Introduction to Industry 4.0 and Industrial Internet of Things

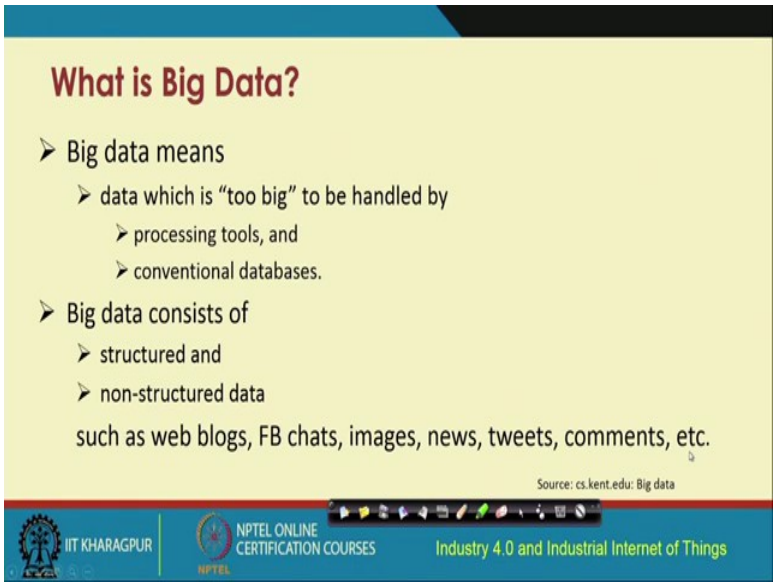
Prof. Sudip Misra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 15
Industry 4.0: Big Data and Advanced Analysis

In Industry 4.0, we are talking about use of IoT, IIoT, Industrial IoT, which essentially are heavily based on the use of different sensors and actuators. So, these sensors typically sense lot of different data and this data are sent continuously after they are being sensed. So, this data will have to be processed. So, what kind of data we are talking about? So, this kind of data in the context of IoT, IIoT, Industry 4.0, exhibit the behavior of something known as big data. So, we have to analyze the big and the traditional forms of data, to gather meaning out of it.

So, when we talk about this analysis whatever we talked in the context of machine learning in AI those kind of methodologies are also applicable over here. So, analytics, statistical analytics, machine learning-based analytics, use of different methods of neural networks, SVM, all of these could be used to analyze the data, in addition to the statistical methods, the multivariate statistical methods that we are all familiar with. So, those could also be used to analyze the data. So, what is important is that you try to gain meaning out of the data that you receive from these IoT devices in IIoT.

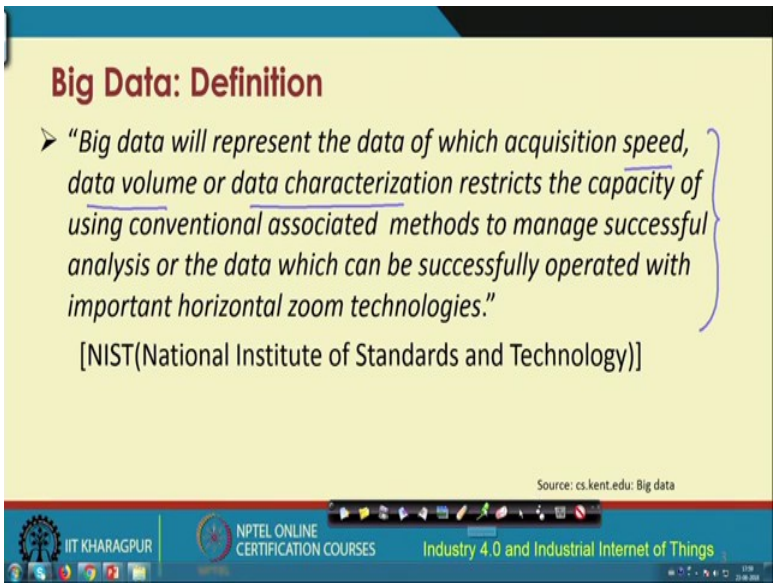
(Refer Slide Time: 02:09)



So, what is big data? So, data which is too big to be handled by processing tools and conventional databases, conventional processing tools. So, big data are typically the ones which exhibit the properties of non-structured data, but they could also have structure data. Structure data means what? Structure data are the ones, which could be stored in the form of tables in different databases; for example, relational tables. So, for example, databases like MySQL, Oracle, they use relational tables.

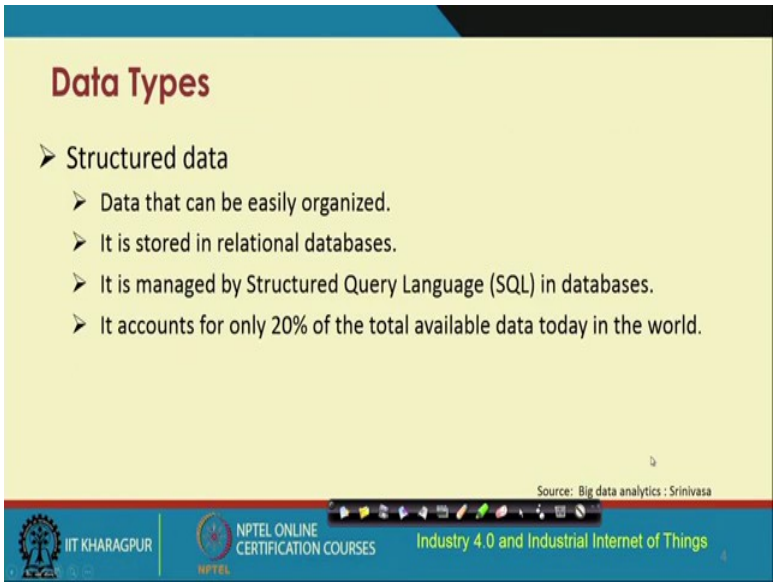
So, you can store those data in these relational tables in the form of these tables you can store this data so that is structured way of storing the data. And not that all kinds of data could be stored in the form of tables. So, certain data for example, Web blogs, Facebook chats, images, the newspaper blogs. So, all of these will exhibit non-structured behavior of data and they cannot be used, they cannot be stored in relational tables in the traditional form. So, these are like big data, non-structured as well as structured data, which will have to be handled in certain ways.

(Refer Slide Time: 03:39)



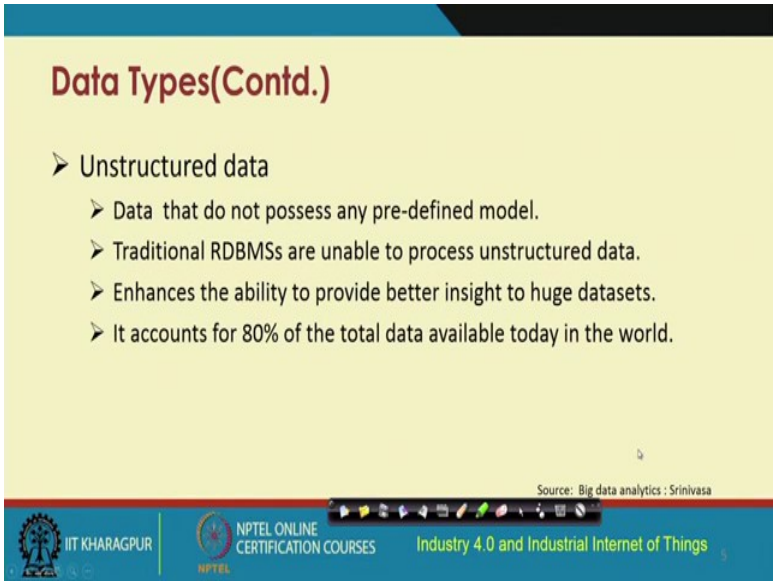
So, let us look at some of the key attributes of big data. We are talking about representation of data, which acquisition speed the data volume, data characterization, restricts the capacity of using conventional associative methods to manage successfully; the analysis or the data, which can be successfully operated with important horizontal zoom technologies. This is as per one of these definitions of big data.

(Refer Slide Time: 04:21)



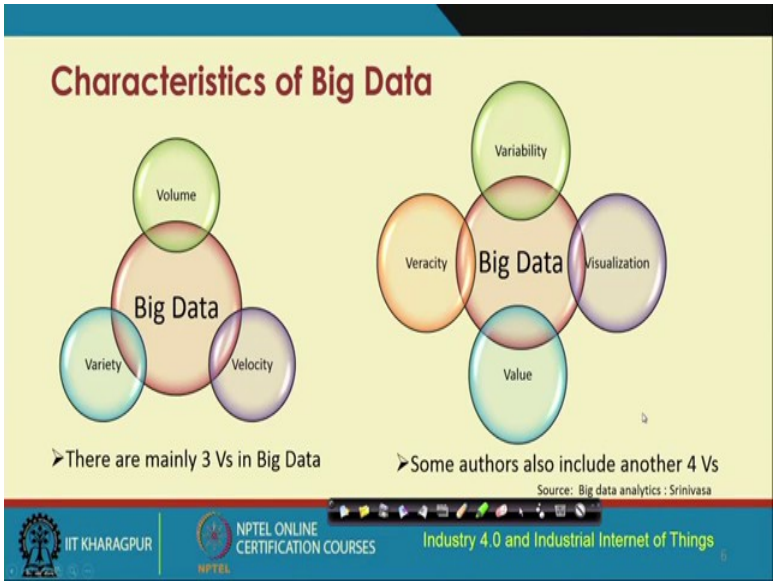
So, I was telling you that in the context of IoT, IIoT, we will get both structured data as well as unstructured data. So, structured data as I was mentioning before are the ones, which can be stored easily in an organized fashion in typically relational tables. So, if you are able to store this data in the form of tables, so you can use a language like SQL, Structured Query Language to query these tables to access the data, that are stored in them.

(Refer Slide Time: 05:03)



However, if you are talking about unstructured data you cannot store this data in the form of RDBMS tables and there is no predefined data model that could be used to analyze this data. So, basically most of the data that we encountered in the present-day world, would be unstructured. So, you cannot use the traditional database techniques that you are already familiar with in order to analyze this kind of data. And at the same time not only that these data are unstructured, but also this data come in huge volumes, so you are talking about huge datasets that will have to be stored in certain way.

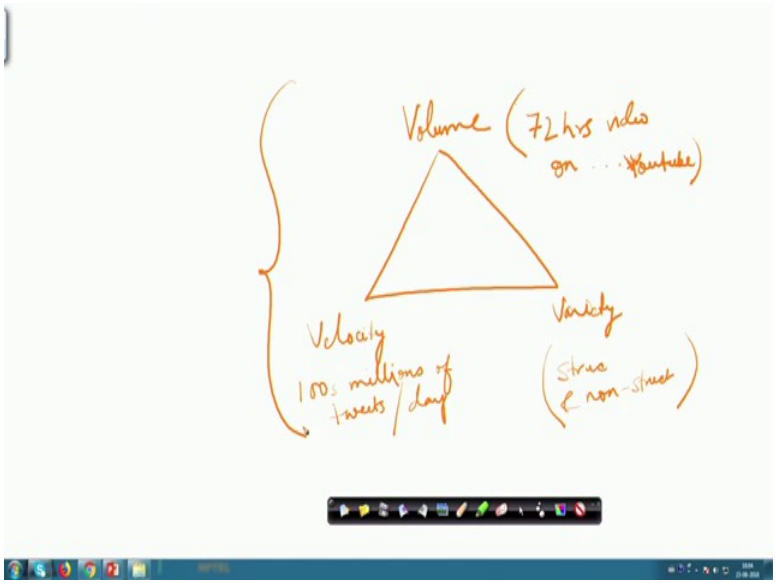
(Refer Slide Time: 05:49)



So, these are some of these characteristics of big data we are talking about volume, variety and velocity. So, these are the main 3 V's of defining big data. Data which is large in volume which exhibits variety and which arrives at high velocities at the processing server or processing agent. These kind of data traditionally were termed as big data, but as researchers what with big data more and more they also included few more different other aspects; other aspects other V's in fact. So, other V's like variability, visualization, value, veracity, so these are 4 additional V's that will also characterize big data. So, traditional 3 V's plus the 4 V's, 7 V's would be used to characterize big data.

So, let us try to understand these in little bit more detail. So, let us talk about the 3 V's first volume, so volume, velocity, variety.

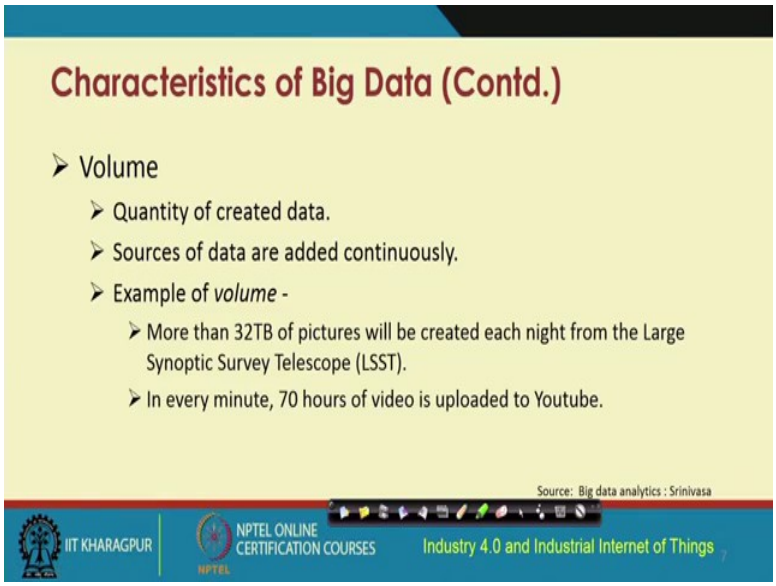
(Refer Slide Time: 07:07)



So, these 3 V's let us consider. So, what is this volume? This volume is large volumes of data we are topically talking about. For example, some 72 hours of video on some server such as let us say YouTube; let us just take for example, it could be anything. This is huge kind of data and huge in volume. Then if you are talking about velocity, if you talk about twitter for example, or similar kind of other platforms we are talking about some 100s of millions of tweets, on average per day. So, as you can understand that handling this kind of velocity of data is very important.

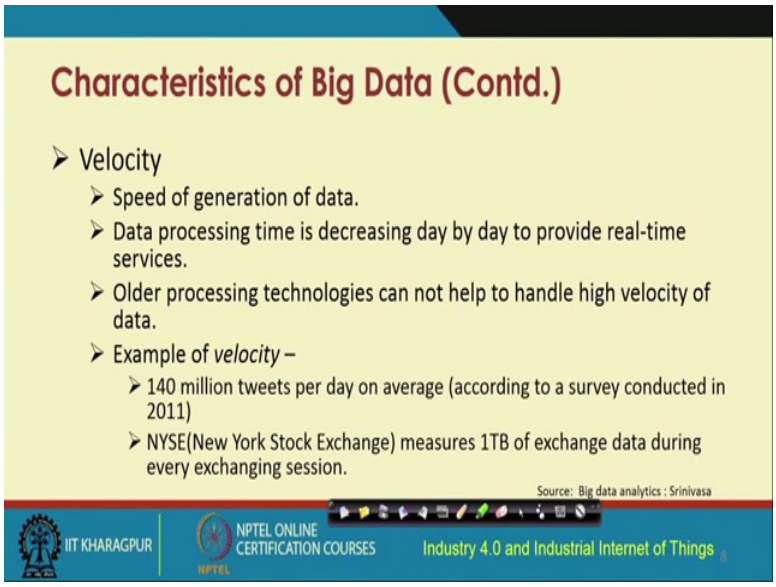
And then variety and this variety could be you can have both structured as well as non-structured data. So, variety of data for example, image data, text data, then video data, all coming together at the same time, that is variety. So, in big data context, we are typically talking about data, which is coming you can think of it as a pipe, some kind of a pipe through which data are coming continuously, huge volumes of data are coming at high velocities and this data the composition of this data are all very varied, text, image video and differ different types. So, all these data coming at the same time, handling this kind of data, using conventional data warehousing, database techniques, etcetera, it is not possible. So, people are talking about how you can have newer methodologies, newer tools in order to address this kind of data. So, let us just go back to what we were discussing in the slides.

(Refer Slide Time: 09:43)



So, we are talking about huge volumes of data, we are talking about data not just in gigabytes and terabytes, we are talking about petabytes, hexabytes and so on of data, how do you store, that kind of data that is very important.

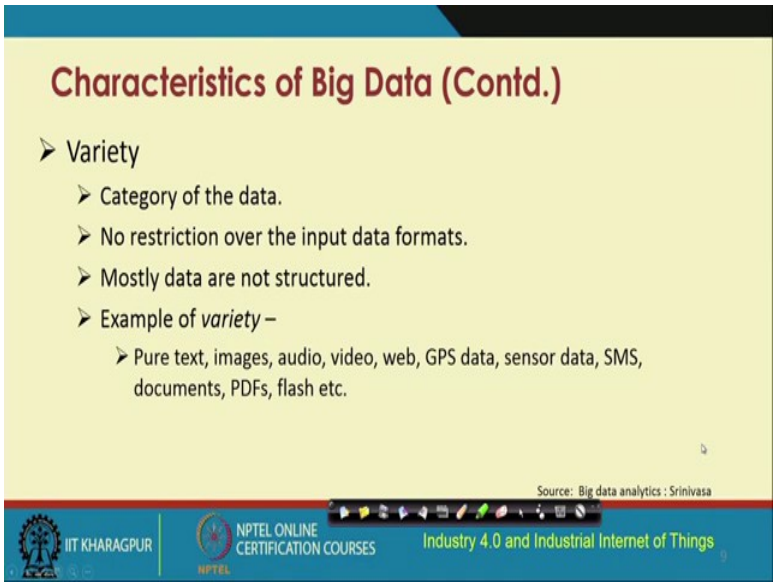
(Refer Slide Time: 10:01)



Velocity, speed of generation, 100s of millions of tweets per day on average that is huge velocity of data coming in, data getting generated. So, if the data is getting generated you have to handle the data and this is a high velocity data that is coming in and that has to be handled.

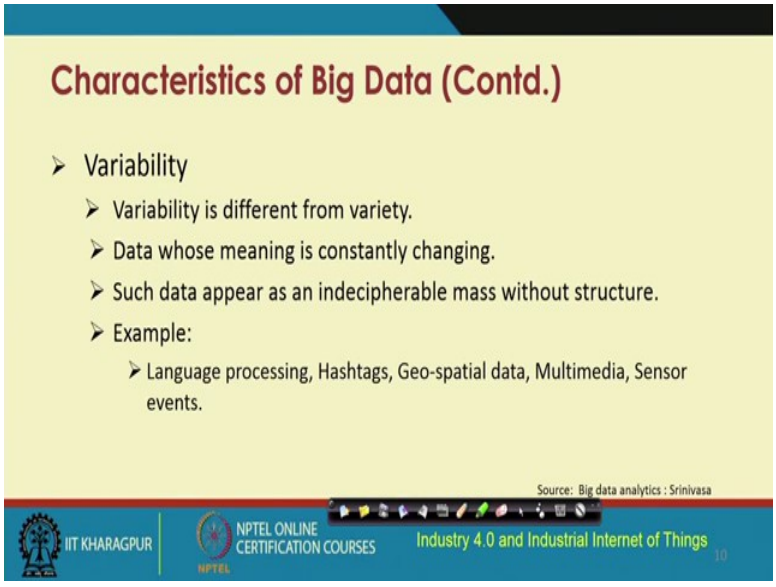
On the other hand, stock exchanges like New York stock exchange. So, they handle large volumes of data coming in high speeds. So, many exchanges, financial exchanges are carried on in the stock exchanges, so the speed is also very important of this kind of data.

(Refer Slide Time: 10:45)



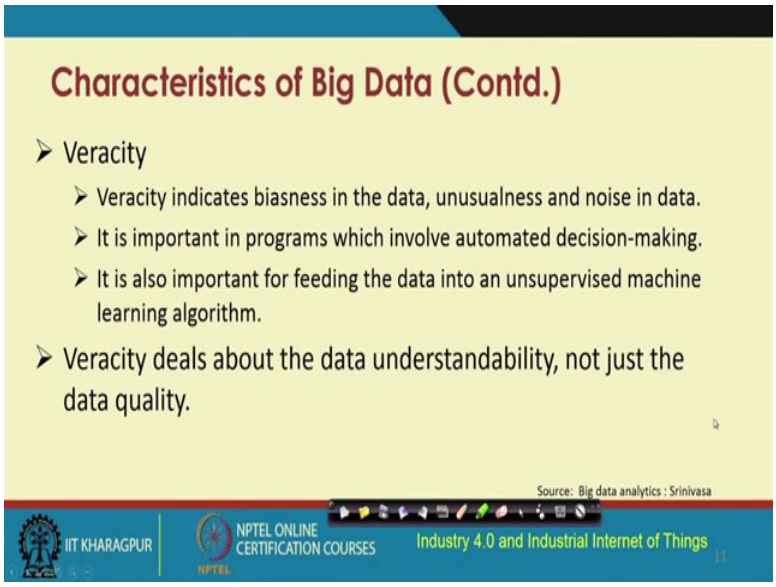
Variety, we are talking about different varieties of data--text, image, audio, video etcetera, etcetera all different types of data without any restriction about the data format.

(Refer Slide Time: 10:59)



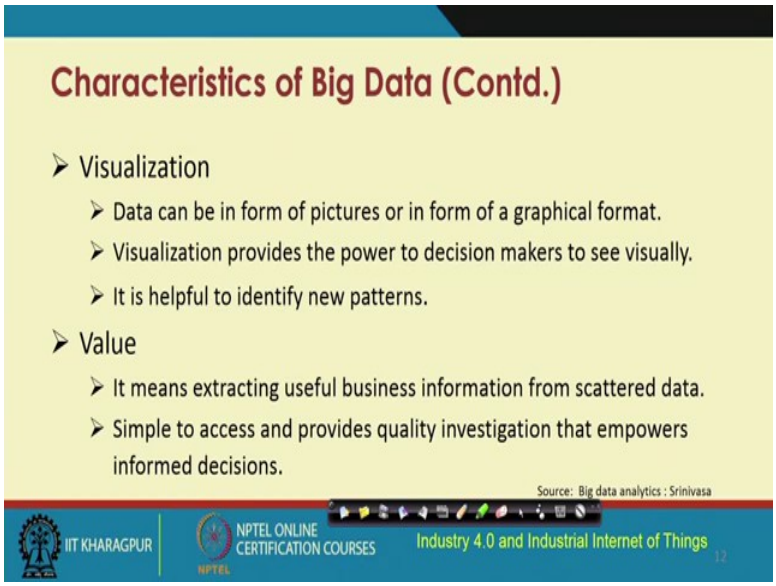
Variability, so here we are talking about it is different from variety. Here we are talking about the data whose meaning is constantly changing. So, example would be language processing, hash tags, geo-spatial data, multimedia data, sensor data, etcetera.

(Refer Slide Time: 11:19)



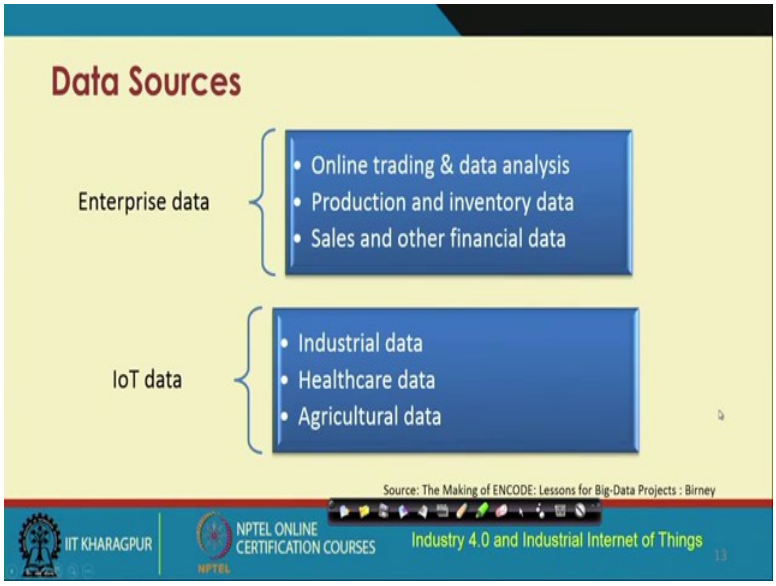
Veracity indicates the biasness in the data and it talks about the unusualness and noise in the data. So, it is important in programs, which involve automated decision-making. Veracity deals with the understandability of the data and not just the quality of the data, understanding the data is important in this thing; understanding the data is important in velocity.

(Refer Slide Time: 11:43)



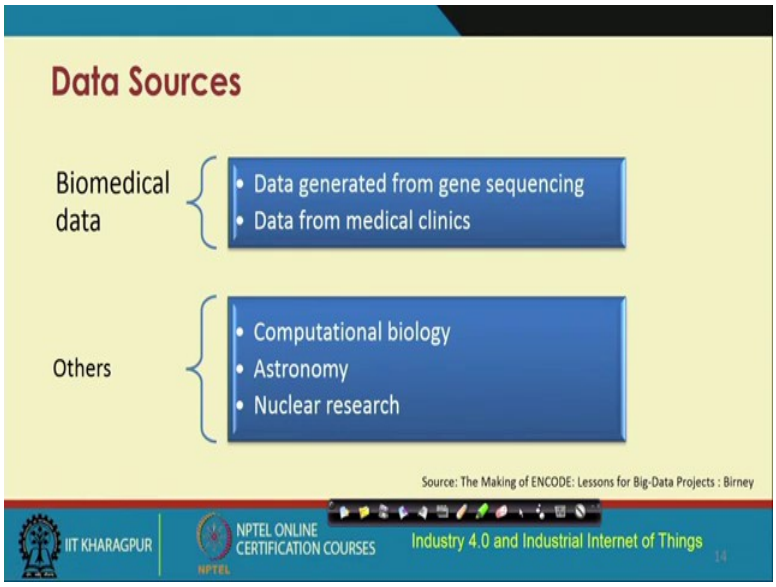
Visualization: visualization we are talking about in what form the data will have to be visualized, in graphical form, in textual form, what are the different patterns that can be visually identified. So, all these different character characteristics of this data the visualization of the data is important. And value, which is basically extracting some business information from the scattered data. So, this is basically the value attribute of big data.

(Refer Slide Time: 12:17)



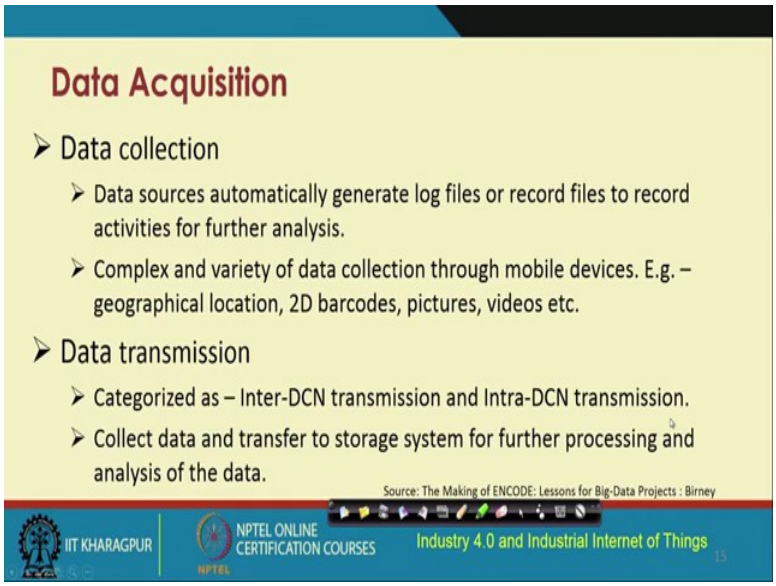
So, what are the sources of data? Enterprise data can be generated from online trading, data analysis, production inventory data, sales and different other financial data. These are some of the examples of enterprise data. IoT data like industrial data, healthcare data, agricultural data, and so on.

(Refer Slide Time: 12:41)



Biomedical data, data that are generated from gene sequencing, from medical clinics. And other data such as computational biology data, astronomy data, nuclear research data, these are the different sources of data. Astronomy, so some of these telescopes have been planted in certain parts of the world to look at the sky continuously, round the clock 365 days, a year, these are scanning the sky. And so as you can understand continuously in the, so much of data are coming from these telescopes. So, these kind of data will have to be handled; these are data which come in huge volumes. So, these are some of the sources of data.

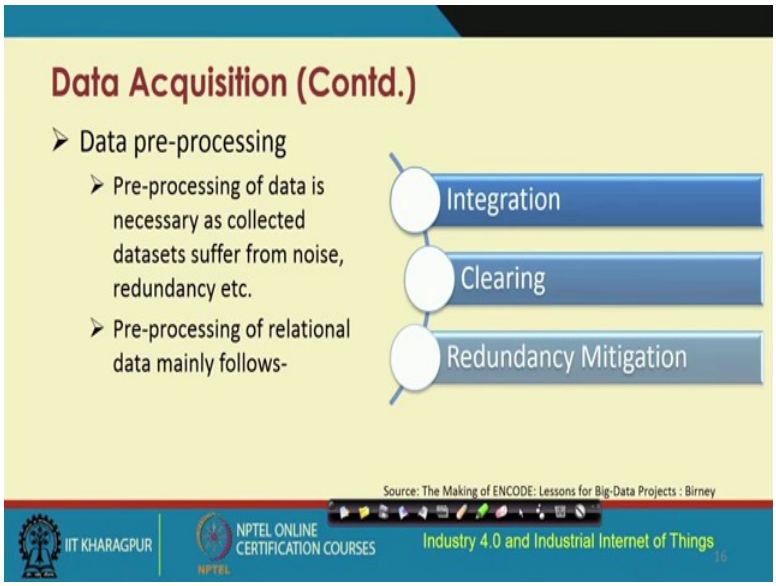
(Refer Slide Time: 13:25)



But in the context of industries similarly these machines, cyber physical systems machines continuously do by virtue of their continuous operation interaction and so on they are generating data, which typically is a combination of structured and unstructured or non-structured data. And these will have to be handled using these different analytical techniques for big data, data collection. So, you have to collect the data. So, the sources could be from log files, record files, from sensors, from different other sources like RFID devices and these could be coming from different sources.

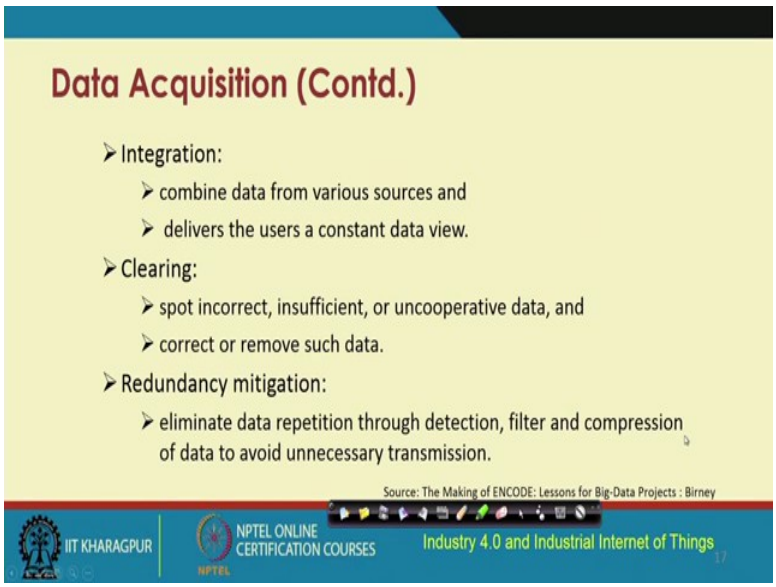
Data transmission is then important. We are talking about data inter DCN transmission and intra DCN transmission. What is this DCN? DCN is basically data center network. Within a data center network of data center; that means, different servers interconnected with each other in a data center within that the transmission of the data and then from one data center to another data center, inter-data center network transmission, that is another type of transmission.

(Refer Slide Time: 14:45)



Then the data will have to be processed. But before that it has to be pre-processed. The data will have to be pre-processed to remove noise redundancy, inconsistency. And this pre-processing of relational data mainly follows integration, clearing, and redundancy mitigation.

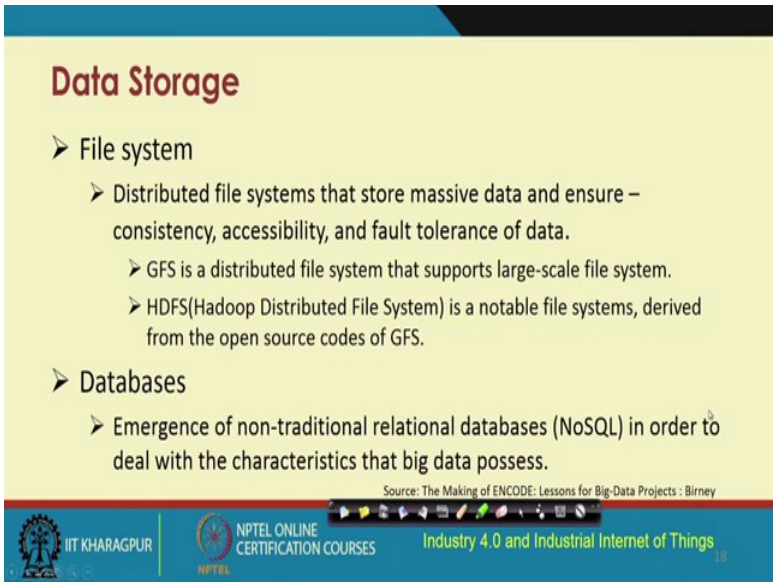
(Refer Slide Time: 15:09)



Integration is basically combining the data from various sources and delivering the users a constant data value. Clearing is spot, if there is something that is spotted to be incorrect, insufficient or uncooperative, then that kind of data will have to be spotted,

corrected or they have to be removed. Redundancy mitigation--if there is any redundancy in the data there has to be filters in place which will in remove the repetition of such kind of unnecessary data from being transmitted, because the bandwidth is also very limited. So, you do not want to transmit, redundant data unnecessarily.

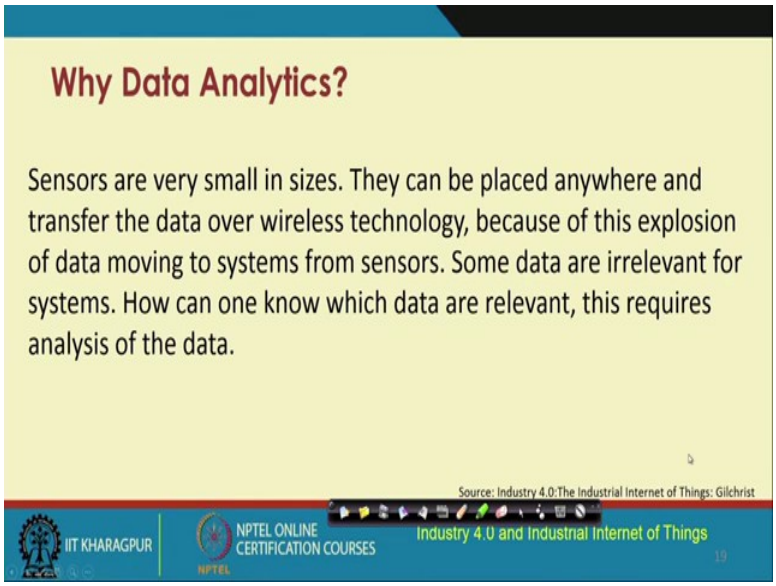
(Refer Slide Time: 15:45)



Data storage is important here we are talking about how the file systems would be for storing the big data. So, GFS is a distributed file system that helps in supporting in the storing, storage of large scale files. Another one is HDFS, HDFS is basically something that is followed in the Hadoop technology, Hadoop system, which is quite popular for use with storage of data, in the big data context. So, HDFS file system is a notable file system derived from the open source course of GFS. So, GFS and HDFS our file systems that are quite popular for use in different contexts of big data.

Databases, databases of different types non-traditional database wearing non-traditional relational database, like NoSQL language is there; like SQL, SQL is for the traditional relational databases use with querying this traditional relational databases and for non-relational databases. NoSQL query language could be used.

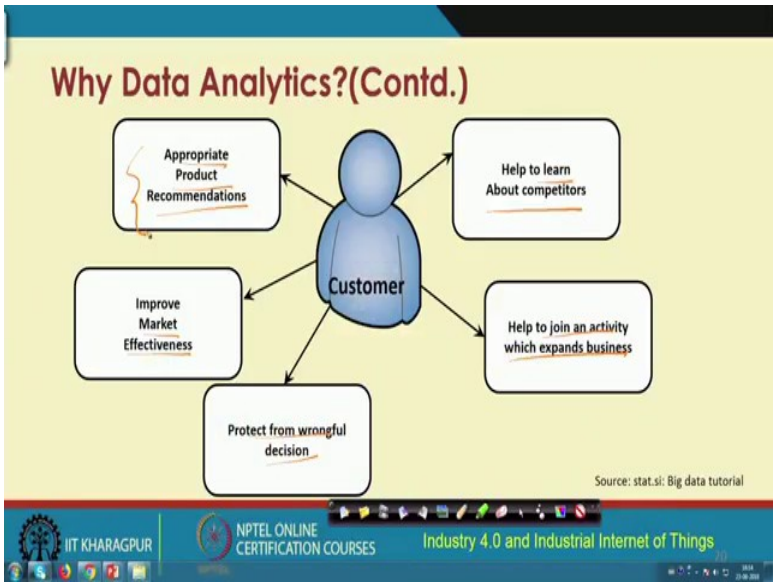
(Refer Slide Time: 17:05)



So, why do you need data analytics? So, in the context of IoT, IIoT, Industry 4.0 in general sensors that are used are very small in size, and they can be placed anywhere and these data from the sensors can be transferred over some networks, typically, wireless in nature. Because of this expression of data moving to systems from systems would be using these different sensors and the data will be moving from one sensor to another.

So, some data are irrelevant for systems. So, how can one know which data are relevant and which are not, so that is why this data will have to be analyzed. And also, there is one more point to it that once you have analyzed, the data you can perform certain actuation on the system, on the physical system, on the physical environment. So, those equations can be performed, based on the results of analysis of the data that is obtained.

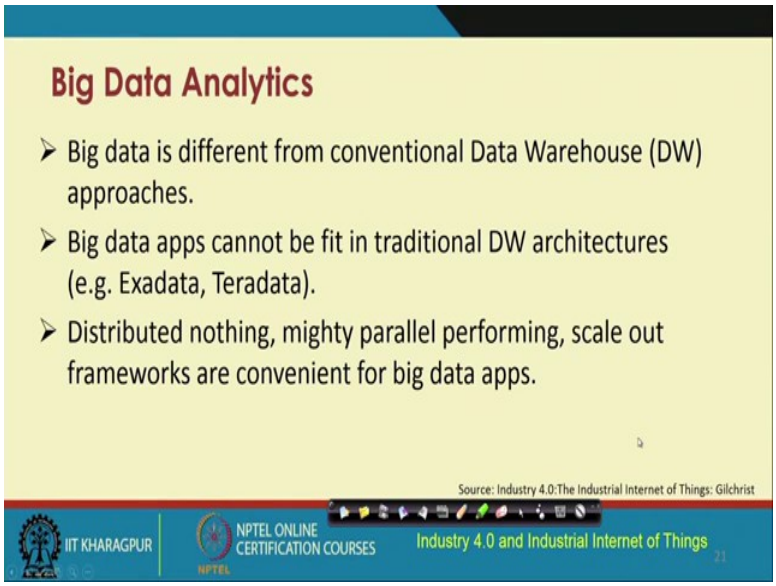
(Refer Slide Time: 17:59)



So, analytics is very important. For a customer, analytics will help the customer to learn about competitors, learn about competitors, help the customer to join and activity, which expands business. For the customer to protect from wrongful decisions, improve market effectiveness, and picking appropriate product recommendations, making appropriate product recommendations.

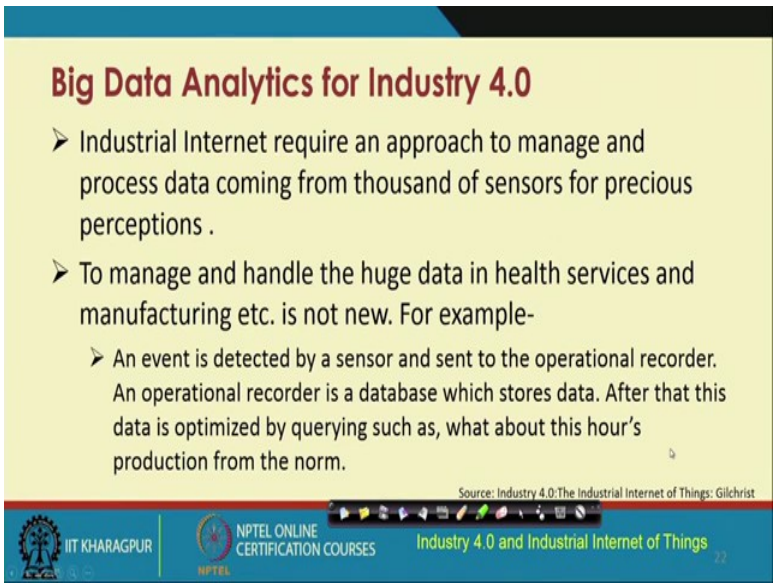
For example, if you are getting into Amazon or eBay or something you must have observed or even like Flipkart, etcetera they use recommender systems a lot. So, recommender systems basically what they do these are also AI based systems. They analyze the data using certain techniques maybe, machine learning or whatever and they will be making recommendations or suggestions to the user about what they could do next. So, these are the recommender systems.

(Refer Slide Time: 19:05)



Big data analytics talks about a different way of handling data from the conventional way, data are handled in databases and data warehouses. Big data apps cannot be fit in traditional applications, cannot be fit big data applications cannot be fit in traditional data warehouse architectures because here we are talking about large volumes of data, Exadata, Teradata.

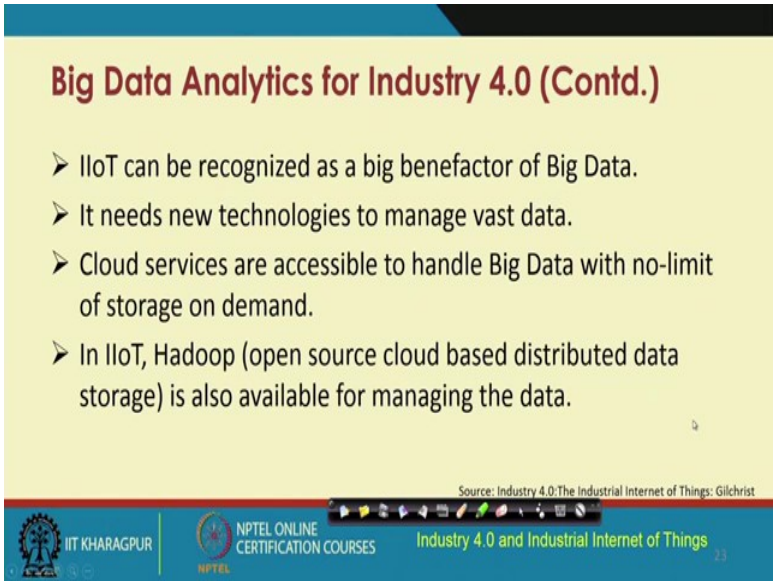
(Refer Slide Time: 19:33)



For Industry 4.0, we are talking about use of internet, industrial internet, which require an approach to manage and process data coming from thousands of sensors for

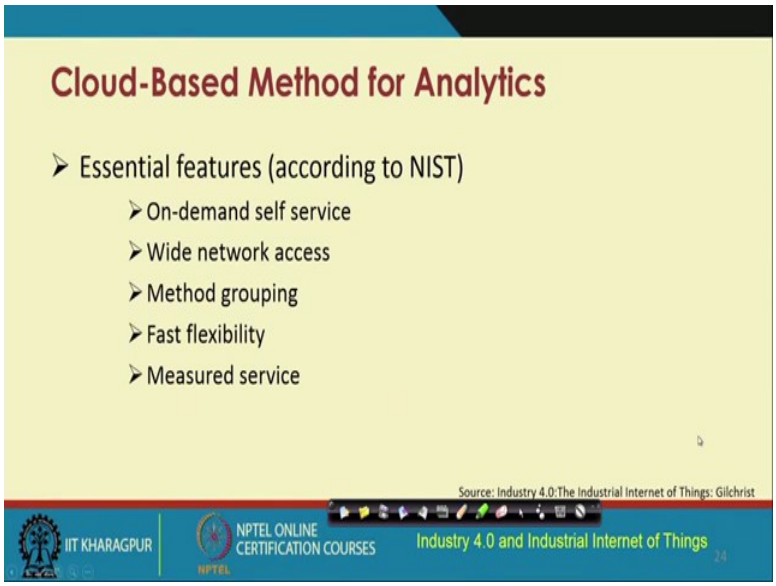
perceptions. To manage and handle this huge data in health services manufacturing other industries agriculture inclusive, is not new. For example, an event is detected by a sensor and sent to the operational recorder and an operational recorder in it is a database, which stores the data. And after that this data is optimized by querying the system about something like, what about this hour's production from the norm from the normal. So, this kind of queries could be made from the data. So, that is where this big data analytics could be used in the context of Industry 4.0.

(Refer Slide Time: 20:25)



So, IIoT can be recognized as a big benefactor or big data. It needs new technologies to manage first data. Cloud can come to the help, because cloud can help in storing the data and also processing the data, because there is as such virtually there is no limit on the storage because if everything the storage is also obtained on demand. So, in IIoT for helping in the management of the data technologies such as Hadoop, which is an open source cloud based distributed data storage platform, cloud can be used for the analysis of this kind of data, which is big data. So, Hadoop is quite popular in the context of big data analytics.

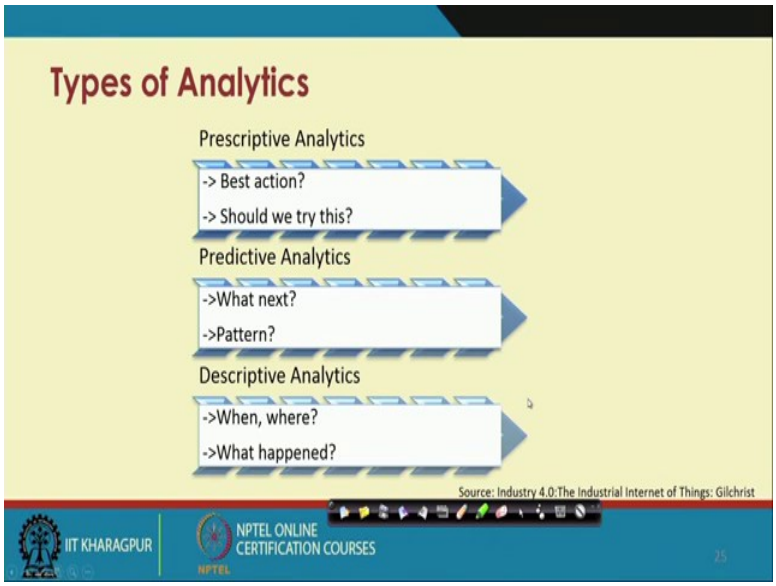
(Refer Slide Time: 21:09)



So, cloud-based methods for analytics would be; so first of all, cloud could be used for handling data big data, more specifically. So, there are different features of these on demand self-service, cloud basically helps in getting through a self-service kind of mechanism computational resources on-demand based on what the user requires. So, the scalability is also much better.

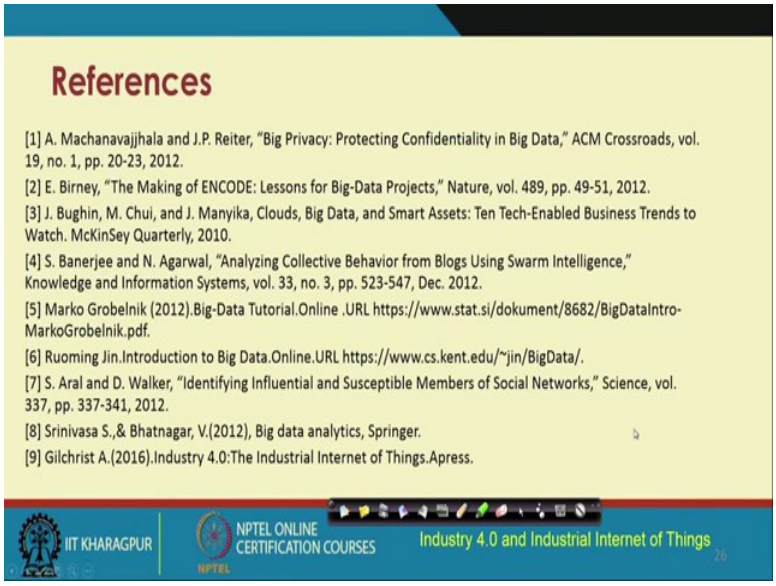
So, this is basically in a cloud offers wide network access, it offers the capability of grouping different methods, faster flexibility, and offering measured service. Measured service means based on the units of usage of cloud, the user is going to be charged accordingly.

(Refer Slide Time: 22:05)



So, there are different types of analytics--prescriptive analytics, predictive analytics, and descriptive analytics. And prescriptive analytics talks about what is the best action, should we try this and so on. Predictive analytics talks about what is next, what is the pattern that is there. And descriptive analytics is when, where, etcetera what happened, these are all descriptive analytics. Predictive, prescriptive, and descriptive analytics are different forms of analytics.

(Refer Slide Time: 22:37)



So, these are some of the references on big data and analytics. So, if you are interested you could go through any of these references, but there are huge number of references on big data and analytics at present and you could go even beyond these lists of references that are there for you.

So, with this we come to an end of analytics. Remember one thing that artificial intelligence, analytics etcetera these are not fields which are separate and are different from Industry 4.0 and IIoT. Without these enabling technologies, like the other technologies, you cannot basically survive to build IoT systems, IIoT systems or to achieve the vision of Industry 4.0. So, these are very important technologies for an analysis of the data, which could be used for analysis of the data, in the big data context in Industry 4.0.

Thank you.