UNIT III

CHAPTER 3

Big Data Analytics Life Cycle

Syllabus Topics

Introduction to Big Data, sources of Big Data, Data Analytic Lifecycle: Introduction, Phase 1: Discovery, Phase 2: Data Preparation, Phase 3: Model Planning, Phase 4: Model Building, Phase 5: Communication results, Phase 6: Operation alize.

Case study : Global Innovation Social Network and Analysis (GINA).

3.1 DATA ANALYTIC LIFE CYCLE : OVERVIEW

UQ. Explain different phases of data analytics life cycle. **(SPPU - Q. 1(b), Aug. 18, 6 Marks)**

UQ. Explain Data Analytic Life cycle. **(SPPU - Q. 1(a), Dec. 18, Q. 2(b), Oct. 19, 8 Marks)**

UQ. Draw Data Analytics Lifecycle & give brief description about all phases. **(SPPU - Q. 1(b), May 19, 5 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. 3.1.1 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.

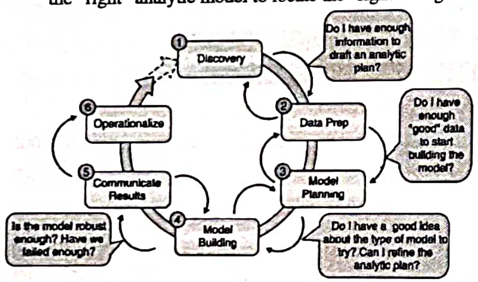

Fig. 3.1.1 : Data Scientist Lifecycle

3.1.1 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.

3.1.2 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.

3.1.3 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.

3.1.4 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.

3.1.5 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.

3.1.6 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

Unit III End Sem.

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.

3.2 CASE STUDY - GINA : GLOBAL INNOVATION NETWORK AND ANALYSIS

UQ. Write a case study on Global Innovation Network & Analysis (GINA).
(SPPU – Q. 2(a), May 19, Q. 2(b), Dec. 19, 5 Marks)
GQ. Write a short note on Case of GINA. **(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.

3.2.1 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 :
  (i) **User of Business, Sponsor of Project, Manager of Project :** Vice President
  (ii) **Business Intelligence Analyst :** Representatives from IT Field
  (iii) **DBA (Data Engineer and Database Administrator) :** Representatives from IT
  (iv) **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.

3.2.2 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.

3.2.3 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.

Unit III End Sem.

- This data collection support the team to test the next two ideas later :
  - (i) **IH8 :** Continuous knowledge growth and transfer events minimize the time required to create a corporate asset from an idea.
  - (ii) **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:
  - (i) Identify the correct milestones for the purpose of accomplishing this goal.
  - (ii) Trace the way by which people shift ideas from each and every milestone towards the goal.
  - (iii) 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.

3.2.4 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. 3.2.1.
- Fig. 3.2.1 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.

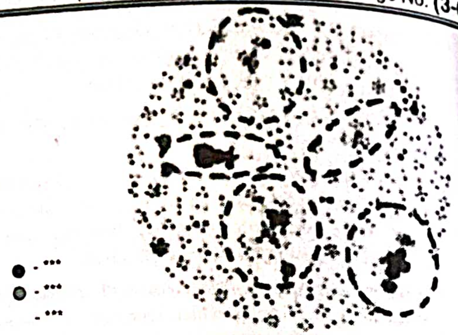

Fig. 3.2.1 : Social graph visualization of idea submitters and finalists

- A hub represents a person having great connectivity.
- The cluster in Fig. 3.2.2 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.

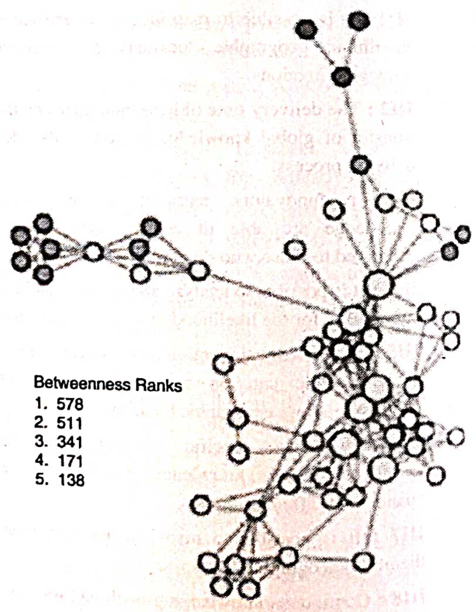

Fig. 3.2.2 : Social graph visualization of top innovation influencers

3.2.5 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.

3.2.6 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 :
  - (i) 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.
  - (ii) 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.
  - (iii) 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.
  - (iv) 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 3.2.1, the team also shows how analytics can drive new insights in projects which are basically traditionally hard to measure and quantify.
- Fig. 3.2.1 illustrates an analytics plan for the GINA case study example :

Table 3.2.1 : 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
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. C) Generated tools to help submitters for the purpose of improving ideas with idea recommender systems.

Chapter End...

□□□

Unit III End Sem