Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

Week-11

Lecture-55

Application of Big Data Analytics in Mining

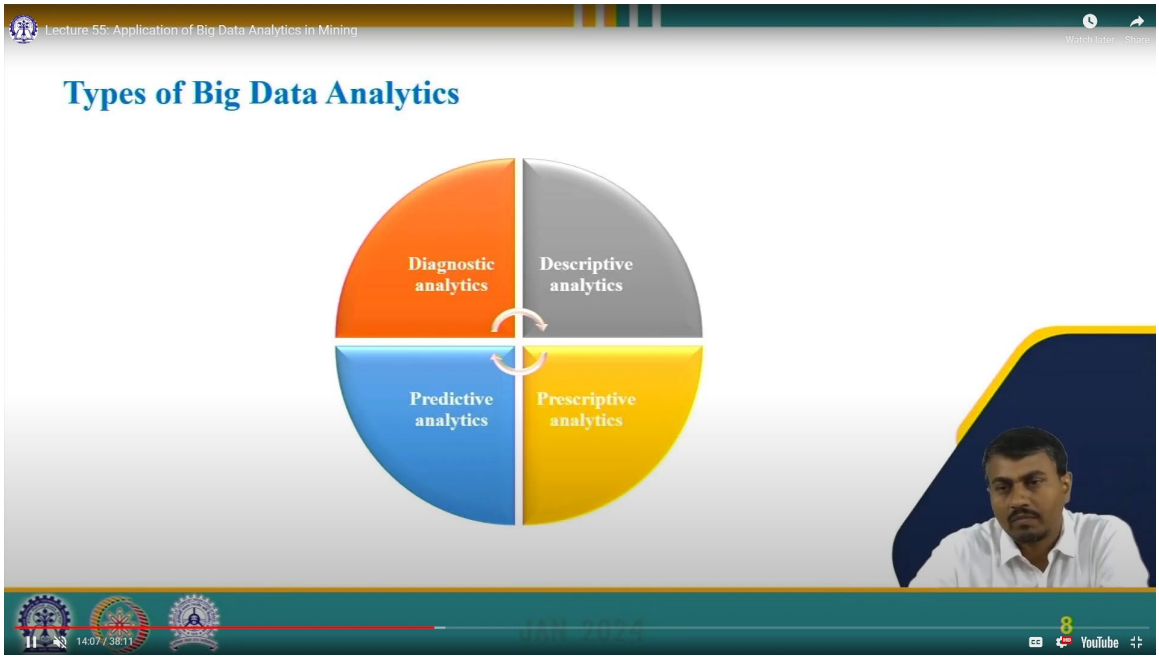
Welcome back to my course on mine, automation, and data analytics. Today, we will discuss the issue of big data analytics in mining. In the first part of this particular course, you have seen that we are integrating various sensing systems with mining operations. There are many more parts to be covered in mining operations that we are not able to cover in this particular course. However, the general engineering, the way sensors sense information, their security concerns, and their structure are basically covered in various examples. So, the next level of operations is to handle this data efficiently and use this data for further optimization of the mining process and mining operation. So, that is basically our subject topic today. What are the issues that arise in handling this big data? So, we will try to focus on various issues today related to big data analytics in the mining industry. So, we will cover the following: What is big data, actually? Then comes big data analytics, and there are different types of big data analytics we will try to cover. Then the three Vs of big data variety, volume, and velocity are very important components. What is the difference between small data and big data? The architecture of Big Data Analytics is a Hadoop-based structure. Then the big data benefits, and finally, we will discuss the challenges of implementing big data analytics.

So, what is big data? Probably, if you remember, when we gave an example, the Automated Haulage System. So, an automated haulage system means that the haul truck operates autonomously. So, there is no human operator, there is no one to steer manually, everything is a sensing mechanism. Now think about how a car basically moves. Now scale down our problem to a normal car. So, a car basically starts with what? Whether the seat belt is on, then the ignition switch will be enabled. Then it will see whether the road is clear to move or not. So, it has to identify whether there is a green signal, a yellow

signal, or a red signal. If it is a red or yellow signal, it will not move. Only at the time of the green signal will it move. So, identification of the color code and the traffic rules is necessary. The second is at what speed it will be able to go, maybe on flat land or on an inclined slope. So, that is basically based on the operator experience and the capability of the engine. So, if, and suppose, during the movement of a vehicle some obstruction or somebody came into it, what the driver does? He or she will suddenly apply the brake. So, to suddenly apply the brake on what basis? Because I identified a person who is just coming into our path. So, that identification is done by the operator who is sitting there. So, imagine a scenario when the automated haul truck is operating without any operator, a physical operator sitting on the operator cabin. So, it basically collects a huge amount of data in real time. There are lidars, there are radars, there are different sensors that monitor tire pressure, and there are different camera orientations in the front as well as back door that always capture the image around the truck or the dumper. Now, with an average camera image, can you imagine a normal picture? It is around 3 to 4 MB. So, 3 to 4 MB, so the number of images when it is in motion, the number of images will be acquired, So, 12 MB per second of data is coming through only the camera. So, 12 MB into 60, 1 minute means 60 seconds. So, 12 into 6, 60 means 720, around half, and half giga means 700, or some 720 MB. So, if it is in hours, how much data can you imagine? So, there are a number of dump trucks operating autonomously. So, the control station is getting all this data, and based on that, it instructs. So, how much of the traffic load, the velocity of the data that is coming, and the instructions are going, you can imagine. So, this is basically a general picture that I want to give you of big data in a mining scenario. In a real-world example, you all, somebody asked, where is Manhattan? Where is the kolar gold field? You will Google it. So, like, crores of people are Googling something. So, Google means search engine, which is going to start in the database. So, a few crores of traffic are in line to search the data. So, there is a big data house. So, you can imagine the traffic. How is this managed? This is well managed, within a second, you are getting the output response. So, this is a very well-managed structure. They are handling very, very big data. So, this big data actually handling is very difficult. But in our real-world case, it is giving us a big advantage compared to our 20-year-old civilization. We are getting benefit from this big data in our day-to-day lives. So, this big data is basically

one aspect of a big volume, sheer volume. It is accumulating at a very high speed and not only that, there are various type of data, various type of data. So, there are three categories that we can think of cumulatively, this is the big data that we are looking into. So, globally, there are different analytics being developed and practiced. So, some of these analytics are used in the mining industry and are allied to the mining industry. We will try to cover that in subsequent lectures. So, the purpose of big data analytics is to extract useful information, useful information because it is a machine. A huge amount of raw data is in the repository. So, until and unless it is used for the specific purpose for which you are building this data repository is useless. So, it is very essential to analyze this data and make connections with the real-world scenario and the data that we are getting. Finally, lending to a fruitful conclusion that gives benefit to the organization for the purpose that we are looking into. So, basically, this big data analytics is for the decision-making process that makes our decisions faster, more efficient, and more beneficial. So, all these organizations are benefiting from it, and we are basically doing more business in very quick intervals compared to the 20-year-old scenario. These are the features that we have introduced. That is, the variety, volume, and velocity of the data are very important. That makes the qualitative strata, or the quality of the data, or the size of the data, or the what aspect, what kind of data structure is required for it, and what kind of processing is required for it. So, these are the three parameters that are very important in big data analytics. So, this basically helps compared to the traditional data that was available on the dashboard. So, it is giving more advantage, getting more gain out of the other players in the market, or all other players are also doing equally. So, the efficiency of your processing and your analytics is basically the gap, and you have to reap the benefits of it. So, managing the quality of the quality of the data is very important because if the data is not of good quality, what will happen? and you are acquiring, or if the data acquisition you are doing is large, So, if the data itself is erroneous, it will land at what? Towards the wrong conclusion. That will damage your business. So, the data acquisition, the frequency of the data acquisition, and the quality of the data are very important. So, poor data is a problem, and we have to address this issue. So, given the error in the data, we have to consider it, rectify it, and take corrective actions so that we do not go out of our aim or come to the wrong conclusions. So, these are the challenges

that we have to ensure. We have to ensure the quality of the data so that we are in the competition.

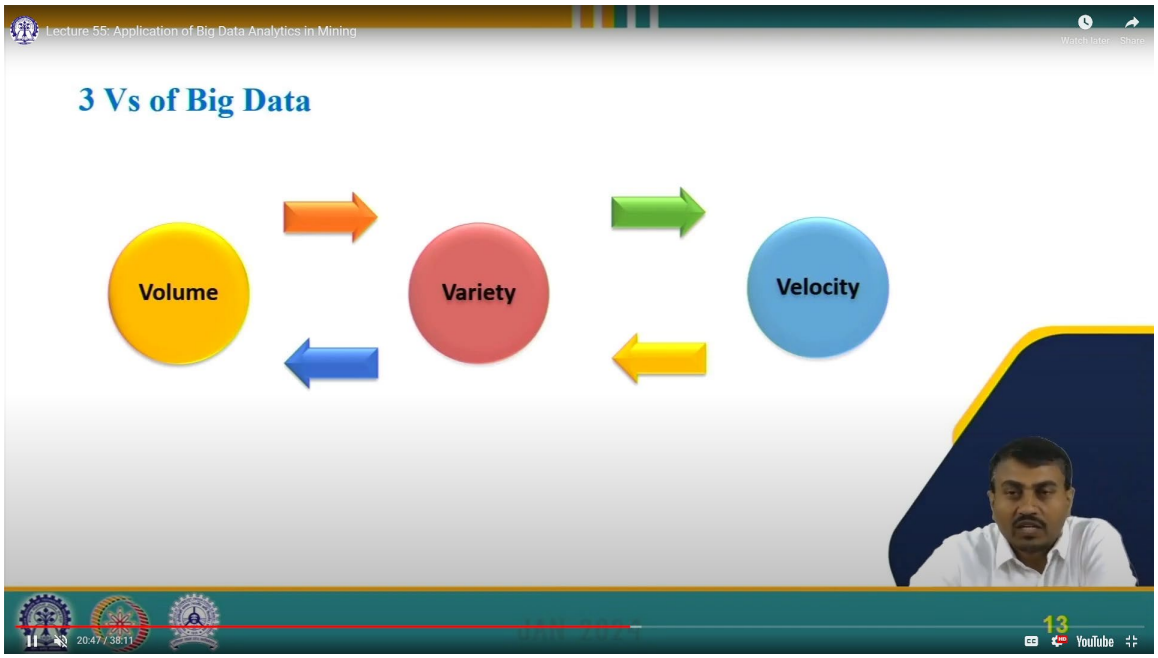


So, the types of data analytics are descriptive analytics, prescriptive analytics, predictive analytics, and diagnostic analytics. So, probably the descriptive analytics you can match with the first part of the statistical tool that we have covered. And basically, the other three prescriptive analytics, predictive analytics, and diagnostic analytics are basically on the second part of the inferential statistics. So, let us see what diagnostic analytics is. So, it is one of the more advanced forms of analytics, and starting with the question, why did it happen? Something has happened in the operations, maybe a sudden market downturn means your product is not able to identify the reason, okay? other parameters were nearly the same, but your production or your sales in the market have drastically decreased. What is the reason? So, based on the data you are collecting, you have to act and identify what possible reason is leading to this effect. Remember, whenever we apply this method, this data analytics, we are accepting the fact that everything related to the business and related to the operations is interlinked and connected with causal relationships. So, this means if something happening, something happening means there is a reason and that reason is physical reason, that reason exist and based on the pattern, it is possible to find out the reason or pinpoint it; we will be able to pinpoint that reason

based on this causal relationship. So, based on this strong foundation, the data analytics is built. So, here, we try to understand the behavior and event from the data, and based on that, we try to find out what possible reason is there that leads to the drastic drop in sales. Anomaly detection, similarly in the machine, particularly, everything is well and good. All sensor data is good, but ultimately, the machine is not performing well. So, there is something, so based on these analytics, we try to find out what could be the possible reason for this engine not working properly. That may be represented by the sensor data, or extra sensing may be required to get extra information about the situation in the machines. So, this is a very useful tool for diagnostic analytics in different applications. So, here what we are doing? We are finding patterns in the data set. We are doing the data filtering and using the probability theory. We are using regression analysis and many more to do these diagnostic analytics.

The second is the descriptive analytics: It is one of the major forms of analytics to monitor current situations and operational performance. If you go to a mine, particularly the fleet management system, the data is shown: the performance, number of shovels, number of trucks operating, how many drilling machines are operating, and their curve. It is a kind of steady growth curve, okay, or a flat kind of curve, which means operation is steady, okay. So, that is basically the descriptive analytics. Based on the collected data, it is represented in a very well-formed graph. From that, we basically find out the trends of the data, whether the trend is good, if something happens, some drop is observed, something idleness has occurred, or data is not coming. So, first, we will try to identify whether it is due to the network. Second, we confirm at the base level from the supervisor whether it is actually a failure or not. So, this descriptive analytics basically helps to maintain a good performance on the mine site. So, this is used for production, finance, and sales. Prescriptive analytics. So, it is basically built on descriptive and predictive analytics, okay? So, this is basically done for the organization for various simulations, okay, based on the past histories, basically an evidence-based approach, okay. This helps to reach a better decision by analyzing the data, analyzing the meaning of the data, and we are going ahead with the decision-making process, which helps the company get better business.

Predictive analytics: This type of data analytics focuses on prediction and future outcomes based on the insights derived from the data. So, there are different sophisticated tools used for that, and mainly machine learning based on statistical tools are used. The future of outcome prediction is that its primary goal is to predict and forecast performance based on historical data and patterns.

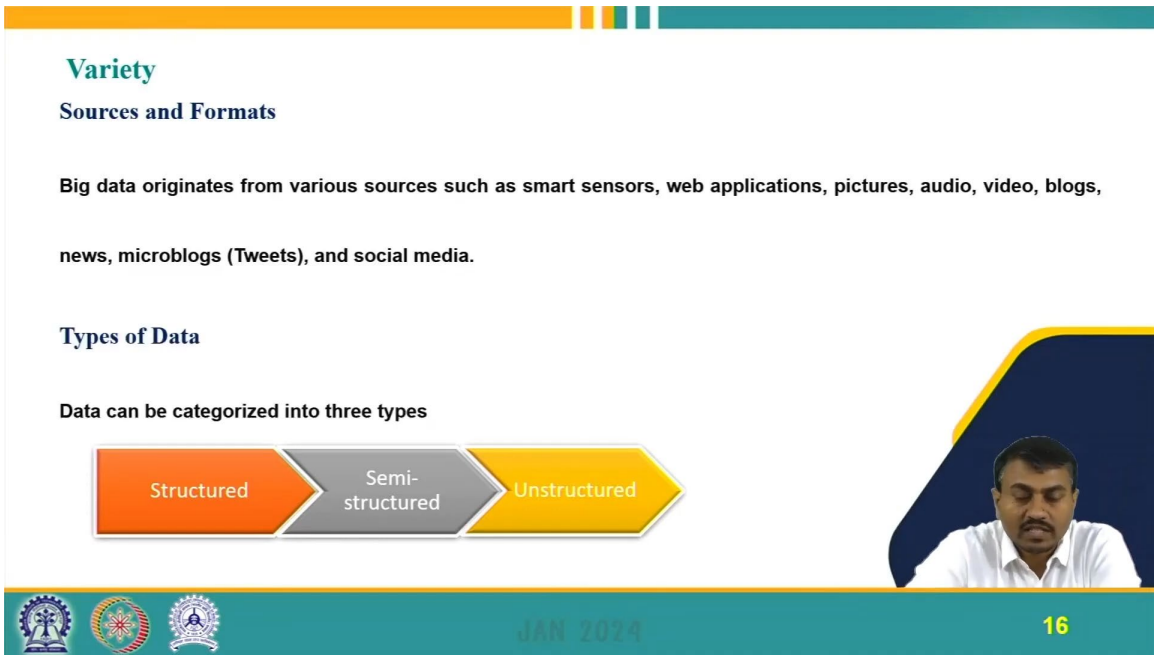


The three Vs are volume, variety, and velocity. In the data analytics, we have already highlighted that these are the three terms that basically define big data. Actually, volume is so high that you cannot handle it with traditional methods. The variety is so high that it is very difficult to apply normal tools, and the velocity of data acquisition as well as transmission is so fast. you have to give the real-time instruction, reaching a decision, and based on that, the operation will move, so the velocity. So, these three are working hand in hand together and give a very complex structure to the big data.

Volume: So, by definition, you can understand that it is a big volume, and our technology is to be evolved and based on an advanced version of the technology. Today's big data may be smaller tomorrow, and so on. So, this scalability is a big issue, and we have to understand this scalability in the prospect of mining engineering as well as other allied industries. This data size will increase in the days to come. So, advanced technologies are required to handle these data. So, processing this large amount of data is important, and

when the size of the size of the data is large, that requires a good amount of time. So, that is also an issue because you need a better kind of processor to handle this kind of data and respond to the data that you are receiving in the decision-making process. So, the abundance of data makes it challenging to sift through and find relevant and meaningful information, pausing, and difficulty in data analysis and decision-making.

So, basically, now we will discuss the data variety. So, there are various formats for the data. More data that we are receiving is through the sensors, lidar sensor, thermal sensor, temperature sensor, pressure sensor, and various camera sensors, which means various sensors. There is data through the web applications for the business, particularly for the customer-oriented business. Pictures, video blogs, news, microblogs, and social media are also nowadays used for these big analytics and business. So, there are various types of data.

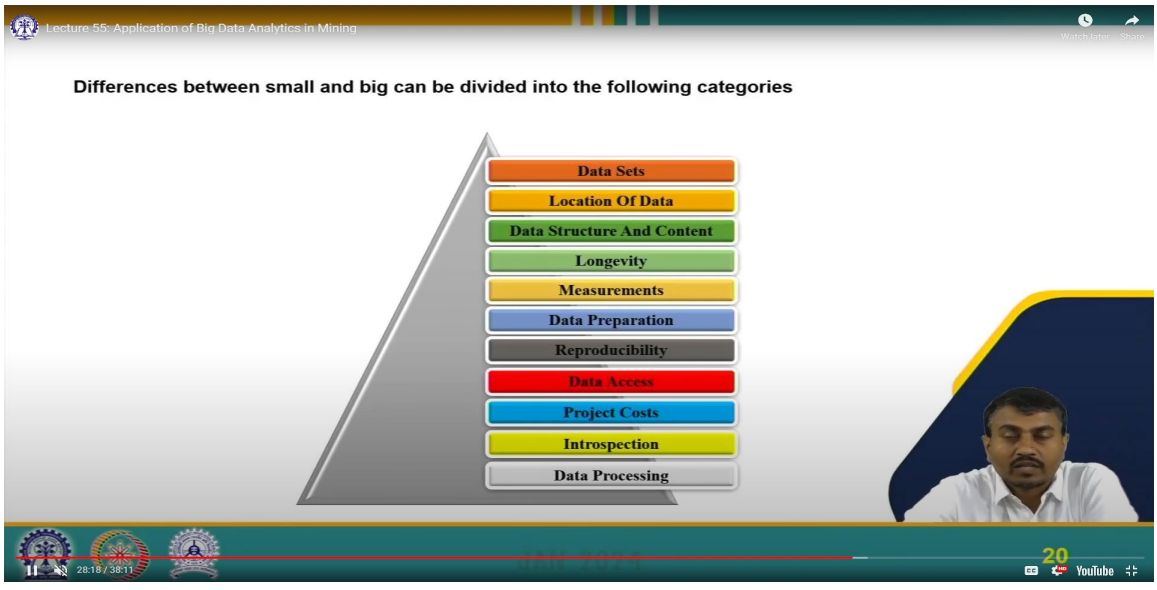


So, principally, these data are subdivided into three categories. First is the unstructured, second is the semi-structured, and third is the structured. So, let us discuss the structured data. So, this represents a fixed structure of the data, particularly when you are receiving data through a sensor. It has a very structured data format row and column, you know it. So, that is a very useful thing, and you can change the data format based on the programming you are doing on the sensor. Finally, it is structured data. Semi-structured

data has certain flexibility associated with the fixed structure; an example is XML documents, where some kinds of data may be presented in multiple ways. Unstructured data is expressed in natural language with no specific structures, particularly our day-to-day experiences, and newspapers, emails, blogs, and power points are basically examples of unstructured data.

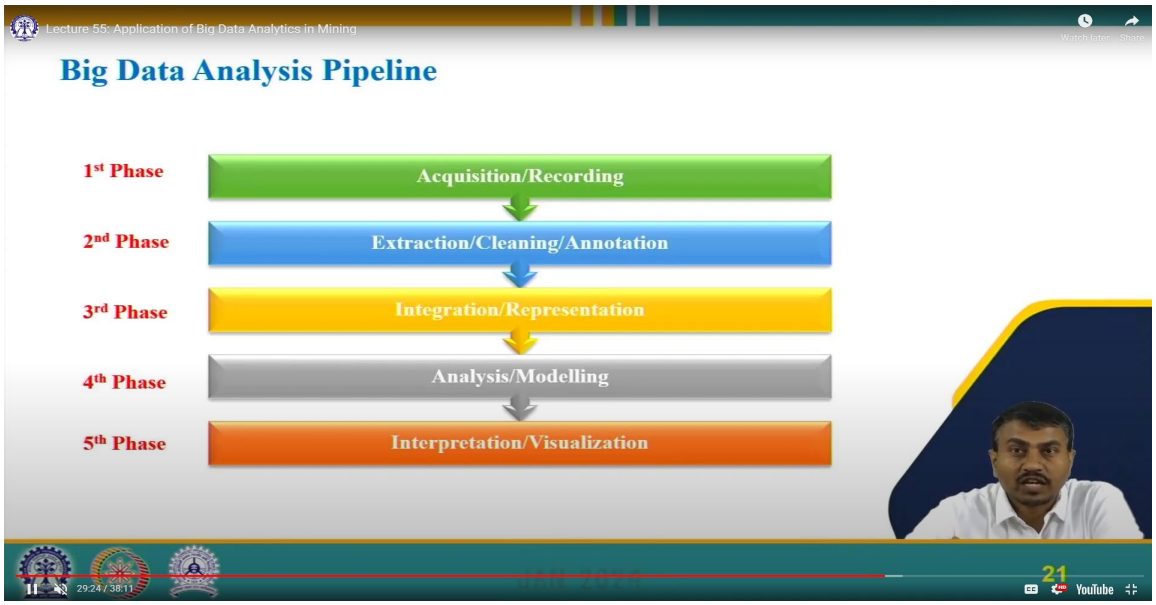
Velocity is a very important consideration. So, velocity refers to the continuous streams of data incoming that need to be processed in real time as they flow into the system and handled. So, the real-time big data is generated from sources such as positioning data, GPS, sensors like motion or picture sensors, and IMU sensor data as well. So, velocity is a very crucial factor in data analytics. The speed of processing the data from acquisition to decision-making is increasing with technological advancements. So, faster data velocity enables businesses to react quickly to changing situations.

So, the difference between small and big data. The data that is used in business analytics can be small or large. The term small data is synonymous with traditional data. Traditionally, data is defined as electronic data that is stored in a database or data warehouse, and big data is electronic data for which management challenges have exploded in three dimensions: volume, velocity, and variety. So, when the big data term is used, it can mean that the amount of data is large, the speed of the data in and out is fast, or the range of the data type or source is wide.



So, these are the types of data you can subdivide into: data processing, data introspection, which is also a kind of analysis. Project cost to handle this data, how much cost involved, then the data access, whether this data is to be accessed by everybody or somebody, to the principal nodes or to everybody that may be taken a decision based on the feasibility and the purpose. The reproducibility of the data is also an important factor. Data preparation is a very important factor. When the data size is small, data preparation cannot be done manually; it needs to be automated. Then measurements; whether the measurements are proper, the checking should be proper. Then the longevity of the data depends on how much time you are keeping the data. For example, with the CCT camera we are fitting for a range, we are storing the data, then we are removing the data. Data structure and content, the location of the data, and the data set. So, these are the aspects that we need to consider when basically dealing with the data.

Now, we will try to understand the data pipeline that is used in big data analytics. So, the first phase is the acquisition, or the same meaning is recording. So, sensor, we do not use



the recording term, but it is ultimately an acquisition; data is captured through different sensing mechanisms. Second, we have to extract meaningful patterns from the data. Before we do, we have to remove some error terms and missing values. So, pre-processing is required here. Then we have to extract the meaning of the data in the second phase. The third phase is integration and representation with your modules and

with your facilities. The fourth phase is analysis and modeling. finally, it is interpretation and visualization in a big sense, in a big graph. So, it is a kind of representation that is done in big, big company headquarters so that everything is visible in quality and quantity terms.

Acquisition and recording: It is the first stage, and you are recording the data and automatically generating the metadata that describes what the data is recorded, how it is recorded, and how it is measured.

Extraction and cleaning: So, whenever you are means of receiving data or collecting data, you need to transform it into a structured form and do some kind of normalization. Standardization may be required, and some amount of cleaning may be required. So, these activities fall under the category of extraction and cleaning.

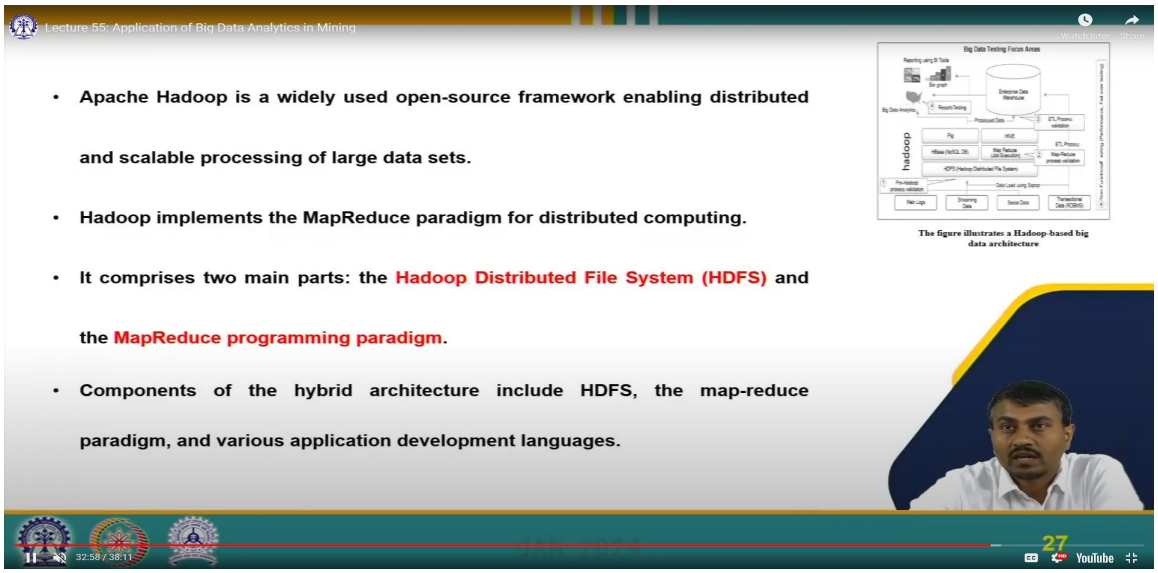
Integration and representation: These basically automated data handling processes, including locating, identifying, understanding, and citing the data, must be automated for effective large-scale analysis. we have to filter the data for much use, and we have to compress. Then, raw data requires processing to enhance usability for analysts. Then small data versus big data. So, small data analysis can handle all data at once. But for the big data, we have to identify a suitable size module that will analyze this part of the data, the next, based on the availability of the processing power.

Data analysis and modeling: So, data analysis and modeling are integrated with the big data analytics problem, and these give us the solution and insightful conclusions. What meaning is there in the data?

Interpretation and visualization: This interpretation and visualization is the most important part, and this represents a compact way of representing the meaning and the performance of the organization in a very short way.

Hadoop-based big data architecture: So, when you are handling big data and the data rate is high and the variety of the data is large, you need to apply some architecture, and based on that architecture, you have to handle that data. So, Hadoop-based data is basically similar to NoSQL infrastructure, and these architectures integrate data warehousing and

Hadoop-based infrastructure in hybrid models. So, there are different varieties of it. So, these basically implement the map-reduce paradigm for distributed computing, and it comprises two parts: Hadoop distributed file system HDFS and the map-reduce programming paradigm. the components of the hybrid architecture include HDFS, the map-reduce paradigm, and various application development languages. So, this is the structure of the Hadoop-based big data architecture.



Benefits of big data: Improve decision-making because ultimately you are getting a benefit from this data. What kind of problem is coming, whether the performance is up to the mark based on this data, and based on that, you are reaching a fruitful decision. So, that helps grow the business. Increase agility and innovation because, with the data and the patterns, you are finding new solutions and new insights that help you get a better meaning from the data and better plan your work. So, that helps to increase the agility and innovation within the structures, and that may allow you to launch new products and add new features to your business. Better customer experience because combining and analyzing a structured data source together with unstructured one provides you with more useful insight for the consumer understanding, personalization and ways to optimize experience to better meet consumer needs and expectations. Continuous intelligence because you are continuously receiving and analyzing the data. So, there is a lot of opportunity for growth and value. More efficient operations because you are saving a substantial amount of time and doing the processing very quickly. So, you are ultimately

saving a lot of time, and that time is basically giving you an extra saving of money. Improve risk management because you know better because you are confident enough in the data showing the health of the organization. So, you can handle risk in a better way. Challenges of implementing big data analytics. Speed of data growth-So, this speed is basically increasing day by day. So, there should be a solid infrastructure to handle this kind of situation, which needs a network, security, storage, and might be difficult. we need some active work on this particular domain. Integration is complex because there are different modules and different kinds of structures. So, data silos within the organizations are there. So, integrating all these is a difficult thing. So, this is what the engineers are working on, and we hope a big solution will come. Problems with data quality-Data quality is a very important factor until and unless you check it and understand it. So, data quality needs to be managed, and we need to guarantee that data quality is managed and that the that the data is accurate. Otherwise, we will land on very wrong conclusions, and that will cause a catastrophe in the business as well as in the organization. Security concerns. So, nowadays, cyber security issues are coming and growing. So, and particularly for this kind of complex data structure and complex data types, it's very difficult to manage the security and frame a policy to secure this kind of data. So, globally, different companies and organizations are working on that so that we can get better security in the future.

So, these are the references. So, let me conclude in a few sentences. So, we have covered the big data, the large volume, its structured and unstructured way, its high volume, high velocity, and high variety. Then we have explored the large data set patterns and its insights to reach fruitful decisions or informed decisions using an evidence-based approach. we have discussed various types of big data: descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. We have introduced the concepts of volume, velocity, and variety. Then we have contrasted the features and challenges of small-scale data versus big data. We have explored the pipeline of big data. We have introduced the Hadoop framework for data processing and storage. We have examined the advantages and disadvantages of big data analytics. And we have discussed the challenges and hurdles of implementing big data analytics in organizations. Thank you.