Computer Aided Decision Systems - Industrial Practices using Big Analytics

Professor Deepu Philip
Department of Industrial and Management Engineering
Indian Institute of Technology Kanpur
Professor Amandeep Singh
Imagineering Laboratory
Indian Institute of Technology Kanpur

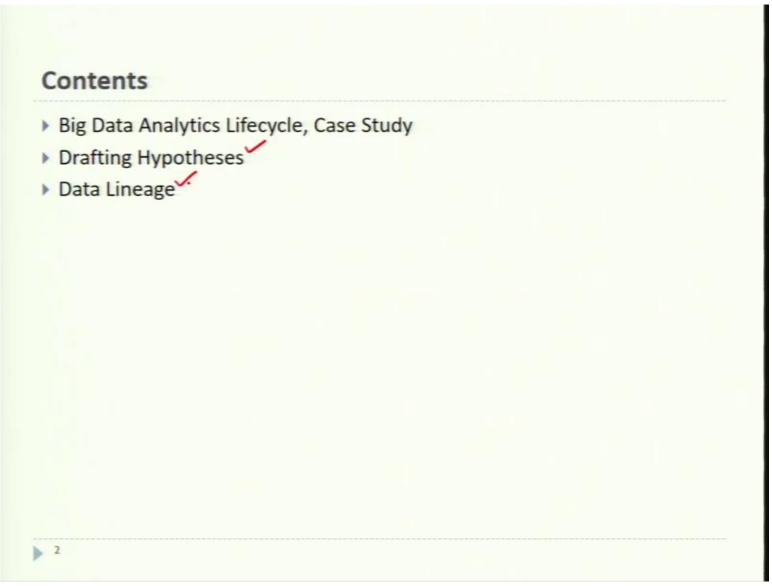
Lecture 27
Big Data Analytics Lifecycle - Case Study

(Refer Slide Time: 0:16)

Big Data Analytics Lifecycle
Case Study

Dr. Amandeep Singh

1



Contents

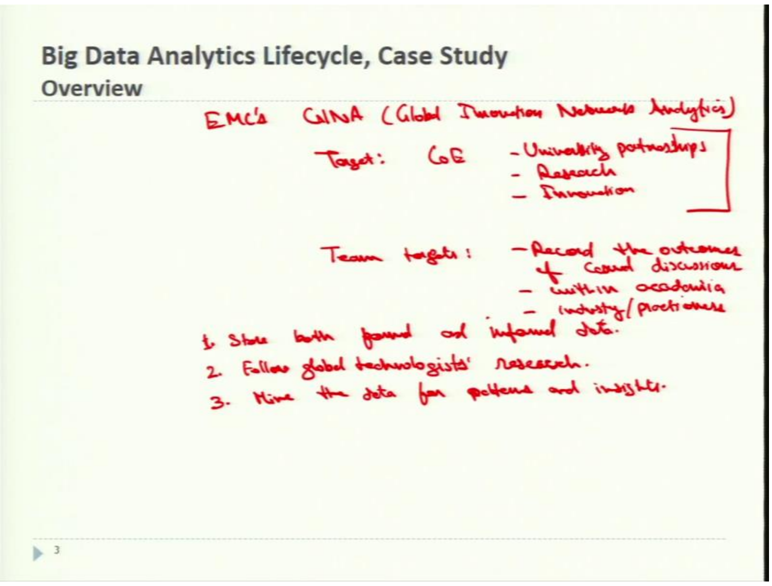
- Big Data Analytics Lifecycle, Case Study
- Drafting Hypotheses
- Data Lineage

2

Welcome to the next lecture on big data analytics life cycle. A case study would be discussed. The case study would be taken on a GINA, which is an EMC's Global Innovation Network Analytics system. So, after that while studying the case study, you will say how do we draft the hypothesis. We try to draft the hypothesis for the present status of the data kind of the present goals that we are trying to show.

Also in the hypothesis, we will have a few of them, which tries to identify whether the kind of the model that we are developing now would be helpful in the future use or not. So, this is also what we will be doing, while doing these data lineage is one of the concepts that I will discuss in some detail.

(Refer Slide Time: 0:57)



Big Data Analytics Lifecycle, Case Study

Overview

EMC's GINA (Global Innovation Network Analytics)

Target: CoE - University partnerships
- Research
- Innovation

Team targets: - Record the outcomes of casual discussions
- within academia
- industry / practitioners

1. Store both formal and informal data.
2. Follow global technologists' research.
3. Mine the data for patterns and insights.

3

So, let us start with the overview of the life cycle of that case study that you are trying to take. So, the case study is in the EMC's GINA (Global Innovation Network Analytics). So, what is this? It is a software framework that bridges the symbolic and connections representation of the world through executable conceptual models. So, GINA works with a mission, with a team, with the centers of excellence, in which the target was to involve the staff from different centers of excellence in advancing universities, partnerships research and innovation, so, they developed a small target that they wanted to develop more effective ways to record the outcomes of its casual discussions. So, with the thought leaders at different EMC or in Academia or in other organizations.

So, the target here was, they wanted the different center of excellences and the representative of center of excellences. They wanted to have a collaboration between

1) University Partnerships
2) Research
3) Innovation.

For this work, a team was developed like I discussed about the team including the business intelligence administrator and to the business or the data scientist. This team wanted to develop more effective ways to record.

- So, the team's target was to record the outcomes of the casual discussions.
- The outcomes within academia or with industry or practitioners, which means there are conferences, there are certain meetings, there are certain collaborations or certain symposiums which keep on bringing the people together to one platform or there might be individual contacts or single contacts between the people in academia or in industry or within an academia or within industry. Whether these informal discussions were also bringing some innovations or not they wanted to test this thing. So, like IIT Kanpur had a discussion with maybe University New South Wales, with the University of Toronto or so. Then they keep writing MOUs or sometimes it is very formal way when the director of both universities participates.

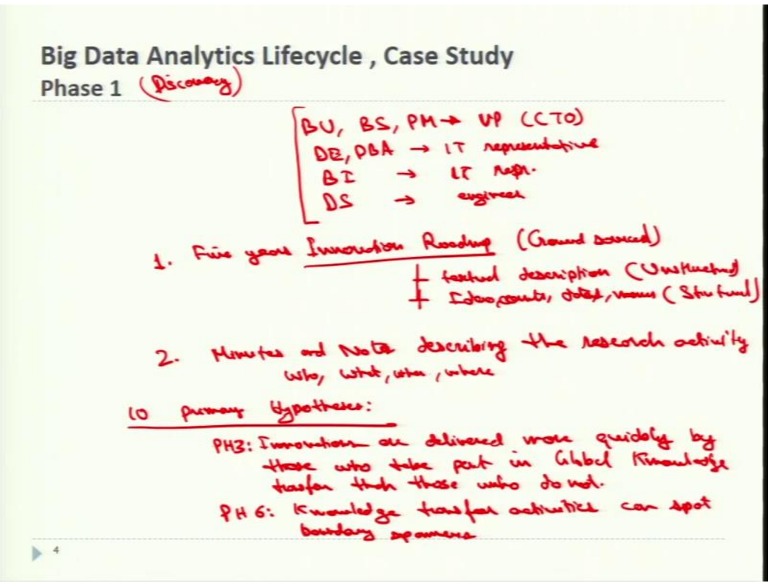
Another part a professor from IIT Kanpur tries to integrate or try to collaborate some of the work with a professor at University of Minnesota. So, this kind of integration also happens. This is all formal. Sometimes informal discussions, there is an idea that the person says okay, this is the fund available for this, they try to collaborate.

So, how does this happen and how do they come up with some innovations out of it and who are the people, who were the most working upon this, they wanted to develop a data repository that will include both structured and unstructured data.

1) So, in order to improve the team's operations and strategy, it was important to store both formal and informal data. Formal data is, what happens in the university, the research paper that you sent to university or the kinds of the abstract of the talks you send to the conference or so. The informal data would be the casual discussions that you have taken or the calls that you have taken with the people who are going to be present or who were present, whom do you follow up later in a different conference or when you meet or maybe the text messages or an informal email those are all unstructured data. Store both informal and formal data was one of the operations or strategies that was targeted upon.
2) Follow up or follow global technologists research that is whether the global technology is being discussed or being researched upon or not

3) One of the targets was also put as mine the data for any patrons and insights.So, a team used a big data analytics system and they wanted to find a way to use cutting-edge analytical techniques to identify key innovators within the company because innovation is typically a challenging concept which cannot be measured directly. So, they try to develop a few hypotheses on that and try to prove or try to work upon that hypothesis whether they are true or not true. They had some analysis around that. So, first phase of this was discovered.

(Refer Slide Time: 7:01)



Big Data Analytics Lifecycle , Case Study

Phase 1 (Discovery)

BU, BS, PM → VP (CTO)
DE, DBA → IT representative
BI → IT rep.
DS → engineer

1. Five years Innovation Roadmap (Crowd sourced)
+ textual description (Unstructured)
+ Idea counts, dates, names (Structured)

2. Minutes and Notes describing the research activity
who, what, when, where

10 primary Hypotheses:

PH3: Innovations are delivered more quickly by those who take part in Global Knowledge transfer than those who do not.

PH 6: Knowledge transfer activities can spot boundary spanners

4

In the 'Discovery' phase, the GINA team or the GINA project team started identifying the data sources. Data sources means for where the data would come, it lacked a formal team. So, they wanted to develop a new team in which they teamed up with a business user, data engineer and database administrator or the IT representatives, so business user and the business sponsor or the project sponsor or the project manager, they were all the vice president of the company or of not the company vice president, they were all a person that who was the vice president from the chief technical office.

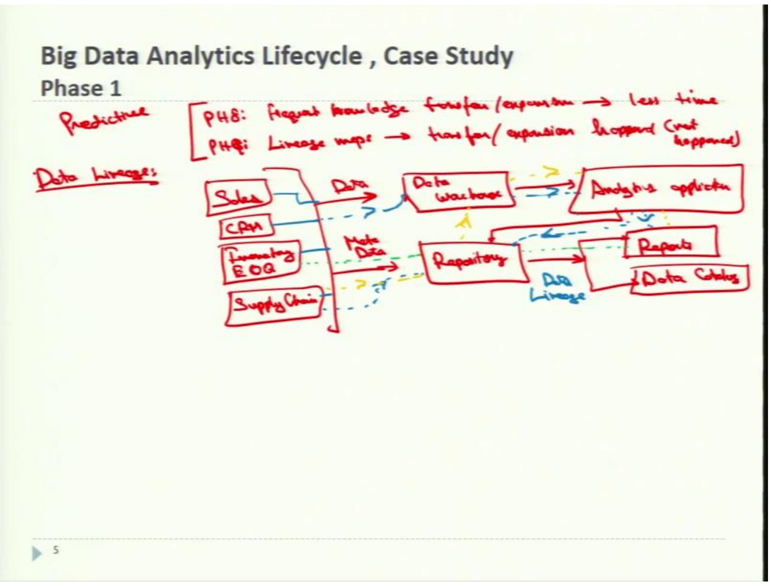
Then, data engineers and database administrators, they were the IT representatives. So, the first step was to make a formal team who would be given a target to work upon. These are the targets; we need to work upon. So, business intelligence analysis was also from IT representatives only, then data scientists. Data scientist was a distinguished engineer and that means he is well versed with the local languages and the social graphs which are displayed. So, he was just an engineer taken from the team who was thought to be intelligent enough to understand.

Now, the project sponsor strategy involved using the social media blogging to promote team for volunteer data. So, lacking a formal team he had to be creative in finding these individuals. The two main categories of the project data which were sort of was

1) The first category was the 5 years innovation roadmap. That is the internal innovation concept or idea submissions within the company implies from all over the world. This is kind of a crowdsource system, where the employees from all over the world of the company submitted ideas through a formal or organic innovation process which is known as innovation roadmap or this was nothing but a kind of a hackathon only. The best concepts were chosen for additional incubation. So, as a result the data became a combination of unstructured content that is a textual description. It is unstructured and the structured data that is the idea counts, then submission dates, names, this all becomes structured data.
2) The minutes and nodes describing the research activity. So, here attributes such as dates, names, places were part of the structure data once again and who was the person? What was he talking about? When was this happening? Where did it happen? these questions that we discussed were all there. So, this information represents the rich data but the knowledge about growth and transfer within the company was present in an unstructured document only. So, then they developed 10 primary hypotheses. I will just list a few of them. For instance, one of the hypotheses,

- ❖ Primary hypothesis 3 or initial hypothesis was innovations are delivered more quickly by those who take part in the global knowledge transfer than those who do not. So, which means people who are participating in the events which are more focusing upon the innovation enhancement. They were able to take or deliver the innovations more quickly. This was one of the hypotheses.
- ❖ Another hypothesis was the knowledge transfer activities can spot boundary spanners for a particular type of research in dispersed areas. For a particular type of research in the specific geographical area the knowledge transfer is happening or the prospect of funding can be examined and accessed for an idea submission, so these were all hypotheses which were developed.

(Refer Slide Time: 13:50)



Big Data Analytics Lifecycle , Case Study

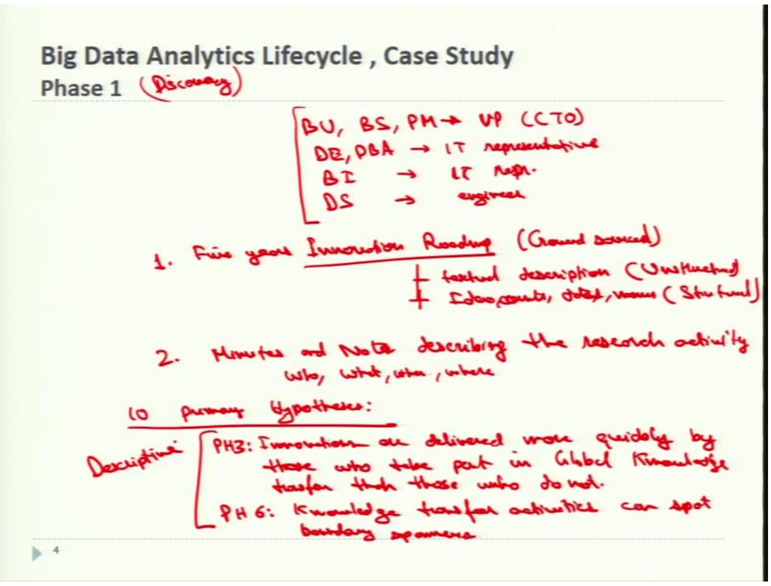
Phase 1

Predictive

PH8: frequent knowledge transfer/expansion → less time

PH9: Lineage maps → transfer/ expansion happened (not happened)

Data Lineages:



Big Data Analytics Lifecycle , Case Study

Phase 1 (Discovery)

BU, BS, PM → VP (CTO)
DE, DBA → IT representative
BI → IT repr.
DS → engineer

1. Five years Innovation Roadmap (Crowd sourced)
   + textual description (Unstructured)
   + Idea counts, dates, names (Structured)

2. Minutes and Notes describing the research activity
   who, what, when, where

10 primary Hypotheses:

Descriptive:
PH3: Innovations are delivered more quickly by those who take part in Global Knowledge transfer than those who do not.
PH 6: Knowledge transfer activities can spot boundary spanners

Two of the hypotheses that I would like to just jot down here that helps a team which they were able to test in the future that means the two theories were developed which could be tested in the future. One of these is that the time it takes to develop a corporate asset from an idea is shortened by frequent knowledge expansion and transfer activity.

- That is to say, if frequent knowledge transfer and expansion happens, this leads to less time to be invested in an idea.
- Another hypothesis is that lineage maps can show when transfer and expansion happen or not.

So, what is data lineage? Data lineage is a concept which talks about how the documents or the data that takes place through an organization information technology system flows between them and gets transformed from different uses along the way, that means it refers to the process of understanding and visualize the data flows from the source to the current location and tracking any alterations made in the data in its journey. For instance, let me say there is a data set or there is a system, in which the data comes from sales, or the customer relationship management of the company, or maybe the inventory we call it economic order quantity, or maybe different parts of the supply chain. This data and metadata, we get data and metadata from here which goes to the data warehouse or metadata goes to its repository. So, then we use it for the analytics application.

Then we try to also produce a repository from the analytics application that we try to use and then we try to have the reports and data catalog. Here comes the data lineage, where metadata is collected as data flows through different systems and then is used to generate data lineage information, which means the lineage tool gives you an instant visibility into the source and the journey of your data that from where is the data coming from whether it is coming through different files how do you extract the data whether it is coming through sales or CRM or inventory technology these all could be tracked here separately and how does it flow. Then who has taken the data, data in the data warehouse what were the maybe if new customer is there, the invoice, the orders, the products or regions the inverse line then transform sales.

So, if I try to maybe conduct or draw small dotted lines here or the networks here, okay, this is the network line here. So, this data was taken, it went here, then from here it went to the analysis, from the analysis or analytics application a report was directly made, sometimes. Or from it was also sent to a repository, if this direction is made. Maybe another set of the data from the economic order quantity, it only went through the repository. So, from the repository later reports were also made. It directly went through this.

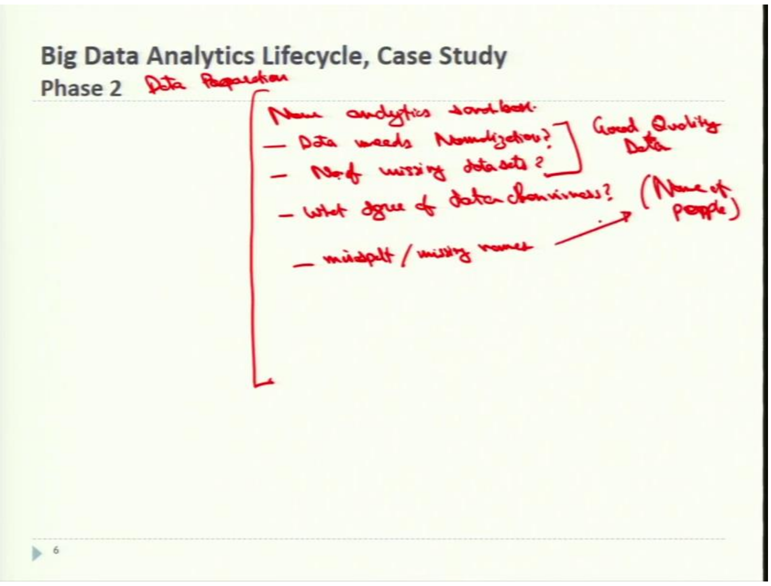
So, these dotted green, blue or yellow lines are representing the different flows of the data. This is data lineage. From where the data has come, how it flows through and where it went to, was it processed or not processed, was it cleaned or not cleaned, this is what data lineage helps us to understand. So, data lineages are nothing like as we do in the ISO applications, in the quality management systems.

When we try to develop medical devices, we try to track, from where has the material of the product come, who was the manufacturer of the material, is the person who has manufactured the material or the person who has supplied the material also ISO certified or not, the person who is working upon the operators is working on, is the person certified to work upon this machine or not. The material does it have a biocompatibility, or the final product that is produced does it have proper testing done upon them based upon the standards, or the flow when the operations when you try to assemble something or so, we try to see whether the assembly is it meeting the standards requirement, or the mechanical pressure requirements or not, this all things could be tracked. So, in data science this process is known as data lineage.

So, this was one of the hypotheses which were designed. That is the data lineage maps can show when the transfer and expansion of knowledge happened or did not happen and when it produced a corporate asset or not.

So, the group now had two kinds of analytics here. In the previous slide we majorly had, this was descriptive. Descriptive analysis which means what is happening right now to encourage more innovation or most teamwork. Then, these were for the future that means they were predictive, that is predictive means, to suggest to the exact management whether it should be making investments in the future or not. So, these analytics or these hypotheses were designed in phase 1.

(Refer Slide Time: 21:19)



Big Data Analytics Lifecycle, Case Study

Phase 2 Data Preparation

New analytics sandbox.

- Data needs Normalization?
- No. of missing datasets ?

Good Quality Data

- What degree of data cleanliness?
- misspelt / missing names

(Name of people)

6

In the phase 2 where the 'Data Preparation' is taken care of, in order to store the experiment with the data a team worked with its IT department to develop a new analytics sandbox which in turn means

- ❖ They tried to have a quick view whether the data needs normalization or not. This was the first question.

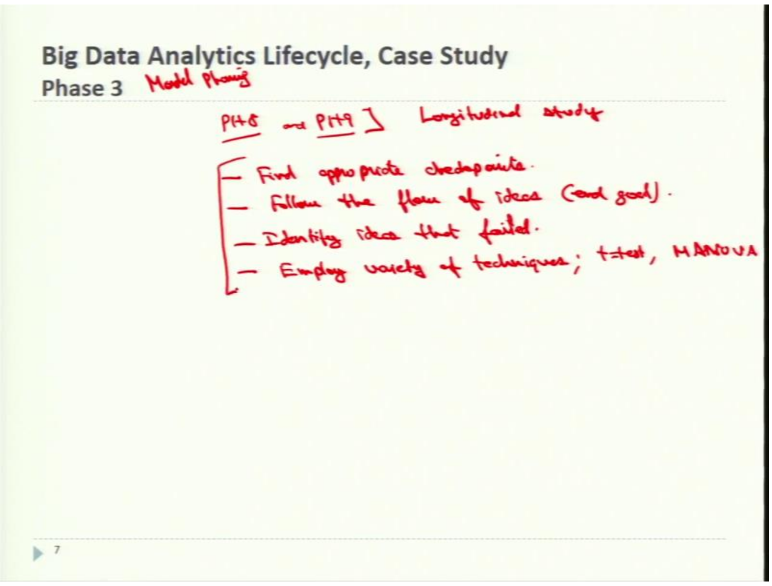
- Then, number of missing points or missing data sets and whether the number of missing data sets is acceptable or not. For instance, if I have sent a feedback email to a person or maybe a follow-up email to a person that I would like to participate in the collaboration that he discussed the last year or the last month whether the person has replied or not, if that reply is received but that is not yet recorded that data set is missing. It should be yes or no. It is received or not received. This data number of missing data sets, if those are missing sometimes whether to be filled with a dummy point or so. So, it became very clear to the team that it would not be able to complete the following steps of the lifecycle process, that is the phase 3 onwards unless they obtain good quality data. So, here is the data preparation itself.

As I said, data preparation is one of the phases where we spend the maximum of the time to make sure that the data that we have got is trying to represent the actual model or not actual model, means data is trying to or data is well enough or we have good enough data to finally have analytics conducted upon them. But the degree of data cleanliness is required. This was also a question.

This is specifically when the data is textual data. Sometimes we just say how, how is your family or so, these are formal informal discussions also happen. So, is this to be clean or is it also required and initial greeting is required or not, greeting at the different occasions is required or not, what level of the data cleanness is required. This was also a question and what quality was appropriate, this was also important.

So, minimum cleansing that is required that the people names or the researcher names were all that is required, that is the name of the people, that was minimal required. Sometimes the names of the trailing spaces were misspelled in the data store which were to also to be cleaned. So, this was the major point or major discussions which were there that is misspelled or missing names that was thought of as very important point because here they were working with the real-time data and they want to contact the people later, maybe to have some insights on the kinds of the innovation activity, they had been participating in the past. So, the name of the person and the right spelling of the person was thought to be an important point.

(Refer Slide Time: 24:54)



Big Data Analytics Lifecycle, Case Study

Phase 3 Model Planning

PH8 and PH9] Longitudinal study

- Find appropriate checkpoints.
- Follow the flow of ideas (end goal).
- Identify ideas that failed.
- Employ variety of techniques; t-test, MANOVA

7

Next comes phase 3, where the model planning happens. In the 'Model Planning' it appeared possible to use social network analyzes. These methods which help the GINA project to examine the networks of innovators within their company, that is EMC, then the dearth of data made it challenging to develop appropriate methods for testing hypotheses. The team decided to start with the longitudinal study for one of the cases. For example, for the lineage hypothesis they try to have a longitudinal study that means an extensive study. So, this was one of these things and for both the predictive analytics that they had drafted.

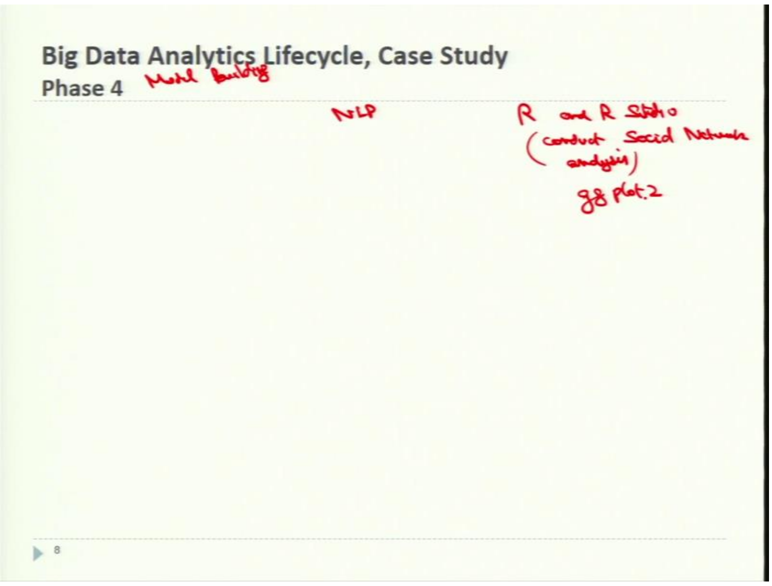
The team now had decided to have a longitudinal study. The group had to decide on goal criteria for a longitudinal study. So, for hypothesis, primary hypothesis 8 and primary hypothesis 9, it decided to have a longitudinal study.

So, since the group had already decided this. It was necessary to pinpoint the purpose of a winning concept, that had to make it all the way. The following factors were taken:

- ❖ Find appropriate checkpoints. That means, to reach its objectives, what are the different checkpoints, where should we stop, where are the milestones.
- ❖ Follow the flow of ideas. This was also one of the milestones that was set the floors, that is that each milestone what is the flow of idea, what is the end goal, that is what I said end goal, we are trying to target whether the data lineage or in the future, are we able to create, are we able to reduce the time of asset or not.
- ❖ Identify ideas that fail, to identify those that succeeded because we were going to talk about the future, whether investment is to be made or not. Identifying the ideas which failed was important to be considered as one of the criteria.

- ❖ Employing a variety of techniques was also thought of. Depending on the kind of data it is, they decided maybe to have techniques from a very simple T Test to a complex MANOVA. That is multiple analysis of ANOVA. When we have more than one objective function, this is in the model planning. Once the model was planned and these targets were put upon, 'Model Building' was the next phase.

(Refer Slide Time: 28:02)



Big Data Analytics Lifecycle, Case Study

Phase 4

Model Building

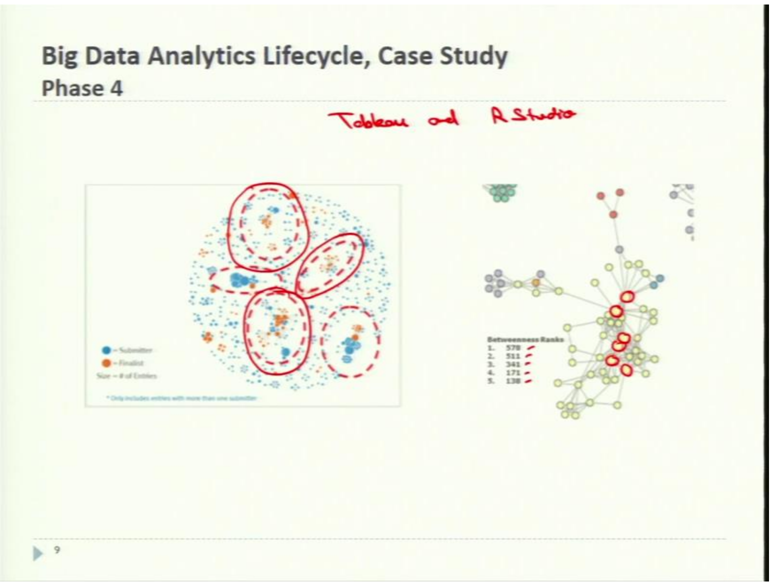
NLP

R and R Studio

(conduct Social Network analysis)

gg plot.2

8



Big Data Analytics Lifecycle, Case Study

Phase 4

Tableau and R Studio

9

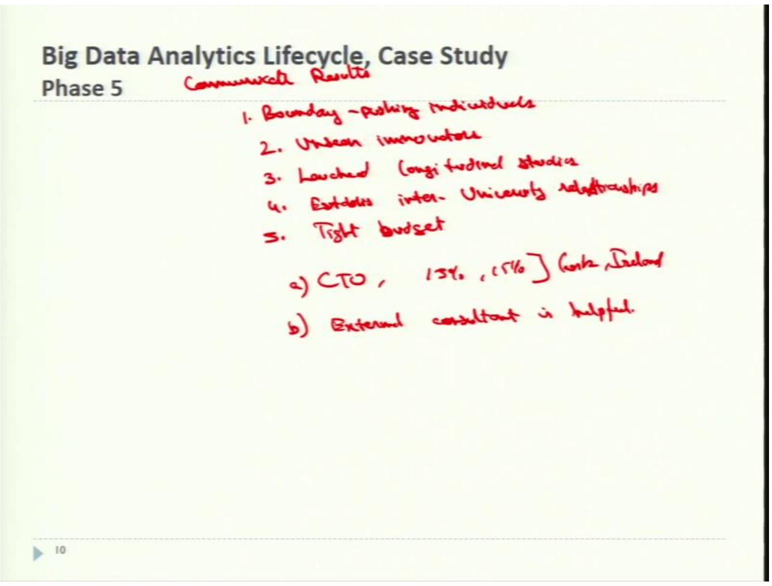
If we try to build the model. In the model building, the GINA teams used a variety of analytical techniques. In this, they included the data scientists works on the textual description majorly and National Language Processing (NLP) roadmaps or methods were used. They used R and RStudio to conduct social network analysis and there is a plot known as g g plot 2 which helps to create a social graph. So, the social graphs were developed. Then it was identified that in these graphs there was one person who was influential and he was thought of throughout the company's global operations, he was doing par apart.

So, this from the social graph, now the hubs are represented by large red circles exactly by dots, these are the hubs here, hub 1 hub 2 and so. So, the submitted and final list are given in different colors. The people who are submitted and those who are finalized were selected. The hub is an individual within a high between a score and high connectivity. The geographic diversity is present in the cluster here in the next figure. So, this is the geographic boundary or the geographic boundary spanners were there.

In second graph, one node stands out from the other nodes by having unusually high scores. So, here you can see between the ranks, how far the people are located but still the nodes stand here. These nodes you can see stood out. It was observed that or it was seen here that the actions revealed from this information, that the research scientist, who showed how influential he was throughout the company's global operations with a business asset, had certain behavior. That is, he went to a prestigious conference in 2011, which devoted to large scale data management, then he paid visit to workers in documentation business units, then he travelled to a conference in California in 2012, then he met a few employees from Russia, from Cairo, from Ireland, for United States, from India, then he also paid visits to research innovators different symposiums at different parts. So, the person, who had been participating in all of these parts, was able to develop more innovations in the company.

So, this result indicated that at least in part the original hypothesis is correct, that is the data can identify innovators, who cut across various business sectors and geographical regions. They used Tableau and RStudio to plot this data visualization and exploration. So, this was phase 4, where the model was built.

(Refer Slide Time: 31:27)



Big Data Analytics Lifecycle, Case Study

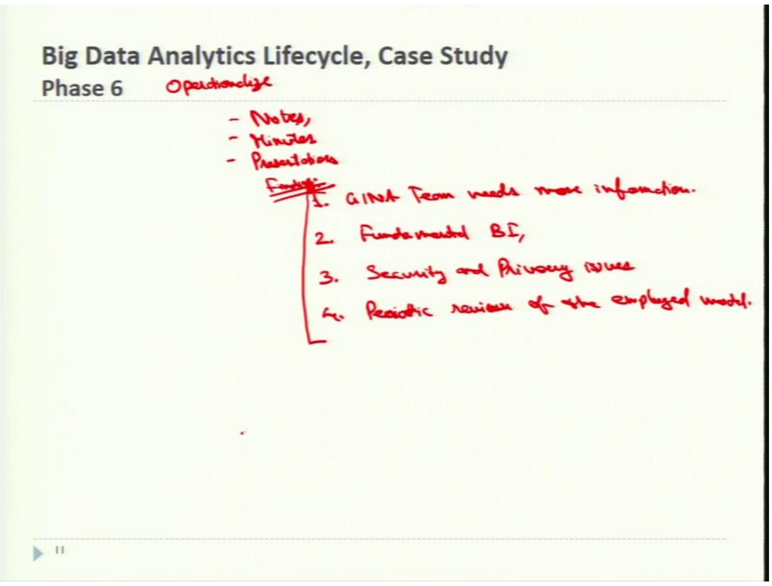
Phase 5 Communicate Results

1. Boundary-pushing individuals
2. Unseen innovators
3. Launched longitudinal studies
4. Establish inter-University relationships
5. Tight budget

a) CTO, 15%, 15%] Cork, Ireland

b) External consultant is helpful.

10



Big Data Analytics Lifecycle, Case Study

Phase 6 Operationalize

- Notes,
- Minutes
- Presentations

Findings
1. GINA Team needs more information.
2. Fundamental BI,
3. Security and Privacy issues
4. Periodic review of the employed model.

11

Now comes phase 5, where the communication of results was to be taken care of. When the results were to be communicated so it was identified that the project is able to identify the

1) Boundary pushing individuals
2) Unseen innovators
3) Launched longitudinal studies
4) It established relationships between universities and for collaborative research
5) The project was completely done on a very tight budget

So, these were the major communications of the results that they planned to

a) Present to the CTO of the company. So, the project's major findings were that for certain places such as Cork Ireland had an unusually high density of innovators. Every year EMC holds a contest for innovative ideas that would create a new value where the 50 percent of the finalists and the 50 percent of the winners were both from Cork Ireland.
b) Investigation revealed that external consultants had provided great input in training the staff for innovation. I would say, not to operationalize there were to be strategies which are to be made. When we say operationalize, operationalize means the project is now to be continued and to be run time and again.

So, in this,

❖ the notes,
❖ then minutes,
❖ then presentations from different innovation activities were made and there were key findings that the CTO office and the GINA will need more information.

1) If your findings were like this of the test, that GINA team needs more information as well as marketing campaign information, or marketing campaign also be connected to persuade people to share their innovation and research activities.
2) In addition to running the models a parallel initiate needs to be created to improve fundamental business intelligence activities are to be run such as maybe dashboards, reporting, queries on research activities globally. So, some of the data is sensitive so the people should be very sure that their dense data whatever they are providing is safe and is classified.
3) Whichever the team needs to take security and privacy related issues to be taken care of.
4) So, after the model is deployed a system must be in place to periodically review the model, periodic review of the employed model. So, these were the insights which the team showed that how analytics can develop new information or new knowledge to the projects that are typically challenging to measure and quantify. In addition to the actions and findings mentioned, they also developed different inputs for the different team members. For instance, for the business intelligence officer, for the CTO or for the database administrator what were the things required, the analytical strategy for GINA case study.

Despite the fact that there are only 4 findings which are listed here, there were many more which they came up with so every business wants to promote his innovation. How do we work upon that Innovation, this was a case study taken by the GINA?

So, what activities, what specific tasks a person should do, how the person should participate, or where should the person participate or whom should it contact, or whom should it keep a network with, should it have an external consultant in the company or so, these were all identified and the hypothesis which were there were tested. Two of the hypotheses only I have discussed which were for Predictive Analytics. So, those were taken, data lineage is also one of the concepts that we have taken here.

Let us meet in the next week where we will try to discuss the Internet of Things, Industry 4.0, then Big Data Analytics in manufacturing. We will take a case study. Thank you.