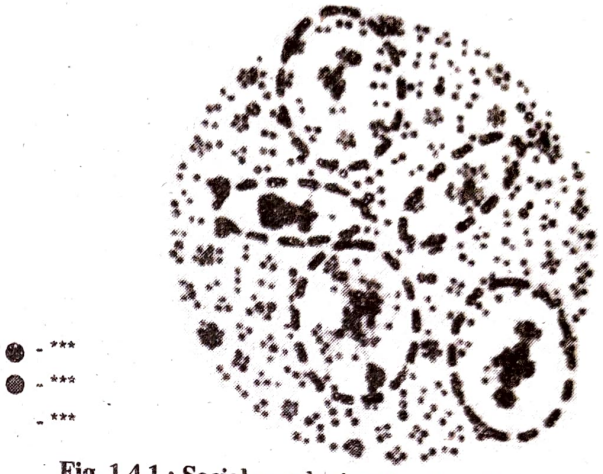

Fig. 1.4.1 : Social graph visualization of idea submitters and finalists

- Fig. 1.4.1 displays social graphs which depict the associations in between idea submitters inside GINA. Innovator from different countries are represented by dots. The large dots with circles around represent hubs.

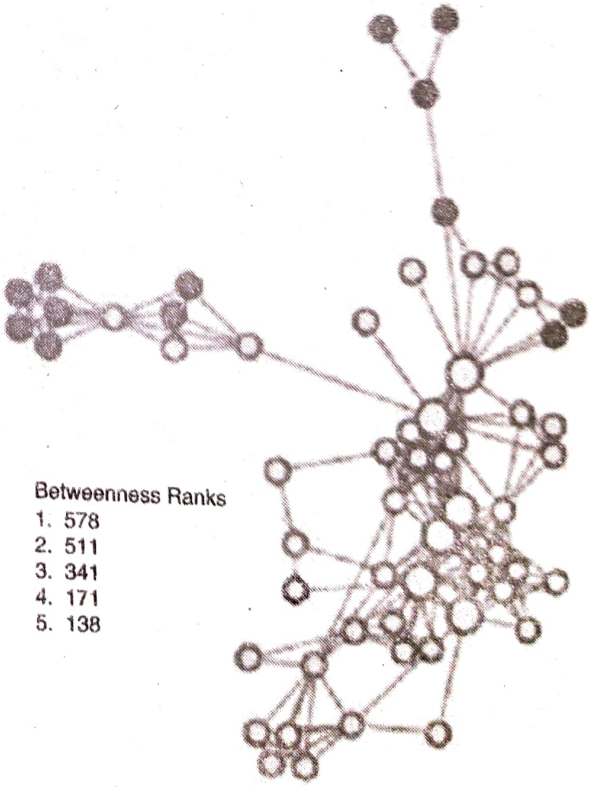

Fig. 1.4.2 : Social graph visualization of top innovation influencers

- A hub represents a person having great connectivity.
- The cluster in Fig. 1.4.2 consists of geographic variety, which is hard to show the hypothesis regarding geographic boundary spanners.
- In this graph, one person posses strangely high score when compared to the remaining nodes in the graph.
- This person is identified by the data scientists and they execute a query against his name within the analytic sandbox.

1.4.5 Phase 5 - Communicate Results

- In Communicate Results phase, the team got various methods to gather results of the analysis and identify the most effective and appropriate findings.
- This project seems to be doing well in the process of identifying boundary spanners and hidden innovators. Consequently, the CTO office establishes longitudinal studies to start data gathering efforts and keep track of innovation outputs for long duration of time.
- The GINA project inspires the concept of knowledge sharing regarding innovation and researchers located at various areas within and outside the company.
- One of the outputs of the project is that there was a strangely great density of innovators in Cork, Ireland.
- Every year, EMC hosts an innovation contest, which was open to all company employees to submit innovation ideas which can drive new value for the company.
- These findings were later on shared internally with the help of presentations and conferences and also promoted using social media and blogs.

1.4.6 Phase 6 - Operationalize

- Implementation of analytics against a sandbox which is basically filled with notes, minutes, and presentations from innovation activities results in high insights into EMC's innovation culture.
- Key findings from the project include :
  - The CTO office and GINA require extra information in the future, containing a marketing initiative for the purpose of convincing people to inform the global community on their innovation/research activities.
  - Some of the data is comparatively very sensitive, and hence the team requires considering security and privacy regarding the data like who can run the models and see the results.
  - In addition to running models, there is need of a simultaneous initiative to enhance the basic Business Intelligence activities like dashboards, reporting, and queries on research activities globally.
  - There is necessity of a mechanism to continually for the purpose of reevaluating the model after deployment. Assessing the benefits is an important goal of this stage, as is defining a process to retrain the model as needed.
- In addition to the actions and findings given in Table 1.4.1, the team also shows how analytics can drive new insights in projects which are basically traditionally hard to measure and quantify.
- Fig. 1.4.1 illustrates an analytics plan for the GINA case study example :

Table 1.4.1 : Analytic Plan from the EMC GINA Project

Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking the growth of global knowledge ensuring efficient knowledge transfer, and rapidly transforming it into corporate assets.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities

Components of Analytic Plan	GINA Case Study
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression Analysis
Result and Key Findings	A) Recognized hidden, high-value innovators and got methods to share their knowledge. B) Informed decisions regarding investment in university research projects.

Components of Analytic Plan	GINA Case Study
	C) Generated tools to help submitters for the purpose of improving ideas with idea recommender systems.