**MES Wadia College of Engineering Pune-01**

**Department of Computer Engineering**

|  |  |
| --- | --- |
| **Name of Student: Khushal Patil** | **Class: TE Comp – 2 (R Batch)** |
| **Semester/Year: 6Th / 3rd Year** | **Roll No: 64** |
| **Date of Performance:** | **Date of Submission:** |
| **Examined By:** | **Experiment No:**  Case-Study (C-1) |

**PART: C) ASSIGNMENT NO: 01**

**Title:** **Case Study**

Write a case study on Global Innovation Network and Analysis (GINA). Components of analytic plan are 1. Discovery business problem framed, 2. Data, 3. Model planning analytic technique and 4. Results and Key findings.

**Introduction**

The Global Innovation Network and Analysis (GINA) team was a strategic group of senior technologists strategically positioned in centers of excellence (COEs) across the globe. Their core mission was to foster a culture of innovation by actively engaging employees throughout these global COEs, with a focus on driving innovation, facilitating cutting-edge research, and establishing strong university partnerships.

This case study delves into how the GINA team effectively applied the Data Analytics Lifecycle to analyze innovation-related data within EMC (now part of Dell Technologies). It's important to understand that innovation is inherently complex and often difficult to quantify. The GINA team recognized this challenge and sought to leverage advanced analytical methodologies to identify key innovators within the company. Their approach was designed to facilitate seamless global knowledge sharing, meticulously track research progress, and comprehensively analyze innovation trends, drawing insights from both structured and unstructured data sources.

The study's primary objectives were multifaceted and aimed to:

* Establish a centralized repository for storing both formal and informal innovation-related data, ensuring that valuable information was captured and readily accessible.
* Systematically track the research contributions of technologists located in various global locations, providing a clear picture of individual and team efforts.
* Employ data mining techniques to uncover hidden patterns and extract actionable insights, ultimately enhancing strategic decision-making processes related to innovation investments and initiatives.
* Improve collaboration: Foster better collaboration among innovators.
* Identify emerging trends: Detect new areas of innovation.
* Measure the impact of innovation initiatives: Determine the effectiveness of programs.

**Background: The Importance of Innovation at EMC**

In the technology sector, companies like EMC recognize that sustained innovation is not merely beneficial; it's essential for survival and growth. The ability to rapidly develop new products, services, and solutions often determines market leadership and competitive advantage. EMC, operating in a dynamic and competitive landscape, understood that effectively harnessing the innovative potential of its global workforce was critical. This context underscores the significance of the GINA initiative. By analyzing innovation data, EMC aimed to:

* Gain a deeper understanding of its innovation processes.
* Identify and leverage its most valuable intellectual assets (its innovators and their ideas).
* Optimize resource allocation for research and development.
* Accelerate the time-to-market for new innovations.
* Maintain a competitive edge in the market.

The GINA team's work was therefore not an isolated project but a strategic effort to embed data-driven decision-making into EMC's innovation strategy.

**The Data Analytics Lifecycle and GINA**

The Data Analytics Lifecycle provides a structured framework for undertaking data science projects. The GINA team followed this process, which consists of six key phases:

* Phase 1: Discovery
* Phase 2: Data Preparation
* Phase 3: Model Planning
* Phase 4: Model Building
* Phase 5: Communicate Results
* Phase 6: Operationalization

**Phase 1: Discovery**

The Discovery phase was a critical starting point for the GINA project, setting the stage for a data-driven approach to innovation management at EMC. The team’s primary focus was to understand the business problem and identify relevant data sources that could provide meaningful insights.

**Framing the Business Problem**

At its core, the challenge was to gain a comprehensive view of innovation activities across EMC’s global organization. To break this down further, the team tackled several key questions:

1. How is knowledge growing within the organization?
2. How can knowledge transfer be made more efficient across teams and locations?
3. How can we accelerate the conversion of knowledge into valuable corporate assets?
4. What is the measurable impact of innovation initiatives?

Rather than relying on anecdotal evidence, the GINA team aimed to build a data-driven framework for tracking, measuring, and managing innovation. They sought to answer critical strategic questions, such as:

1. Where are the most promising ideas originating?
2. Who are the key individuals driving innovation?
3. How can collaboration be improved to maximize the impact of new ideas?

**Key Stakeholders and Their Roles**

A diverse set of stakeholders contributed to this phase, each bringing unique expertise:

1. Business User, Project Sponsor, and Project Manager – Led by a Vice President from the Office of the CTO, this group defined business objectives, secured resources, and ensured alignment with EMC’s overall innovation strategy. They provided the high-level vision and direction for the project.
2. BI Analyst (IT) – This role focused on leveraging existing data infrastructure, identifying potential data sources, and ensuring compliance with EMC’s data governance policies.
3. Data Engineer & Database Administrator (IT) – Responsible for technical execution, these experts handled data acquisition, storage, and management. They ensured data quality and set up the analytics sandbox for analysis.
4. Data Scientist – A distinguished engineer in this role was in charge of developing analytical methodologies, selecting appropriate techniques, and deriving meaningful insights from the data. They worked closely with business users to translate innovation challenges into analytical questions.

**Identifying Data Sources**

The team identified two primary data categories for the GINA project:

1. Innovation Roadmap – This structured database tracked formal innovation projects, proposals, and research initiatives. It contained key details such as project descriptions, timelines, milestones, team assignments, and funding allocations, providing a structured view of the innovation pipeline.
2. Minutes and Notes – This unstructured data source captured informal aspects of innovation, including:

Transcripts from brainstorming meetings and discussions.

Notes from research presentations and workshops.

Emails and instant messages between innovators.

Documents outlining preliminary research findings.

By combining structured data (Innovation Roadmap) with unstructured data (Minutes and Notes), the team aimed to develop a holistic understanding of EMC’s innovation landscape.

**Formulating Initial Hypotheses**

The team developed initial hypotheses to guide their analysis, categorized into two major areas:

**Descriptive Analytics** – Focused on understanding the current state of innovation at EMC. The goal was to analyze and visualize data to identify trends, patterns, and collaboration opportunities.

1. Where is innovation activity concentrated geographically?
2. What types of ideas are being generated?
3. How much collaboration exists between different teams?

**Predictive Analytics** – Aimed at forecasting future innovation trends to support strategic decision-making. Using statistical models and machine learning, the team sought to predict:

1. Which ideas had the highest potential for becoming valuable intellectual property?
2. Which research areas showed the greatest promise for future growth?
3. Which individuals were likely to generate high-impact innovations?

**A Key Hypothesis: The Role of Geographic Knowledge Transfer**

One specific hypothesis the team explored was:

“An increase in geographic knowledge transfer improves the speed of idea delivery.”

This suggested that when innovators from different locations share knowledge effectively, ideas are developed and implemented faster. Testing this hypothesis could help EMC optimize collaboration and innovation workflows across its global network.

**Phase 2: Data Preparation**

Once the data sources were identified, the next step was to prepare the data for analysis. This involved setting up the necessary infrastructure and ensuring that the data was clean, consistent, and of high quality.

* **Analytics Sandbox:** The IT department played a crucial role in setting up a new analytics sandbox. This was a dedicated environment where data scientists and data engineers could work with the data without affecting production systems. The sandbox provided the necessary computing resources, software tools, and storage capacity for data exploration, manipulation, and analysis. The purpose of the sandbox was to:
* Provide a safe and isolated environment for experimentation.
* Enable rapid prototyping and iteration.
* Protect sensitive data.
* Facilitate collaboration between data scientists and data engineers.
* **Data Conditioning and Normalization:** The data scientists and data engineers discovered that the raw data from the Innovation Roadmap and the textual notes required significant conditioning and normalization. This involved a series of steps to clean and transform the data into a usable format. Specific examples of data conditioning and normalization might include:
* **Handling missing values:** Determining how to deal with incomplete data (e.g., imputation, deletion).
* **Correcting errors:** Identifying and fixing inaccuracies in the data (e.g., typos, inconsistencies).
* **Standardizing formats:** Ensuring that data was consistently formatted (e.g., dates, units of measurement).
* **Removing noise:** Filtering out irrelevant or extraneous information.
* **Text processing:** Cleaning and preparing the textual data for NLP analysis (e.g., removing stop words, stemming, tokenization).
* **Data Quality:** The team recognized early on that data quality was paramount. They understood that if the data was not of sufficient quality, or if they could not obtain good quality data, it would be impossible to perform meaningful analysis in subsequent phases of the lifecycle. High-quality data is essential for:
* Ensuring the accuracy and reliability of the analytical results.
* Building trustworthy models.
* Making sound business decisions.
* Avoiding misleading conclusions.

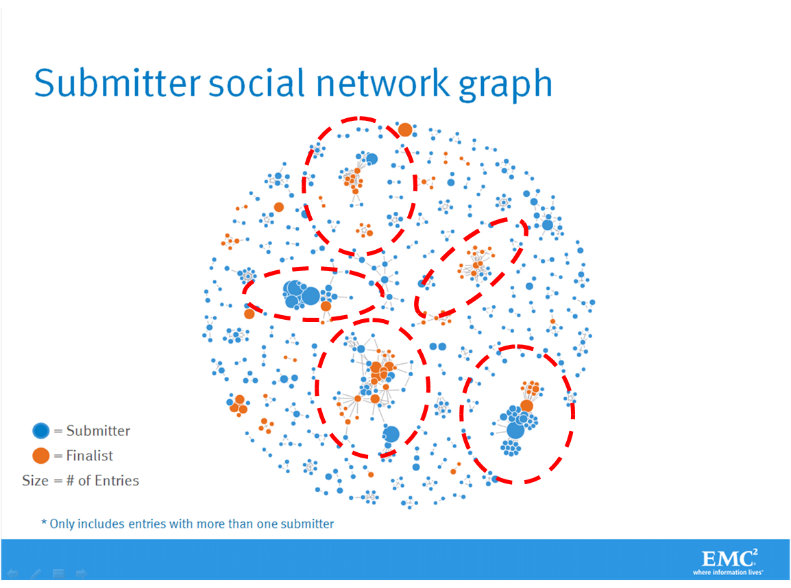
The team needed to define what level of data quality and cleanliness was "sufficient" for the project. This involved establishing data quality metrics, implementing data validation procedures, and working with data owners to improve data quality at the source.

**Phase 3: Model Planning**

* The team made a decision to initiate a longitudinal study to begin tracking data points over time regarding people developing new intellectual property.
* The parameters related to the scope of the study included the following considerations:
  1. Identify the right milestones to achieve this goal.
  2. Trace how people move ideas from each milestone toward the goal.
  3. Once this is done, trace ideas that die, and trace others that reach the goal. Compare the journeys of ideas that make it and those that do not.
  4. Compare the times and the outcomes using a few different methods (depending on how the data is collected and assembled). These could be as simple as t-tests or perhaps involve different types of classification algorithms.

**Phase 4: Model Building**

* The GINA team employed several analytical methods. This included work by the data scientist using Natural Language Processing (NLP) techniques on the textual descriptions of the Innovation Roadmap ideas.
* Social network analysis using R and RStudio.



* Fig shows social graphs that portray the relationships between idea submitters within GINA.
  1. Each color represents an innovator from a different country.
  2. The large dots with red circles around them represent hubs. A hub represents a person with high connectivity and a high "betweenness" score.
* The team used Tableau software for data visualization and exploration and used the Pivotal Greenplum database as the main data repository and analytics engine.

**Phase 5: Communicate Results**

* This project was considered successful in identifying boundary spanners and hidden innovators.
* The GINA project promoted knowledge sharing related to innovation and researchers spanning multiple areas within the company and outside of it.
* GINA also enabled EMC to cultivate additional intellectual property that led to additional research topics and provided opportunities to forge relationships with universities for joint academic research in the fields of Data Science and Big Data.

**Phase 6: Operationalization**

* Deployment was not really discussed.
* Key findings:
  1. Need more data in the future.
  2. Some data were sensitive.
  3. A parallel initiative needs to be created to improve basic BI activities.
  4. A mechanism is needed to continually reevaluate the model after deployment.

**Components of Analytic Plan - GINA Case Study**

| **Components of Analytic Plan** | **GINA Case Study** |
| --- | --- |
| Discovery Business Problem Framed | Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation. |
| Initial Hypotheses | An increase in geographic knowledge transfer improves the speed of idea delivery. |
| 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 | 1. Identified hidden, high-value innovators and found ways to share their knowledge.  2. Informed investment decisions in university research projects.  3. Created tools to help submitters improve ideas with idea recommender systems. |

**CONCLUSION:**

GINA (Global Innovation Network and Analysis) is a tool that provides insights and analysis to support innovation and technology development. It enables companies and organizations to stay up-to-date with the latest industry trends and technology advancements, allowing them to make informed decisions and stay competitive in their respective fields. The platform offers a range of features, including technology scouting, competitor analysis, and IP portfolio management, and can be customized to suit the needs of individual companies. Overall, GINA is a valuable resource for companies looking to innovate and stay ahead of the curve in their industries.