

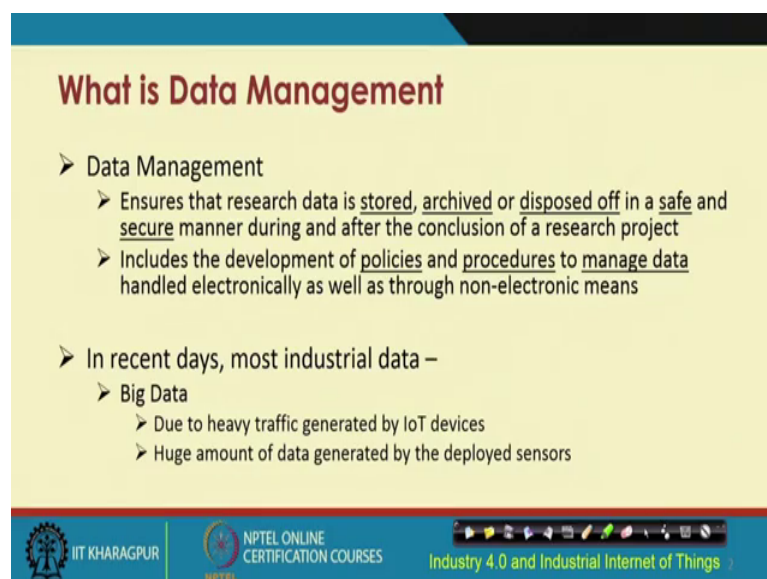
Introduction to Industry 4.0 and Industrial Internet of Things
Prof. Sudip Misra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 43
IIoT Analytics and Data Management: Data Management with Hadoop

In this particular lecture we will focus on data management. So, what is data management overall not just in the context of IIoT and industry 4.0. So, what is data management overall is first what we are going to understand and thereafter data management with hadoop is what we are going to understand. So, basically this is this particular module as you have noticed focuses on analytics right analytics with the help of different different technologies and so on.

So, we are focusing on cloud computing different different you know analytic techniques including machine learning etcetera and also we have understood machine learning AI etcetera we have understood. So, how do you manage this data before even you can do all this analytics rights, how we manage this data. So, data management overall is what we are going to understand and data management in the context of IIoT is what are going to understand next and thereafter we are going to switch our gears and we will understand what is data management in the context of use of hadoop, that is what we are going to focus on thereafter.

(Refer Slide Time: 01:26)



What is Data Management

- Data Management
 - Ensures that research data is stored, archived or disposed off in a safe and secure manner during and after the conclusion of a research project
 - Includes the development of policies and procedures to manage data handled electronically as well as through non-electronic means
- In recent days, most industrial data –
 - Big Data
 - Due to heavy traffic generated by IoT devices
 - Huge amount of data generated by the deployed sensors

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Internet of Things

So, data management basically when we are talking about as the name suggest as the term suggest focuses on different attributes such as storing the data, archiving the data, then once the date has been used and is no longer required disposing of the data, securing the data, keeping the data in a safe manner secured manner at the end of the project completion requirements etcetera, these are some of the different requirements of data management.

This basically also includes development of policies procedures to do all these above things that I mentioned just now either you are storing the data in electronic or non electronic manner and so on. So, these are these different aspects of data management and as I told you at the outset that, this is something that what is data management in general and not necessarily in the context of IIoT or data management with hadoop. Those are the things that we are going to discuss next, but this is the overall per view of data management so the different attributes of data management and so on.

So, in recent years as we have seen before we had a lecture which was completely dedicated to big data so we have understood what big data is, but what we have understood is also that big data at present is an important technology because big data is what most of the data, that we are dealing with at present particularly in the industrial context, particularly in the IIoT and IoT contexts, the nature of the data that we deal with are typically unstructured and having the characteristics of the big data.

So, if you recall that in the context of big data in the lecture on big data we earlier discussed about the different characteristics of big data, we talked about the definitions of big data starting from 3 V's, through 5 V's, through 7 V's and so on. Data having high volume coming in high velocity, having high variety, variability, velocity and so on and so forth, so many different attributes all these different V's where used to characterize the big data and big data is unstructured typically unstructured data. So, how do you deal with all these unstructured data is what big data concerns.

So, managing this kind of data, managing this kind of data and managing this kind of data in this particular lecture we are going to focus on to get an overview of use of hadoop to manage this kind of data right. Of course, you need to have this cloud enablement and cloud enablement is, what we have already discussed earlier and we

have we are also going to talk about data centre data centre networks and so on in another lecture.

So, putting everything together is how you are going to manage the data and thereafter use the data for analysis, processing, analysis and deriving intelligence or meaning out of this data. So, typically industrial machinery having fitted with large number of different sensors, actuators and so on, all these different IoT devices. In most of these IIoTs settings you are going to have this industrial machinery with these different sensors actuators and so on basically generating large volumes of data having all this big data characteristics. So, this kind of data will have to be dealt with adequately.

(Refer Slide Time: 05:03)

Data Management: Technologies

- Cloud computing
 - Essential characteristics according to NIST
 - On-demand self service
 - Broad network access
 - Resource pooling
 - Rapid elasticity
 - Measured service
 - Basic service models provided by cloud computing
 - Infrastructure-as-a-Service (IaaS)
 - Platform-as-a-Service (PaaS)
 - Software-as-a-Service (SaaS)

The slide also features logos for IIT KHARAGPUR, NPTEL ONLINE CERTIFICATION COURSES, and the text 'Industry 4.0 and Industrial Inte'.

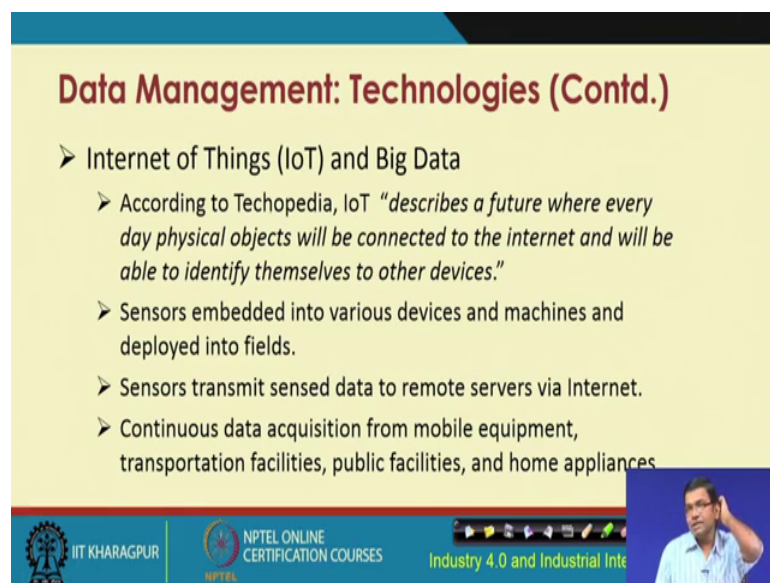
So, cloud will help you with the help of it is different models such as the infrastructure as a service, platform in service and software as a service that we have understood in the cloud computing lecture earlier.

Cloud with the help of all of these different cloud models and architectures will help you to offer on demand self service, on demand you know service of computation, computation services on demand will be made available to the end users. Broad network access, resource pooling, rapid elasticity, measured service, cloud is a measured service so depending on the units of usage of this computation resources one will first of all one will be able to measure the units of use and then accordingly one is going to be built. So,

billing is going to happen depending on the units of measured service that are going to be used.

So, these are some of these different characteristics of cloud computing which makes it popular in the context of data management. So, cloud models all these cloud models IaaS, PaaS, SaaS along with it is different characteristics that we just spoke about and to be discussed at in depth in or the cloud computing lectures so, together we are going to use it for the data management.

(Refer Slide Time: 06:28)



Data Management: Technologies (Contd.)

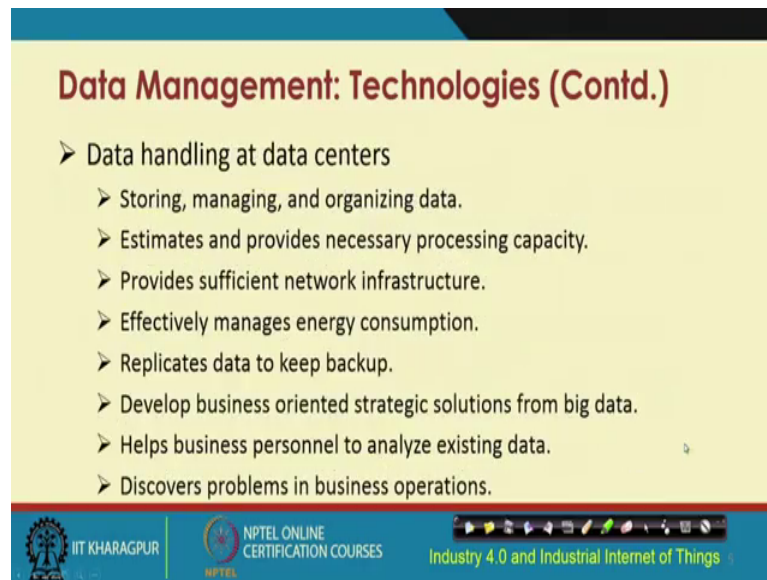
- Internet of Things (IoT) and Big Data
 - According to Techopedia, IoT *"describes a future where every day physical objects will be connected to the internet and will be able to identify themselves to other devices."*
 - Sensors embedded into various devices and machines and deployed into fields.
 - Sensors transmit sensed data to remote servers via Internet.
 - Continuous data acquisition from mobile equipment, transportation facilities, public facilities, and home appliances

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Inte

So, just as a recap in the context of IoT big data basically describes use of or management of this kind of data that are coming from the day to day physical objects being used for serving day to day activities and so on, which are interconnected together typically through the internet and this devices the IoT devices, the sensors etcetera which are have their own different identity and which generate lot of these different types of data.

These sensors fitted to this different machinery, industrial machinery devices etcetera which are working in the field and so on. And each of all of these basically connected to the through the internet work or the internet, generating lot of data, how do you handle this data, IoT data, IIoT data, how do you handle it is, what we are talking about in the context of IIoT data management.

(Refer Slide Time: 07:27)



Data Management: Technologies (Contd.)

- Data handling at data centers
 - Storing, managing, and organizing data.
 - Estimates and provides necessary processing capacity.
 - Provides sufficient network infrastructure.
 - Effectively manages energy consumption.
 - Replicates data to keep backup.
 - Develop business oriented strategic solutions from big data.
 - Helps business personnel to analyze existing data.
 - Discovers problems in business operations.

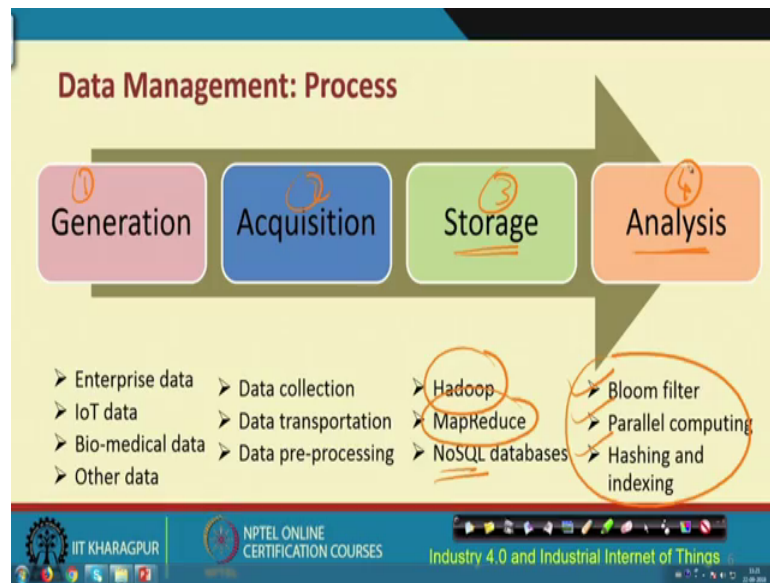
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Internet of Things

So, data centre we are going to talk about in depth in another lecture, but handling this kind of big data at the data centres is very important, IIoT data IoT data having different characteristics that we spoke about just now will have to be handled properly handling with respect to the storage of the data, management of the data, organisation of the data, estimating and providing necessary processing capacity this will also have to be done.

So, data centres are useful in all of these things that you see listed in front of you, storage management organisation of the data estimating and providing necessary processing capacity, providing sufficient network infrastructure, effectively managing energy consumption, repeating the data to keep the backup, developing business oriented strategic solutions from this kind of data the big data, helping the business personal to analyse the existing data and discovering problems in the business operations.

These are some of the data handling aspects that are very important in the context of IIoT and data centres which we are going to talk about in another lecture is going to help a lot in catering to these data handling and data management requirements that are there in the context of IIoT.

(Refer Slide Time: 08:50)



So, this is the overall process of data management it starts with the generation of the data, then acquisition of the data; that means, getting the data, storing the data and there after analysing the data. So, for each of these 4 steps in the data management process the corresponding attributes the corresponding characteristics are mentioned.

The only thing that I would like to highlight are 2 things, number 1 is to storage, storage is very crucial over here for storage you are going to use the cloud, but not just the cloud itself, but we are going to use this hadoop and enabled with MapReduce etcetera which are also going to help in performing NoSQL queries and so on. So, NoSQL is basically querying the databases you know this is a query language which can help you in querying databases which store unstructured data.

So, if you have structured data stored in relational table traditionally so, there you can use your SQL, SQL language for querying those tables, but if you do not have the data in the form that can be stored in the form of tables then this NoSQL will help. So, NoSQL is useful for big data is useful for unstructured big data and so on and NoSQL works very well with hadoop, MapReduce etcetera.

So, this is basically your storage and queries and so on and the other thing that you like to highlight in this particular slide is basically the analysis, for analysis you have large number of all these computational techniques that are there, including use of bloom filters, parallel computing techniques, hashing indexing, machine learning, clustering,

then classification use of neural networks SVM all of these can help in different ways for this analysis. So, it starts with generation of the data, acquisition of the data, there after storage of the data and finally, the analysis of the data.

(Refer Slide Time: 11:08)

Data Sources

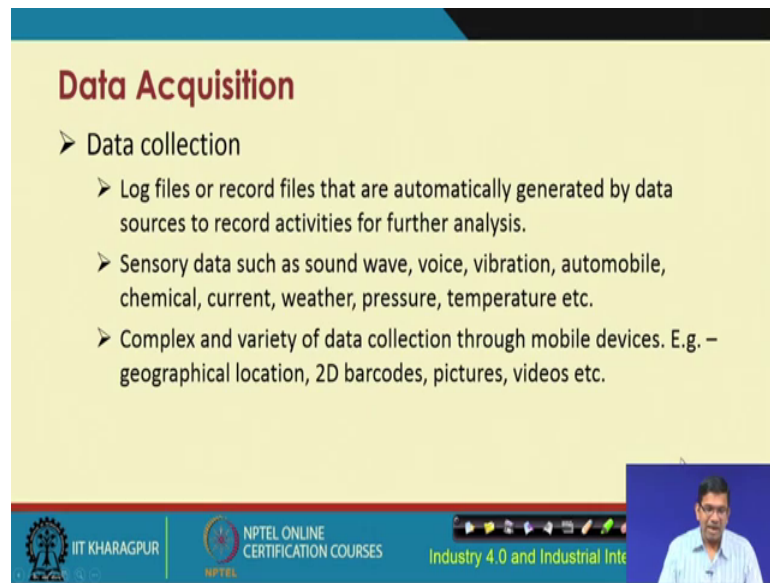
- Enterprise data
 - Online trading and analysis data.
 - Production and inventory data.
 - Sales and other financial data.
- IoT data
 - Data from industry, agriculture, traffic, transportation
 - Medical- care data,
 - Data from public departments, and families.
- Bio-medical data
 - Masses of data generated by gene sequencing.
 - Data from medical clinics and medical R&Ds.
- Other fields
 - Fields such as – computational biology, astronomy, nuclear research etc

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Internet of Things

So, data can be generated from different sources can be enterprise data, which talks about you know trades you know trading data, data coming from different machinery from different systems different parts of the systems, sales data, financial data productivity data, inventory data, all kinds of enterprise data. IoT data that means, all these machinery in these different industry, manufacturing industry, transportation, agriculture, etcetera the sensors and IoT device that are fitted they continuously basically generate lot of data so that in that IoT data that we are talking about in the context of industries.

So, then bio medical data generated from the medical clinics, medical R and Ds you know through different biotechnological you know different bio technological methods such as gene sequencing etcetera so all of these are data sources for this big data. Other fields also generate lot of data nuclear power plants, astronomical devices such as big big telescopes, which look into the sky continuously 24 7. So, those computational biological fields also generate lot of these kind of data which are unstructured and having this big data characteristics and so on these are the data that we have to be managed.

(Refer Slide Time: 12:27)



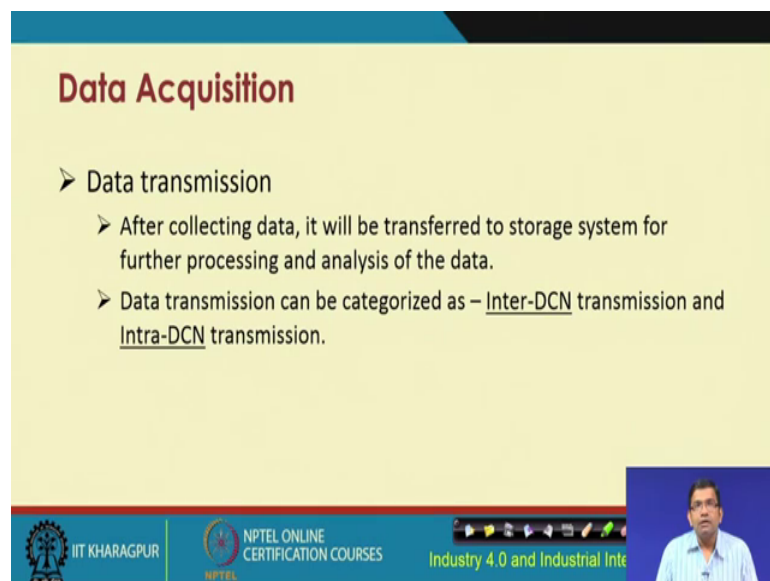
Data Acquisition

- Data collection
 - Log files or record files that are automatically generated by data sources to record activities for further analysis.
 - Sensory data such as sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature etc.
 - Complex and variety of data collection through mobile devices. E.g. – geographical location, 2D barcodes, pictures, videos etc.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Inte

Acquisition of the data in terms of collection to you know login the data and recording the data automatically generated from this data sources and how do you basically collect and log this data that is important. Sensory data such as sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature, etcetera, etcetera, etcetera these are all these different types of sensory data which will have to be acquired. And complex and variety of data collection is possible through the use of different mobile devices such as geographical locations, 2 D barcoding, pictures, videos etcetera.

(Refer Slide Time: 13:07)



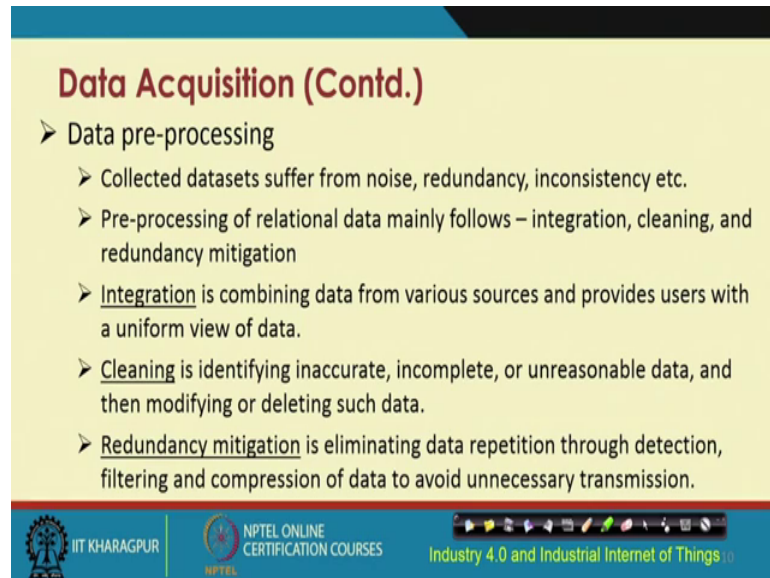
Data Acquisition

- Data transmission
 - After collecting data, it will be transferred to storage system for further processing and analysis of the data.
 - Data transmission can be categorized as – Inter-DCN transmission and Intra-DCN transmission.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Inte

Data transmission through the network through the system and we will see in another lecture, how the data can be sent through the network through different interconnected data centres which will have inter data centre traffic data centre traffic data traffic and also intra data centre data centre data traffic. So, transmission of the data within and outside the data centre networks is what is very important.

(Refer Slide Time: 13:38)



Data Acquisition (Contd.)

- Data pre-processing
 - Collected datasets suffer from noise, redundancy, inconsistency etc.
 - Pre-processing of relational data mainly follows – integration, cleaning, and redundancy mitigation
 - Integration is combining data from various sources and provides users with a uniform view of data.
 - Cleaning is identifying inaccurate, incomplete, or unreasonable data, and then modifying or deleting such data.
 - Redundancy mitigation is eliminating data repetition through detection, filtering and compression of data to avoid unnecessary transmission.

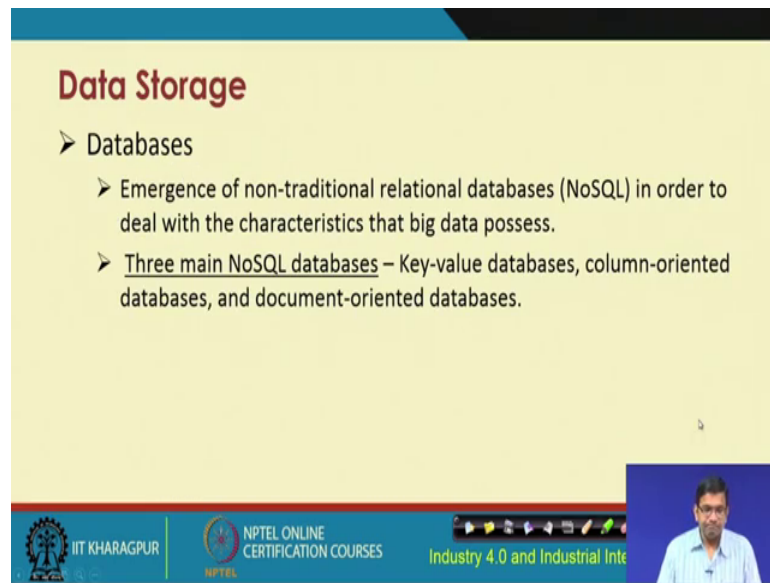
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Internet of Things

Pre processing of the data connecting, the data removing noise redundancy that might be existing inconsistencies, that might be found out. So, these are some of these pre processing tasks that are relevant. It is also very important to pre process the data for serving this different needs integration cleaning and redundancy mitigation.

Integration talks about combination of data from arriving from different sources and providing users with unified integrated view of the data, even though the data is originated and is coming from different different channels. Cleaning of the data to remove all these incompleteness, in inaccuracies, incorrectness that might be there, if there is any unreasonable data that might be there modifying that particular data or deleting that data.

And also mitigating the redundancies such as eliminating data, repetition through detection, filtering, completion of data to avoid unnecessary transmission of data through this limited resource limited or resource constant networks is what is done as part of data pre processing.

(Refer Slide Time: 14:53)



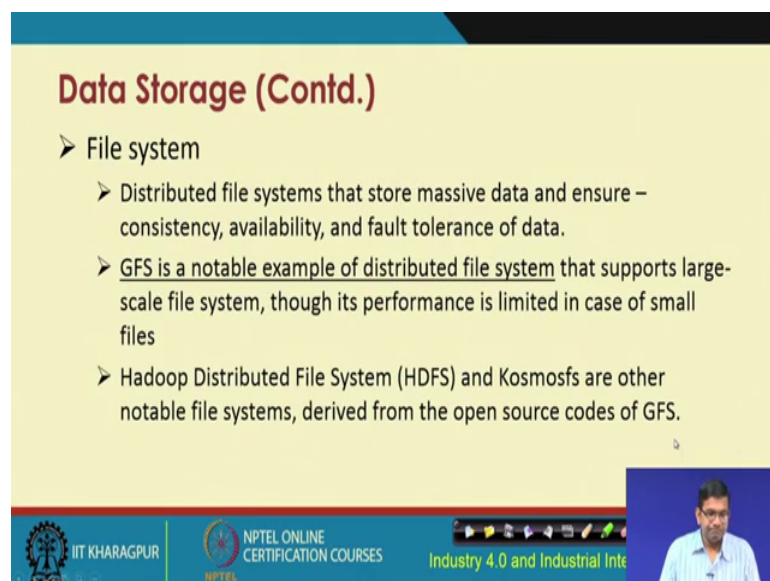
Data Storage

- Databases
 - Emergence of non-traditional relational databases (NoSQL) in order to deal with the characteristics that big data possess.
 - Three main NoSQL databases – Key-value databases, column-oriented databases, and document-oriented databases.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Inte

Storing the data in different databases traditional or non traditional data bases such as the NoSQL databases that I mentioned earlier is very important. NoSQL databases, we will support different thing such as key value database, column oriented database and document oriented database and their corresponding techniques of how to handle the data in each of these different databases is what is of concern in the context of data management.

(Refer Slide Time: 15:23)




Data Storage (Contd.)

- File system
 - Distributed file systems that store massive data and ensure – consistency, availability, and fault tolerance of data.
 - GFS is a notable example of distributed file system that supports large-scale file system, though its performance is limited in case of small files
 - Hadoop Distributed File System (HDFS) and Kosmosfs are other notable file systems, derived from the open source codes of GFS.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Inte

Data storage in the files GFS, Google file system is a notable example of distributed file system storing large scale file system data store, data store in different file systems HDFS in this is an another example, hadoop distributed file system, then Kosmosfs is another example these are different examples of the file systems used for data storage.

(Refer Slide Time: 15:48)



Industrial Data Management

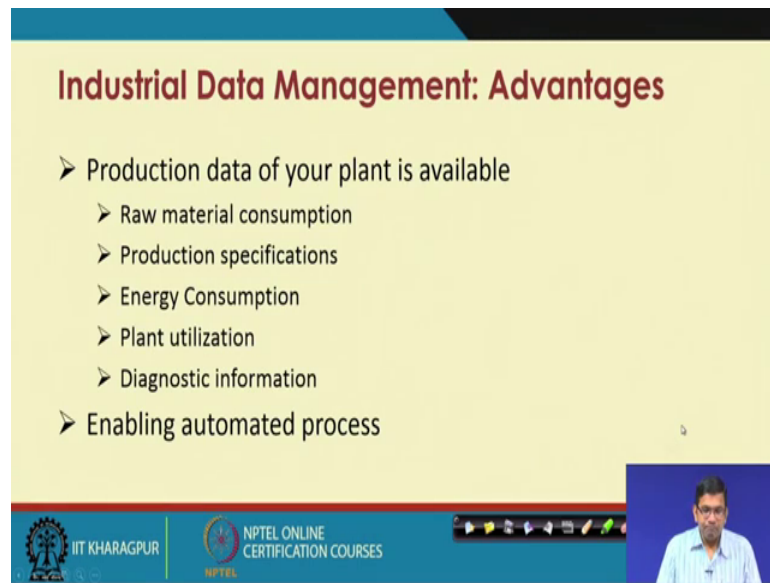
- Incorporates data generated in
 - Manufacturing plants
 - Processing plants
- Management done in entire value chain
- Data availability is ensured
- Enables decision making process easier

The slide is part of an NPTEL presentation. It features the IIT Kharagpur logo and the text 'NPTEL ONLINE CERTIFICATION COURSES' at the bottom. A small video inset in the bottom right corner shows a man in a light blue shirt speaking.

So, industrial data is what is concerns IIoT industrial data managing such kind of data using all these different techniques that I just mentioned will have to be done. Incorporating industrial data management basically will incorporate data that is generated from different processing plant manufacturing plants and so on and the management is done in the entire value chain. So, you know basically what it means is that the data industrial data will have to be handled properly in order to deliver value to the end users properly and this value chain is very value chain consideration is very important.

Availability of the data has to be ensured in industrial data scenarios. So, availability of the data in order to derive intelligence later on because if the data that is continuously coming etcetera etcetera is not handled adequately, then it is meaningless basically it is meaningless to derive any intelligence if the data is not available properly then you cannot do any further intelligence on it. So, this is basically will have to will also help the higher availability of the data will also help in enable proper decision making whenever it is required.

(Refer Slide Time: 17:12)



Industrial Data Management: Advantages

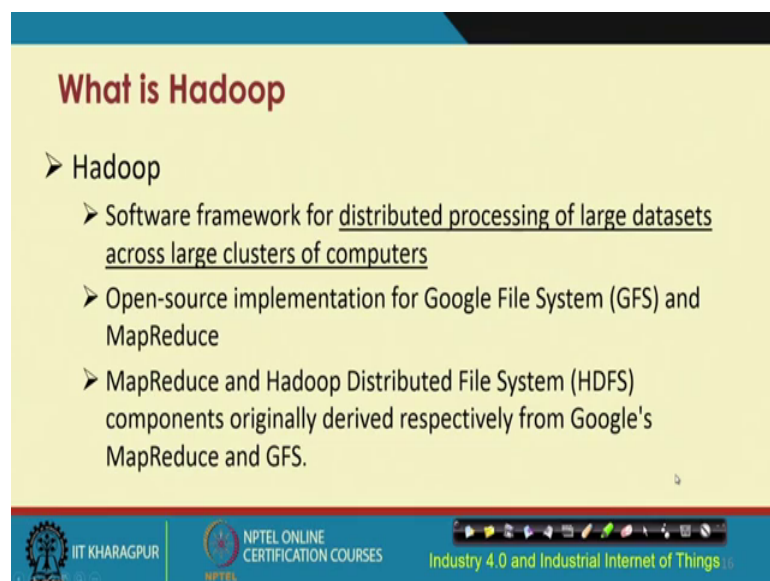
- Production data of your plant is available
 - Raw material consumption
 - Production specifications
 - Energy Consumption
 - Plant utilization
 - Diagnostic information
- Enabling automated process

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

Industry 4.0 and Industrial Internet of Things

The advantages of industrial data management are that production data of a particular manufacturing plant is made available through such kind of activities such as raw material consumption and production specifications, energy consumption, plant utilisation, diagnostic information and so on and so forth and then for industrial management we need to have an automated process implemented which will do all these data management activities autonomously.

(Refer Slide Time: 17:47)



What is Hadoop

- Hadoop
 - Software framework for distributed processing of large datasets across large clusters of computers
 - Open-source implementation for Google File System (GFS) and MapReduce
 - MapReduce and Hadoop Distributed File System (HDFS) components originally derived respectively from Google's MapReduce and GFS.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

Industry 4.0 and Industrial Internet of Things

So, let us now take a specific example of how hadoop a popular technology could be used for data management in IIoT scenarios. So, what is hadoop? Hadoop is basically a very popular technology from apache, which gives a software framework for distributed processing of large data sets you have large data sets distributed processing of those data sets in a cluster of computers is what hadoop basically specifically gives you the framework for. So, open source implementation for GFS and MapReduce and MapReduce and HDFS components, these are all the different aspects of hadoop.

(Refer Slide Time: 18:28)

Building Blocks of Hadoop

- Hadoop Common
 - A module containing the utilities that support the other Hadoop components
- Hadoop Distributed File System (HDFS)
 - Provides reliable data storage and access across the nodes
 - Rapid data transfer among the nodes
 - Fault tolerant

The slide also features a navigation bar at the bottom with logos for IIT KHARAGPUR, NPTEL ONLINE CERTIFICATION COURSES, and Industry 4.0 and Industrial Inte, along with a small video inset of a presenter.

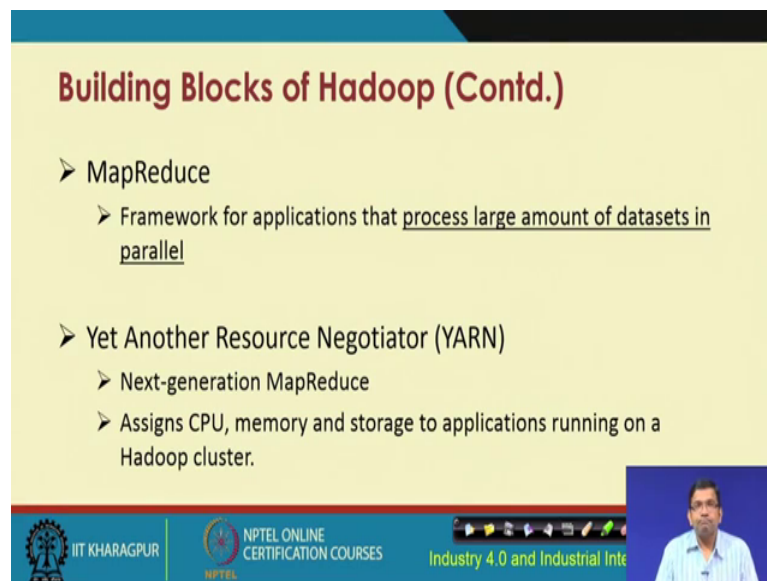
So, there are different building blocks of hadoop. So, number one is hadoop common which is basically a common component which is basically a module that contains the utilities that would support the other hadoop components like the once that I am going to mention next. So, it is basically a common is basically a module that will help other modules to work together in a connected fashion.

The HDFS is the central thing in hadoop HDFS hadoop distributed file system is the core of hadoop. It provides reliable data storage and access across different nodes in the system, the rapid data transfer among the nodes is going to be possible with the help of hadoop distributed file system HDFS and fault tolerant fault tolerant season attribute that is inherent to HDFS architecture.

This HDFS architecture as I will show you later on, HDFS architecture basically has different layers and in one of these layers basically what you have are something known

as the blocks which actually contains the data. So, what happens in HDFS is this blocks are replicated. So, basically the different data nodes which I will tell you later on, this data nodes basically have different blocks and each of these blocks is replicated across multiple data nodes. So, this basically ensures fault tolerant; if something goes wrong with one of these different blocks the other blocks the replicas of these different blocks are there in the other data nodes.

(Refer Slide Time: 20:07)



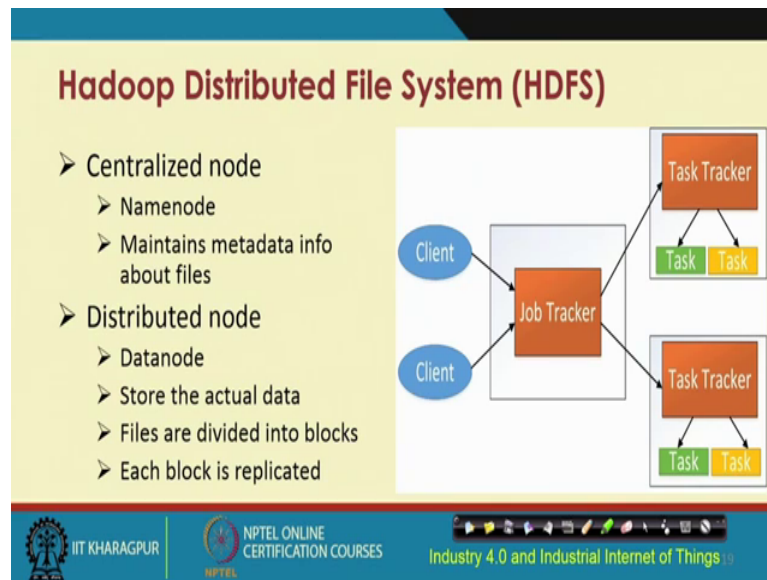
The slide is titled "Building Blocks of Hadoop (Contd.)" and lists two main components:

- MapReduce
 - Framework for applications that process large amount of datasets in parallel
- Yet Another Resource Negotiator (YARN)
 - Next-generation MapReduce
 - Assigns CPU, memory and storage to applications running on a Hadoop cluster.

The slide footer includes the IIT KHARAGPUR logo, NPTEL ONLINE CERTIFICATION COURSES, and the text "Industry 4.0 and Industrial Inte". A small video inset shows a man in a blue shirt.

So, the other building blocks of hadoop input the MapReduce which is like a framework for processing large number of large amount of data bases in parallel. This is a MapReduce, MapReduce is also core to hadoop, but MapReduce the algorithms that are there large number of different types of algorithms and their it is just a philosophy, it is a framework that hadoop basically also uses. So, YARN is basically the next generation MapReduce which assigns CPU, memory, storage to different applications running on the hadoop cluster of different computers.

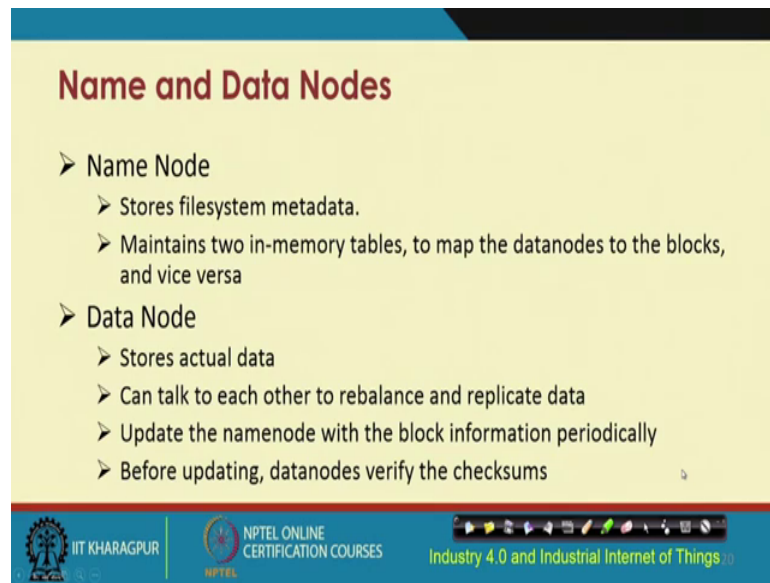
(Refer Slide Time: 20:45)



So, this how this HDFS which is central to hadoop looks like. So, in HDFS you have 2 types of nodes, one is the name node, the other one is the data node. The name node is centralized node so, this particular job tracker for example, is a name node right so, it is being done in the name node. So, this is a centralised node and then you have this different other task trackers for example, which are basically being executed in the data nodes.

So, this name node is basically the centralised node maintains all this meta data about the different files different meta data are basically maintained about the different files in the name node the centralised node, whereas the distributed node these task tracker the data nodes etcetera store the actual data and these files are divided in these into different blocks and each of these different blocks is replicated and this is what I was telling you earlier, this different blocks have their own the data, the data are replicated across this different data nodes and so on in this HDFS architecture.

(Refer Slide Time: 21:50)



Name and Data Nodes

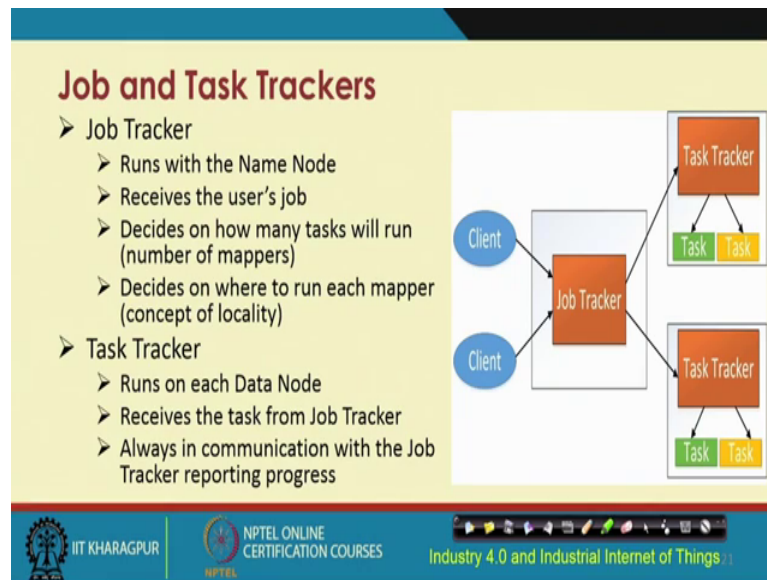
- Name Node
 - Stores filesystem metadata.
 - Maintains two in-memory tables, to map the datanodes to the blocks, and vice versa
- Data Node
 - Stores actual data
 - Can talk to each other to rebalance and replicate data
 - Update the namenode with the block information periodically
 - Before updating, datanodes verify the checksums

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Internet of Things

So, the name node basically is something which is the centralised entity which stores the metadata about the file system, it maintains two in memory tables to map the data nodes to the blocks and the vice versa. So, name nodes are connected to the data nodes which are actually the once where the storage of the actual data is done.

So, this data nodes also are interconnected with each other, they can imbalance, they can replicate the data across each other in this data node layer and they update the name node with the block information periodically. So, that the name node has a proper you know metadata and the updated metadata in place. So, before updating the data nodes would verify that check sums for ensuring the integrity of these data.

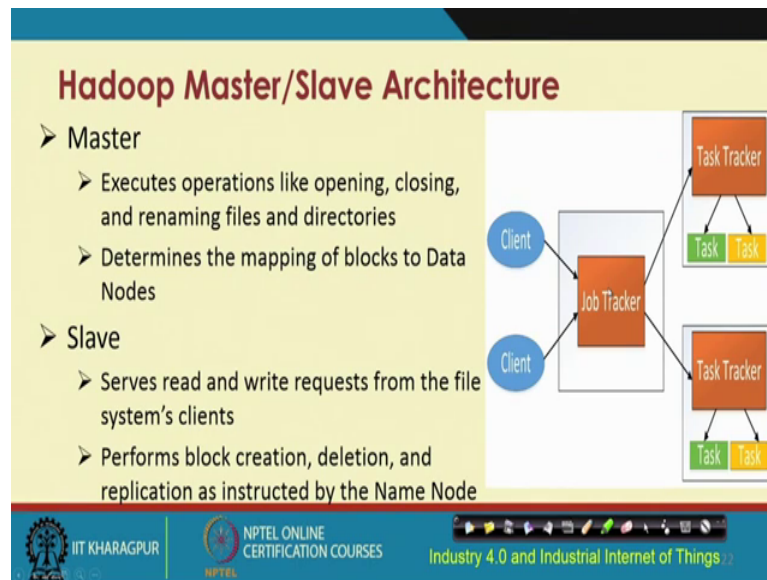
(Refer Slide Time: 22:41)



So, are these concepts of the job tracker and task tracker, the job tracker are basically running at the name nodes and the task trackers are running in the task in the data node right. So, these name nodes these job tracker will receive the users job will decide on how many tasks will run using, the concept of mapping and how many jobs and which jobs are going to run that mapping is going to be done and it is going to also decide on where to run in each of these different jobs.

In the task tracker on the other hand will run at each of these data nodes so, this is one data node, this is another data node. So, this task trackers are running on them, receiving the data receiving the tasks basically from the job tracker. So, the tasks this tasks that are going to be executed over here in this data nodes are going to be retrieved are going to be received from the job tracker and these are going to always be in communication with the job tracker, the task tracker is going to be in communication with the job tracker and these are always going to also report the progress to the job tracker.

(Refer Slide Time: 23:52)



So, basically it is a master slave architecture, where basically the master executes the operations like opening, closing, the renaming the files and dictionaries and determines the mapping of the blocks to the data nodes. Slaves are the once which read write request from the file systems clients and perform block creation, deletion, replication and so on.

So, this is basically this becomes your master node, the name node becomes a master node, these are like the slave nodes and slave nodes are basically continuously being they basically give the different tasks from the job tracker which will have to be executed at each of these different data nodes and basically these task tracker after completion of the task or in between also they would be updating the status to the job tracker. So, this is basically the master and this becomes your different slaves.

(Refer Slide Time: 24:49)

MongoDB in Data Management

- Database ↔ NoSQL
- Ensures
 - Performance
 - Scalability
 - Availability
- Creates a similar view of data across the enterprise

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

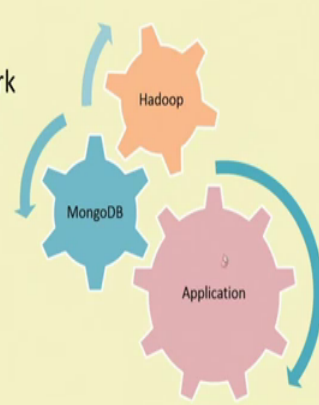
So, another thing that I would like to highlight over here is something known as the MmongoDB. So, MongoDB is a data management tool which basically supports data bases, different databases and particularly in this context of IIoT we are talking about NoSQL database is because of the unstructured data that we are typically experiencing

So, NoSQL database is supported by MongoDB, it ensures MongoDB is database basically it is a database it is a NoSQL database and it works in conjunction with hadoop. So, this particular database will ensure performance, improvement, performance ensuring performance overall good performance is ensured, scalability, availability and so on and creating a similar view of data across the enterprise.

(Refer Slide Time: 25:38)

MongoDB with Hadoop

- Hadoop adds a powerful framework to MongoDB for complex analytics
- Applications:
 - Batch Aggregation
 - Data Warehouse
 - ETL (Extract, Transform, Load) Data



The diagram illustrates the integration of Hadoop, MongoDB, and an Application. Three interlocking gears are shown: a blue gear labeled 'MongoDB', an orange gear labeled 'Hadoop', and a pink gear labeled 'Application'. Blue arrows indicate a clockwise flow of data or interaction between the components: from Hadoop to MongoDB, from MongoDB to Application, and from Application back to Hadoop.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 24

So, MongoDB basically works hand in hand with Hadoop. So, Hadoop basically adds as a powerful framework to MongoDB for complex analytics and different applications are supported by MongoDB such as batch aggregation, data warehousing and ETL data; that means, extract transform and loading of data, this is basically common term ETL in the context of databases so, ETL data handling. So, batch aggregation data warehousing and ETL data handling these are the different characteristics of or the different functionalities of the MongoDB along with Hadoop.

(Refer Slide Time: 26:19)

References - I

1. R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 603-630, Dec. 2012.
2. M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 707-734, Dec. 2012.
3. S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," *Science*, vol. 337, pp. 337-341, 2012.
4. A. Machanavajhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," *ACM Crossroads*, vol. 19, no. 1, pp. 20-23, 2012.
5. S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 523-547, Dec. 2012.
6. E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," *Nature*, vol. 489, pp. 49-51, 2012.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Internet of Things 25

So, with this we come to an end of this particular lecture we have this different difference is given to you, for you to benefit from and if you are interested you know there is a lot to do with data management and if you are interested particularly hadoop is very popular, MongoDB is very popular, MapReduce is also very popular, these are once which work hand in hand and this can help you in proper data management of big data that is being that is experience typically in the context of in the context of in the context of IIoT.

(Refer Slide Time: 26:57)

References - II

7. S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.
8. J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinsey Quarterly, 2010.
9. D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.
10. <http://hadoop.apache.org/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Industry 4.0 and Industrial Internet of Things 26

So, with this we come to an end and this is the assortment of different references that is listed for here.

Thank you.