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 18
Big Data Analytics Team

Welcome back to the course on Computer Aided Decision Support Systems and application using Big Data Analytics. Professor Deepu Philip has given a broad introduction about the course about Big Data Analytics, what it is and generally you have also studied some of the topics like Normalization, the SQL commands and some Database Management Systems, I will have to focus more on what is Big Data Analytics? The lifecycle of the Big Data Analytics systems, the team and how it differs from the normal data science or data analytics tools.

(Refer Slide Time: 00:55)

So, this week would focus more on the Big Data Analytics Lifecycle. The tiles broadly I will cover in this week would be:

- Information Technology vs Big Data
- Big Data Analytics team, who are the stakeholders, what are their roles and responsibilities, what should be the size of the team depending upon the organization depending on the size of the data or the kind of the future prospective outputs you need to have.
- In the Big Data Analytics Lifecycle, we keep on going in the forthcoming lectures, we will discuss different phases of it and how the phases are the planning is imperative is crucial is vital, before starting any project, this will be discussed this week itself.

(Refer Slide Time: 01:42)

Now, Information Technology versus Big Data. I will have to differentiate data science or data analytics from Big Data. Let me divide this into three parts.

i) Big Data

ii) Data science

iii) Data Analytics

When we say the word ‘Analytics’, it is the general term, which can broadly or visually consider three major types:

a) Descriptive Analytics- Descriptive means how do we describe the data mean, mode, median, variance, the measures of central tendency, measures of the dispersion. All these are describing the data, what data do we have, we have primary secondary sources of the data. The secondary sources are one which is available in some databases that we learned in the previous lectures, how to retrieve them, how to delete some of the points, how to recover, how to return to the positions. Primary data is the one that we create for our own purpose. So, these kinds of data are there in the Descriptive one.

b) Predictive Analytics- Predictive is what is the forecast in the future, what do you think would come based upon that previous data mapping on that current scenario, what will we get in the future? It is predictive.

c) Prescriptive Analytics- Prescription is when you have causality, when you have something

like regression models, like some heuristics to be written then we prescribe back away this will be the future. In this, these steps would lead to these kinds of output. These kinds of algorithms are written.

Big Data is the one, where we have a large amount of data analytics and analytics and data science are part of it. And, Big Data is an extension to the existing analytics tools.

- Data science is a field that refers to collective processes. It talks about theories, the concepts, the tools and technologies that enable us to analyze, review, and extract valuable knowledge or information from the raw data. So, we get further knowledge from the raw data. So, whenever you come to the different positions or the roles, then the analytics team, which is there, will talk about data scientists and data engineers. The basic role or the basic purpose of the basic skills, the technical skills are to be there with data scientists who understand what is the raw data, who understand what is programming, who understand all the commands that you are trying to learn in this course, that is a starting position of data science.

A data engineer is the one who does Data Analytics. He broadly understands the things, little more than data scientists only because data scientists majorly know the subject matter, expertise, analytical solutions or techniques or so. The data engineers should understand this as SQL, data management that is broadly the things that it does.

- Data analytics is the process of examining the data sets. So, in order to draw some inferences about the information, what kind of input or information does this data have? So, they have some specialized systems, which help to get the output or the conclusions out of the existing data sets.
- Big Data refers to the voluminous data. Voluminous data, that is we talked about a few points here in that previous lecture velocity, then volume, then variety is for the structure data. And for the unstructured data as well sometimes the things are not in a very streamlined way, but still the data is available in a volatile manner (volatility) and variances (variety of data) is available. So, when this kind of data is available, it is an

extension to the next step when you go to Big Data. And, that is all different.

So, a few examples of simple data science would be digital advertisements, the internet search that you actually do, the recommender system in the internet that is there. The next movie or the next program, next purchase that you should do, that could be helpful with the data size itself. If you are a small startup, you need to understand, okay, what kind of products are available in the market. What kind of similar product?

For example, if you need to sell a medical device that you have developed. I will take an example of the oxygen concentrator development in the forthcoming weeks as an example in the Big Data to risk management systems. So, how do we manage risk while developing a medical device? For example, if you are developing an oxygen concentrator, who would use it? In what scenario could it be used? These could be taken from the existing data points. So, this could be simply done by a simple internet search, so that we could come into data science. However, the detailed study could go into Big Data. So basically, the data science person should understand the basic technical languages such as Python, such as they should understand the platform such as SAAS, SPSS, the basic commands or SQL. I would also put SPSS with data analytics.

So, data analytics or the data engineers could understand SPSS. If I talk about open-source software like R or some tools like Tableau could also be put in a skill set, then we have Spark. Here comes the most used software, or the most popular is 'Hadoop', which the big pioneer, bigger tycoons have used like Netflix, like Walmart. These big tycoons have been using Hadoop because they started the data science, it started the data analytics and the Big Data Analytics in its incipient stage itself.

We understood the volume and the potential that Big Data could bring to the market to their sales to their overall output. So, they started using these things, Hadoop was the one of the bases or the one of the initial software which were used. Now NoSQL databases are available. Similarly, multiple databases are also available. These are the major differences between Data Sciences, Analytics and Big Data Analytics.

(Refer Slide Time: 09:24)

What does the platform of Big Data look like? If I talk about the problem-solution system, you will see a traditional or just a data Analytics systems generally what has is,

- A processor, in which the input data comes, let me say this is database 1, we have database 2, this database could be an Oracle based data, or maybe it could be a NoSQL data.
- It goes as input in the system, one has to clean the data first, data cleansing.
- Then we do Predictive Analytics.
- Then once the Predictive Analytics is taken, it goes to the, or it creates a new database which can be presented to the project sponsor or the upper hierarchy level that is we use some visualization tools. With relation, I will discuss the data visualization, what kinds of tables, what kinds of graphics to be used to present the data? There are different ways to choose, when to choose a pie chart, when to choose a line diagram, when to choose a histogram or maybe some more detailed charts such as, we have box notes or so when and where to choose them this also we will discuss. Because data once it is analyzed and we have the output ready with us as a scientist as an engineer, you will be able to understand what output you are trying to present, but the team, the expert or the financers would like to understand that in a visual form.

To make the upper level of hierarchy, understand what data or what output are you trying to

present to them, visualization is vital, what kinds of charts you represent that we will see. ❖ So, after visualization, you have here a report that is the output. This is a scenario for the general data analytics system.

❖ Now comes, your Big Data Analytics systems, in this case also, we would have a few databases, database 1, database 2, again, it could be Oracle, NoSQL Cassandra or anything like that, this goes inside the processing system. Now, here I will put the word Extract Transform and Load (ETL). I will talk about Extract Transform Load Transform (ETLT) also. What are the differences that we will discuss in the forthcoming lecture this week itself? Now, database 1 and database 2 goes as an input to the processing system, it does not go to the data cleansing directly as was in the case one above. So, this is case two, case two actually, the process two which is for the Big Data Analytics system, in which this input goes here to the processing system. Along with this we have archived data.

❖ Archive data is a previous data which is generally thought of no use in the general analytics system, because in the general analytics database system, we only opt to have very small wrenches of the data, a very few lines of the input or very small amount, which we think are only significant would only be used, but data which is not recovered, which is archive, which is previous, maybe only the recent weeks are taken, recent times are really taken in the general data analytics system.

Yes, recent times, recent weeks data, let me take in the case of supposed retail in the forecast or so is important for the data analytics in the Big Data Analytics as well. But the previous data, the previous year's data, previous month's data, a few decades back, is also taken to understand what are the trends, how the season changes, how the things would impact the forthcoming sales as well. So that is why archive data is also taken as an input in case two. So, this is input here, this input goes to the processing system, so my processing system is this large box.

❖ So here, ETL or ETLT systems help us to have different Predictive Analytics. Then similarly, I do have similar Predictive Analytics and Predictive Analytics on the next one. This goes to the Big Database and this is an enclosure in our overall Big Data Analytics system. This is my database again. These inputs go here. Finally, we have data

visualization

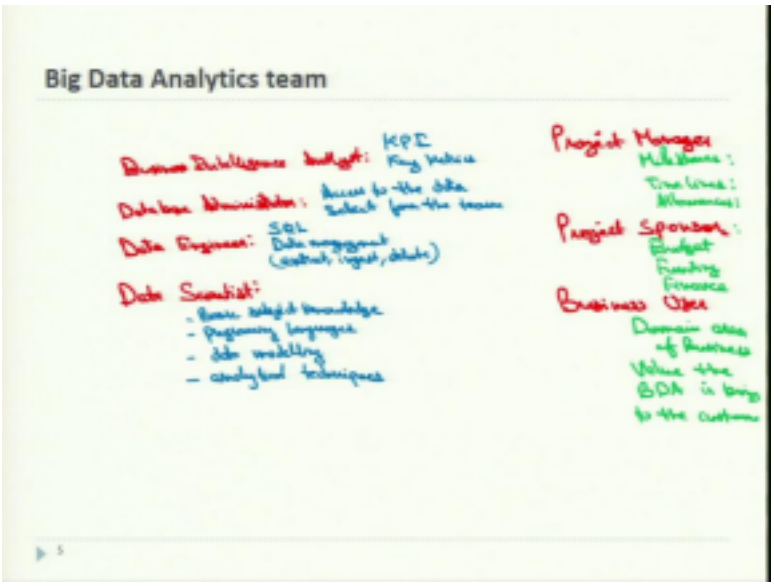
and a report. So, this is my Big Data Analytics system.

So, broadly if I say, in the Big Data Analytics system the archive or the historical data is also taken this is one. Second is we have not cleaned all the data, we just extract, transform and load the data as and when required we keep on taking points and the whole process is divided into small work breakdown structures. So, this small data that is suppose, if I designed a small algorithm in which out of these, let me say 20 petabytes of data is there, I only selected 20 gigabytes of data, made a small program, this program will be repeated time and again, once this program is tested, validated, so we have to define training, testing and validation data, how do we do that? we will come to the points. So, it is repeated multiple times to process the whole available 20 petabytes of data which is a large and big amount that I am talking about. So, this is how Big Data Analytics work.

So, what are these? Let us talk about the life cycle. Before that let us talk about the roles and responsibilities of the Big Data Analytics team because it is Big Data, we are talking about the large or volume of data with the velocity could also be high, in which variety could also be high, the team vary from the number of people from 3 to 20 in general.

It depends on the big companies like Walmart, which has more than two million employees and can have a team of 100s or 1000s of people, maybe 100s of people who are specifically working in the Big Data Analytics part of the company. So, the team's basic roles would be similar.

(Refer Slide Time: 17:54)



So, I have now delineated roles into different ways here. For instance, first role could be 1) Data Scientist- As I talked about this in the previous slide, as well data Scientist is the one who provides us subject matter expertise.

2) data engineers- Should be able to understand the SQL queries or data management, data extraction then support for data ingestion into the Analytics sandbox, so these things the person should know.

3) Database administrator- database administrator is one who figures out the database environment to support the Analytics needs of the working team.

4) Business Intelligence Analyst- We will talk about this further.

So now, let us talk about data scientists. To jot their roles that we put here are, data scientists have

- ❖ basic subject knowledge
- ❖ programming languages
- ❖ data modelling
- ❖ analytical techniques.

Now, we move towards data engineersing. A data engineers should be able to

use SQL or similar languages he should know for data management.

Management is a very wide term here, by management I say,

- ❖ extract, ingest, delete, etcetera, the person should know. So that he keeps on guiding the data Scientist, what are the things the person needs to do.

Just these two roles are interchangeable in a way in a small team of data engineerss and Data Scientist both the jobs could not be done by a similar person or the same person.

Now, database administrator is the one, who has more broad information his database administrators who might have two or three data engineerss working under him and he should be able to take the responsibility for providing the

- ❖ access to the data for the data engineerss or scientists.
- ❖ Then, key tables or key databases which are required, he should be able to select them as well for the data team.

Now comes the Business Intelligence Analysts. Business Intelligence Analysts is the one, who understands what is the business, what are the key metrics. So now, which is something known as

- ❖ KPI (Key Performance Indicators). KPIs are defined, KPIs are measured, KPIs are quantified, for instance, applying this algorithm the sales should rise by 100 pieces per day. This is a Key Performance Indicator. Applying this technique, the employee retention in our company should be increased by 20 percent. This is a Key Performance specifically 20 percent 100 pieces per day, these performance indicators are only defined by Business Intelligence Analysts. So Key Performance Indicators are defined by him.
- ❖ Also, they put key metrics. These metrics are pulled from the business intelligence viewpoint, that is how to create different dashboards and reports and have knowledge about the different data feeds and resources.

Now, understanding the data visualization sometimes data scientists might miss those points. data visualization what to present, what kind of graph to present, one figure could give us an information that 1000 words could not, where to put the different data points, what are the different metrics, how do we label them, Business Intelligence Analysts should understand because he has read multiple reports multiple books, and he should be able to understand and try

to present this to other roles. So, these are all supported by the person who is trying to manage everything.

Project manager might not be a core expert in Big Data Analytics. But he is overall managing the project, he has seen what are the objectives and milestones, these are all to be met in time, this is his role. We cannot run any project without a sponsor. Project sponsor provides the finances, always includes the people who are going to use the data we provide and also the business user. For instance, in a manufacturing concern itself, I had a chance to work with Hero Cycles. Whatever bicycles were produced, first, they were given only to the people who are within the company itself. Maybe the milk vendors or the small vendors who used to deliver newspapers with the milk within the hero cycles plant themselves at Ludhiana. So, these internet users give us the input or the feedback, what is the overall performance of the bicycle? If we try to use it in a very bashing way? How were the breaks or not?

So, these kinds of business users are identified, sometimes their internal users, sometimes they are the people who are identified as users only, but they are paid for it. For example, Netflix pay people to watch movies and tack some points wherever they people pause, people stop watching auditing. So, this is something that is very important for the overall business as well.

- So, the key rules of Project manager's milestones are set. The Key Indicators are the Key Performance Indicators given by the Business Intelligence Analysts. But milestones are met in time, timelines, then alliances. These are all given by the project manager alliances, meaning if your timeline is one week, it could be one week plus minus one day, something like that.
- Project sponsor understands what the total budget is. The budget project manager also knows, but the financing is to be done by the project sponsor, from where the funding is coming, or who is going to finance it. Project Sponsor actually understands what the market is, what does the market need? So, he is investing in something, he does not know anything or maybe he knows very little about what is Big Data or what is analytics or so. The person who is putting their money to the project sponsor knows that, by the past experiences or their example company who has used this kind of the procedure, these

kinds of algorithms to come up with a profit of this level O. So, that profit plus minus 5 percent or so, he would only expect to have. So, he understands what market there is. Big Data, specifically technical information the person might not have, but as an executive summary, he knows everything that is going on in the system.

- Now, next is a Business user. Business user is the one who understands the domain area of business. So, this is a person, who can consult or who can advise the project team on the overall context of the project or what value the Big Data Analytics (BDA) is bringing to the customer. So, these are the key roles or Key Analytics team members, seven roles are there majorly that I bought. The positions or the name of roles could be different, it could be project based or product sponsor, it could be the financer, it could be the owner, it could be the managing director or so. So, business users could be the final chooser or maybe in the gaming business. The person who plays the game for the company and tries to give the feedback on the new inputs that could come. So different kinds of roles, different positions are there, but majorly these seven rules are there. In a team of three a few roles at the lab. For example, data engineers, data Scientist could be one Business Intelligence Analyst and project manager could be one. Sponsor and project manager could also be one.

In a team bigger, we can have multiple engineers or scientists, maybe 20 or 30 scientists working under a single product, engineer or I would say data engineers, these roles are there in general. So, with this, I would like to have a pause in this lecture. And I will continue in the second part of this week, in which I will try to explain the Big Data Analytics Lifecycle. Thank you.