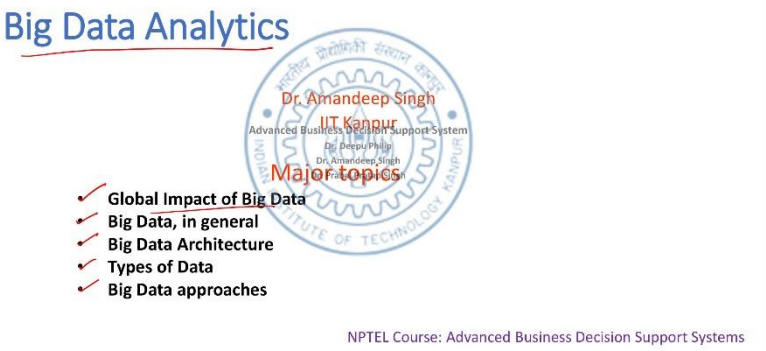
Advanced Business Decision Support Systems

Professor Deepu Philip
Department of Industrial Engineering and Management Engineering
Indian Institute of Technology, Kanpur
Professor Amandeep Singh
Imagineering Laboratory
Dr. Prabal Pratap Singh
Indian Institute of Technology, Kanpur

Lecture 40

Big Data Analytics

Welcome to the last week of the course Advanced Business Decision Support Systems. You have learned a lot of things about the different kinds of the models for the Decision Support Systems, learned something about the Python programming.

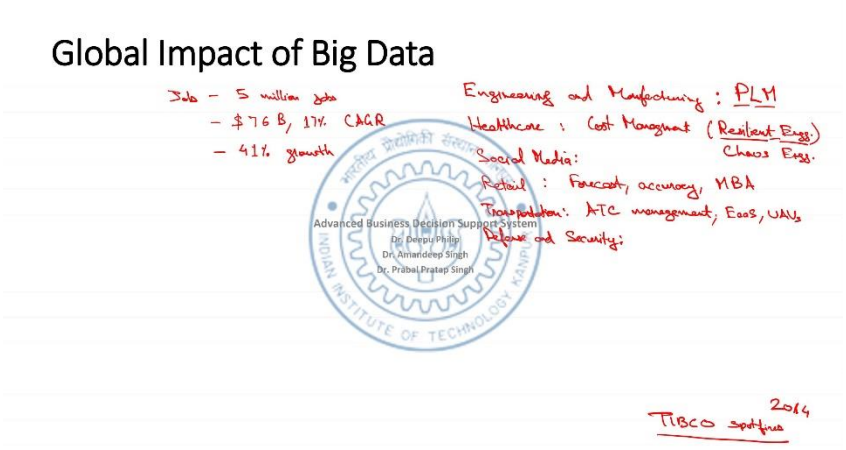


I would just like to give a brief Introduction to Big Data Analytics in this week. Though, the term is very common now for you to understand, but still a brief introduction to what is Big Data, what is Big Data Analytics, what is architecture of that. So, I will just talk about the Global Impact of Big Data, and in general what is Big Data, what is Big in the Big Data, and Architecture of a Big Data, and the layers of that, types of data we will discuss about, and we will see some of the Big Data Approaches that is Big Data Analytics Approaches.

What actually is Big Data when the data is available in large amounts that is the word Big comes into the play. Along with the amount of the data or the volume of the data, the variety of data is also high. For instance, let me say the data that is generated in the aircraft engines when we are trying to have the proper sensors installed at each point, each second data is recorded sometimes, at each millisecond sometime also the data is recorded. For instance, how the heat patterns are changing, how the engine is behaving or so. So, this generates a lot of data in the whole running of the engine for maybe, suppose

8 hours of a day.

For this thing, the data that is generated, is to be used somewhere, is to be used somewhere, data is to be converted into the information, information is to be then analyzed so as we get something usable out of that, something we can understand out of that. For that there are certain techniques, we have map reduce, we have certain other approaches for Big Data Analytics that I will just give a small introduction to. Then, I will also like to talk about one of the Big Data Analytics or maybe generally data mining techniques known as Market Basket Analysis. I will give an introduction to that, and we will also try to learn on the excel program how we conduct the MBA, that is Market Basket Analysis.



Let us talk about the global impact of Big Data first. If I talk about Big Data globally, there is a study given by TIBCO Spotfire that says that if I talk about job market around 5 million jobs are being created for analysts, for computer scientists, for mathematicians, for operational researchers, for people who deal with the Big Data, and if we talk about the amount, it is talking about a business of around 76 billion dollars that is according to this study which was conducted in 2014. The Big Data Analytics market was expected to grow 17 percent CAGR till the year 2020.

Then, they estimated 41 percent growth in the next 3 years from this time for the business to use Big Data Analytics. The growth I am talking about, then the data I am talking about here includes almost everything. We are talking about the modern media, maybe YouTube, the OTT or so. We are talking about data from social media, data from real-time information, we are talking about the data from different influxes of the data coming from different technologies as well.

So, if we talk about different industries in a way, there are certain industries which are impacted by the large amount of data being involved, engineering and manufacturing. Nowadays, we talk about product life cycle management. When we talk about product cycle management, we talk about the life cycle costing of the product, we talk about the cradle to grave data storage of the overall tracking of

the product even before the product has come to the market, the overall disposal of the product, how it is going to happen that is also discussed.

This is a part of product life cycle management. So, engineering and manufacturing has now gone beyond what it was thought of in healthcare as well. To reduce reaction time to the clinical events, to have better care, to better cost management, post COVID, there is something known as Resilient Engineering that came through the healthcare systems, but it has entered into the engineering and manufacturing systems.

Now, Resilient Engineering means when a complete system is missing the data or the strategies for it, this is also known as Chaos Engineering. So, then what algorithms would work, disaster management also, then we have social media, that is how do we capture the customer sentiments, reduce service cost to the customers, feedback on products and services, improved public opinion or so. In the retail industry definitely, the examples I am going to take in this week only regarding the Market Basket Analysis.

So, in the Retail sudden Forecast Accuracy, then we have Anticipation of the Demand Changes, Inventory Management Systems, Pricing and Market Basket Analysis, that what kind of the specific set of the basket should be set for the customers or for the future customers which would be going to purchase based upon the previous or the current purchase that they have taken.

Then, we have transportation, when it is airport transportation, we are talking about the ATC management. We are talking about fleet tracking; we are talking about the RR system. The RR system is Rolls Royce system, that is Engine As A Service, EAAS system.

Then, UAVs are there, then definitely in defense and security. In defense and security, the advanced modeling and analysis of defense scenarios, life cycle analysis of the defense systems, then we need to develop next generation weapons, surveillance, and so many things like this. These all contribute to what is the real need of Big Data.

Big Data, in general

So, in general what we have is, we have data, we need to get information out of it. What is data? Data is facts and figures which are recorded just for the sake of recording. So, it is not structured, it is not put in a regular way, it is not put in an organized way.

So, data if I put it as a definition is, recorded truth from a point of time, that is, it is a snapshot of an event. It is facts, figures, that really something specific, but they are not organized in any way, they provide no further information regarding the patterns or any context. If I am trying to take some pattern out of the available data, out of the available data would be maybe the engine heat that is generated at the peak running hours of the engine, what is the heat at the start of the engine, what is the heat that is being generated, this is only data being generated.

If I get maybe the average heat that is being generated each day during the peak running hours, this becomes a piece of information. If I take the standard deviation of the engine heat being generated for the last one month, that is the information.

So, information is we take data from an event, and put it into a narrative. A narrative put on the taken data, that is we convert or term the data point into something that informs your business. I am putting the word something useful, so this useful is we get some information only, it informs.

So, data becomes information when it is contextualized, when it is categorized, so as we give it a relevance and a purpose, so that we are able to tell some story out of it. Then, out of the data and information, we take out intelligence. This is where data analytics comes into play, it is one step further where information is used to drive decisions.

Instead of just telling a story that what is the information or so, we try to paint a picture of the story, how the story is being developed, what is the use of the different heat points or heat generated through the whole month during the peak hours of the engine, and how is going to affect the overall performance of the engine in the next five years. If we try to paint this story, the data is intelligence

that we take out of the data and the information. That is data, information, intelligence, these are all part of the same continuum when developing a solution to the business problem. All three elements, I will number them, should be part of our strategy. So, only then we can be able to use the data, that is the big amount of data. So, now this data could be if I say 1A and 1B, it is data, we can talk about the internal and external data.

Internal data is the company website, the comments within the company, the charity in the company, how to know our customer, the account balance within the company, the pay that we are giving to the employees or so, the internal behavior aspects of the people who are employees who are working or so.

External data is something that we need from external agencies, for example, government agencies, from maybe other banks, from maybe the competitors, from the social media, regarding the customer sentiments, so all these are part of the external data.

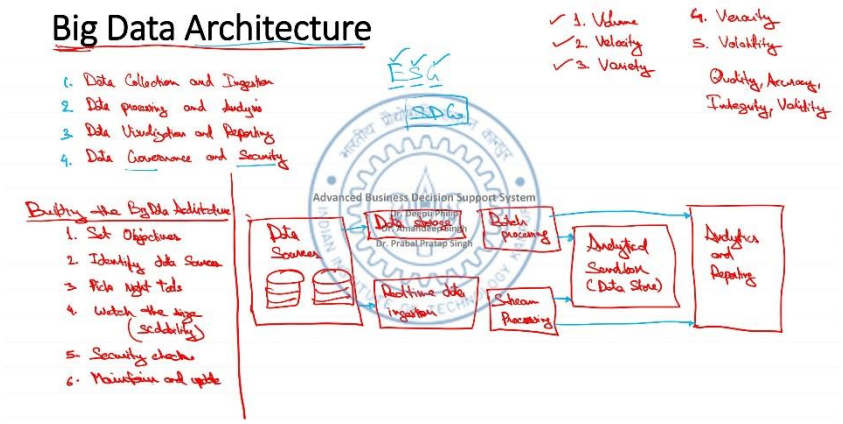
So, two kinds of the data sets could be obtained. So, another classification for one would be data could be, I will call it 1C and 1D, it could be structured data or unstructured data. Structured data, by the name it suggests it is data stored in databases with an orderly column and rows.

So, means the data items are clearly defined, the analysis requires less computational power here, but when the data is unstructured that is when the data is just available as a total data which is not available in an organized manner. So, it is not organized, so it could be texts, images, videos, audio files from device sources, and the dilemma is that 90% of the data that is available in the market is unstructured data.

Here comes the need of the data analysts to convert the data into information, so that we can get some business intelligence out of it. Another kind of data if I say, large amounts of data or I would say too big data or data could be too less, both of them are the problem. If I talk about too much of the data, the dimension reduction classification, the billing etcetera, and the Market Basket Analysis is not manageable, the size of the data set is sometimes not manageable, then patterns and behaviors, states are to be diagnosed, so many biases do come if the data is very big.

On the other hand, if the data is too less, that is the parameters are available in very less amount or maybe very less number of the data points are available, then we need to use some of the statistics tools.

For instance, if very less number of data points are available, we are only having a uniform distribution or we are only having something where only mode is available, we might have to use beta distribution or triangular distribution that statistical tools are used for. But, still here we need to manage errors, because the amount of data that is available is very less here.



So, let us talk about Big Data Architecture now, when I talk about Big Data, it only means, majorly the word literally only means, large amount of data, but there is something more to it. So, there are three measures of the large set of data. Number one is volume of the data, that I have already talked about, the large amount of data, then we have velocity of the data.

Velocity means the way I talked about the data points that we are taking for the engine heat that is being generated. Not only heat, we are also talking about the engine performance, we are talking about the engine vibration and so, a lot of the data is coming, and data is coming at very high speed at each second data is being generated. So, that becomes the velocity of the data, because we are talking about heat, we are talking about vibration, we are talking about the change in performance or so.

So, that means the variety of data is also very high here. So, these three parameters describe what is Big Data. So, data is of extremely large volumes of data, it is having high velocities, it is having wide varieties. Along with it, two more measures are there which are veracity, and volatility which means we are referring to the data itself, but also, we are talking about the technologies that perform all the functions of the varied collections of the data to solve the complex problem.

So, here when I am talking about the veracity of the data, veracity means the data quality, the accuracy I will put it here quality, accuracy, integrity and validity. Validity means the data that is there available with us for how long is this data valid for the specific period of time or not. That is the data is volatile as well, the data that was generated for the last month might not be usable this month itself.

For example, in Market Basket Analysis for the month of October, and November in India when there is a Diwali season, the kind of the baskets that we need to set should be separate or are separate. So,

the data that is generated for this month should not or might not be relevant in the month of December.

So, that means the data could be volatile, and a lot of veracity of data is also there. So, now we need to have an architecture to deal with Big Data in this case now. So, when we talk about the Big Data Architecture, that is how the Big Data Framework is there, and how it defines different components, different processes, different technologies which need to capture, store and process the data so that we analyze it. So, Big Data architecture typically includes four of the layers. These four layers are data collection and ingestion.

Then, we have data processing and analysis. Then, we have data visualization and reporting. And, the fourth layer comes the data governance and security. If I put them into an architecture so I can say I have data sources, different hard disks are there that contain data. We have data storage from here, and we have along with data storage that is the secondary data we have the real time data ingestion as well.

We have an analytical data store or we call it analytical sandbox or data store which takes data from the stored, and the real time in various forms. It could be batch processing, a specific set of the data, because the variety is large, because the velocity is large, specific batch that is to be analyzed. So, stream processing could be there. Let me say I would only need the heap data only, only this specific stream only. So, that is stream processing.

So, this helps later after going through a certain analytics process by producing, and applying certain mechanisms over it, we try to get analytics and reporting. And, if I try to connect the systems, I can connect data sources to data storage, and the batch processing, and stream processing can give data to the analytical sandbox here. It can directly give the validation data to the analytics and reporting. So, this overall system is known as Architecture of Big Data Analytics. So, these are the layers we have a data ingestion layer.

This layer is responsible for collecting data, for storing the data from various sources. Data ingestion is very important because when we extract the data from various sources, and load it into the data repository the data ingestion is a key component of Big Data Architecture. As I said, all the points, data, information, intelligence is required. So, data is to be ingested, and whatever data is required specific streams, specific batches are to be ingested. This becomes layer 1.

Then, comes data processing, that is the collection, data cleaning, and preparing the data for analysis. This is critical to ensure that the data is of high quality, and is ready to be used. Then, comes the third layer, that is data visualization which is responsible for creating the visualization of the data. That is the graphical representation of the data, so that humans or the people who need to understand them, the executives, the engineers, the operators are able to understand that. So, this layer is important to make data accessible to the people who are actually going to use it.

So, then comes the fourth layer that is data governance, and security because nowadays we are talking about ESG, ESG is Environmental Social Governance. So, when we are talking about the ESG, the companies are talking about the sustainable development goals, and there are certain agencies which help us to get the data from the different viewpoints from the environment viewpoint, from the social viewpoint, from the governance viewpoint, so that we are able to meet the sustainable development goals set by the United Nations.

So, data governance and security also become one of the major layers now of Big Data Architecture. So, when I am talking about the architecture, for certain steps which I could see we need to first set objectives. When I say set objectives, we need to understand what is the architecture, what kind of decision we need to know, and what kind of stream or batch of data we need to pick. So, then we need to identify our data sources.

Along with it we need to pick the right tools. There are certain online data storage tools available, data analysis tools available. Do we need to use them or can we use very basic languages such as python that we have learnt in this course to ingest the data, and to have small analysis done over it. When we try to do it, we need to also plan or see the scalability. Watch the size, when I say size, I am talking about scalability. If I am able to analyze, maybe for the Market Basket Analysis, the 5000 pieces of specific product which are sold in one year or so.

So, this could be scalable to 10 years, this could be scalable to 20 years or so, this 5000 should be multipliable to be 5 lakhs to 50 lakhs or so, scalability point is to be considered. So, these are the specific points that we keep in mind. Along with this anyway, a security check is to be taken, and we need to keep on monitoring, and maintain along with updating. So, this is building Big Data Architecture. So, with this I will take a rest, and I will meet in the next lecture where we will talk about Market Basket Analysis.