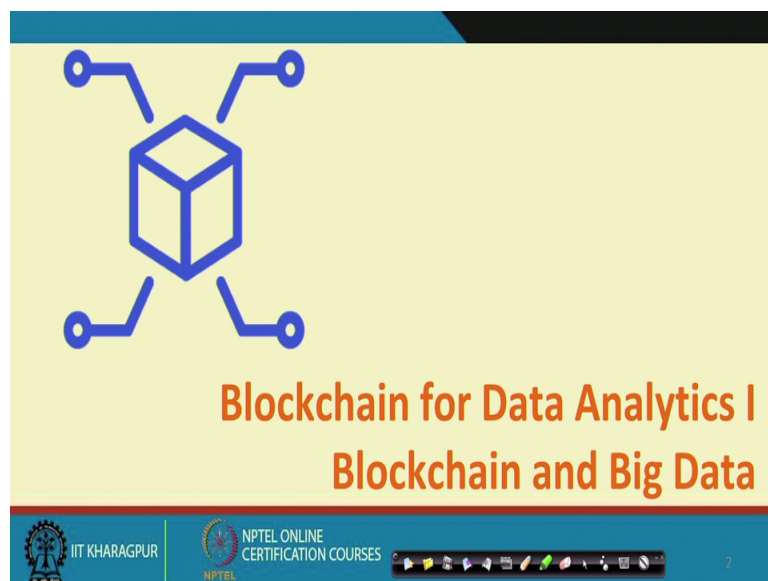


Blockchains Architecture, Design and Use Cases
Prof. Sandip Chakraborty
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 54
Blockchain for Data Analytics – I (Blockchain for Big Data)

Welcome to the course on Blockchains Architecture Design and Protocols. We are discussing about different use cases as well as the research topics on blockchain.

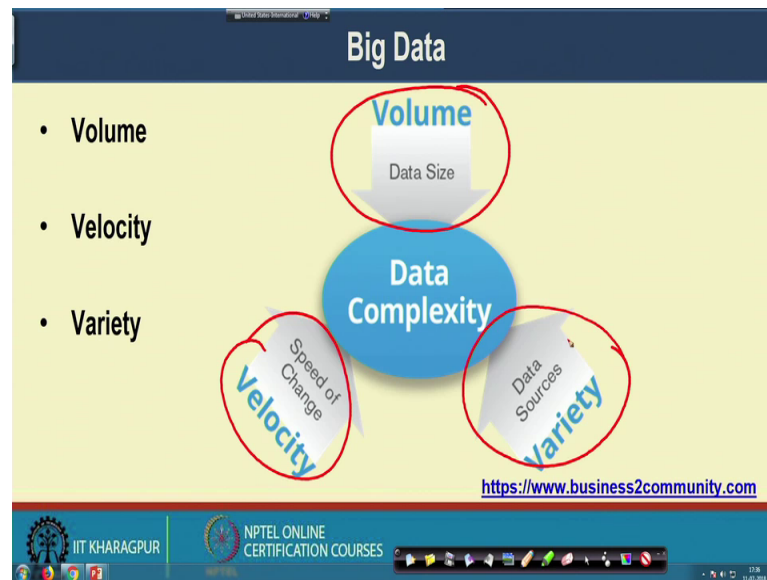
(Refer Slide Time: 00:33)



So, today we will touch upon another interesting use case of blockchain, and the research topics and aspects around that; the use of blockchain for data analytics. So, how can you utilise the concept of blockchain technology for effective data analytics under various environment with the help of blockchain. So, we will take 2 lectures on that. So, in this first lecture we will discuss about blockchain and its application over big data technology.

So, let us look into what is meant by big data.

(Refer Slide Time: 01:09)



So, big data, it is classified or it is identified by 3 v's, 3 parameters; volume, velocity and variety. So, the volume the first parameter it denotes the data size the amount of data that we have. So, if you just think about the typical data centre. Like, say YouTube data or Google data centre, the Amazon data centre; in any of this data centre the amount of data which is been generated it is in the order of trillion of bytes. So, the question comes that how will you manage this huge amount of data. Another interesting example is that social networking sites like Facebook. The just think of the amount of data which is generated from Facebook every day; it is even in the order of few 100 terabytes if I assume correctly.

So, that way managing this huge amount of data and running application on top of this huge amount of data is a challenging task. So, the question comes that with this huge amount of data how can you design effective mechanism or effective algorithm to process the data. Then the second perspective of a big data technology is the velocity of the data. Like the speed of change of data. Again think of the example of a social networking website like Facebook. So, the amount of change of data it is again in a larger scale. Say for example, if you just think about the data which is being generated from a personal profile, a personal Facebook profile, it again changes possibly in every day.

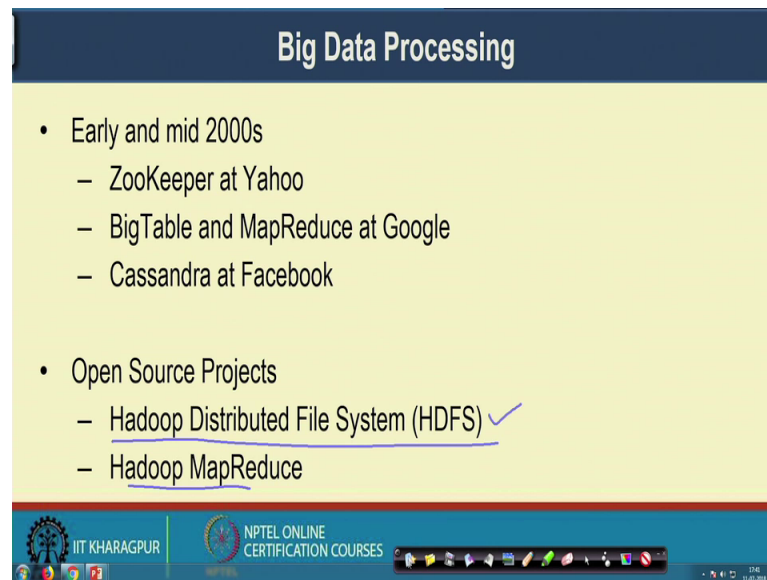
Another interesting example is the video streaming websites like YouTube. So, you can just think of that everyday people are uploading data in YouTube, and the way this data is changing and the information is changing that is the tremendous. So, if you want to do some analytics on top of this streaming media data, if you have done certain analytics today, again you have to run the analytics under next day and day after to find out a correct result of that or to predict the trend of change. So, this velocity of the data is the second important aspect. And the third important aspect which is again crucial for applying blockchain for big data application is the variety of data.

So, the variety of data means say just think of a kind of geospatial applications. Say, you want to throw a geospatial query. Now if you want to throw a geospatial query, you have to possibly require data from land registering, you possibly require the emerging data from ISRO you possibly require the metrological data, the weather data.

So, different varieties of data which are there all together, and your query need to be executed on top of that variety of data. So, this gives a challenge, this gives multiple challenges for developing an application over the big data platform. So, the question comes not only in terms of query processing or analysing the data, rather managing the data, sharing the data among multiple peers and then analysing the data.

How will you handle all these things in a effective way? So, we will first look into that what are our challenges for managing this kind of data in a traditional platform that we are using nowadays. So, we will identify certain problems from there, and then we will look into that how we can design a solution for those problem with the help of the blockchain technology.

(Refer Slide Time: 05:35)



The slide is titled "Big Data Processing" and is divided into two main sections. The first section, "Early and mid 2000s", lists three technologies: ZooKeeper at Yahoo, BigTable and MapReduce at Google, and Cassandra at Facebook. The second section, "Open Source Projects", lists two projects: Hadoop Distributed File System (HDFS) and Hadoop MapReduce. The slide also features logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES at the bottom.

- Early and mid 2000s
 - ZooKeeper at Yahoo
 - BigTable and MapReduce at Google
 - Cassandra at Facebook
- Open Source Projects
 - Hadoop Distributed File System (HDFS) ✓
 - Hadoop MapReduce

So, let us look into the traditional way of processing big data. So, this volume of data that started increase after internet become global and every people started having a personal computer and then a smart phone through which people are generating data every day. So, in the earlier mid 2,000 people had started taking about that what type of technology are platform, we should require to handle a big data aspects.

So, multiple technologies were developed such as ZooKeeper at Yahoo, then big table and MapReduce platform at Google. So, then Cassandra at Facebook so, all these different kind of technologies that was for developing applications or developing analytical method over big data platform, or developing methodologies to effectively analyse data in a time sensitive way. So, if you want to process the trillion of petabytes of data, then time is a huge constant.

So, if you just run a single processing machine with some fixed amount of RAM and processor power, it will possibly take some even millions of years to process your data. So, that is why people are exploring the technology for parallel processing through MPI based techniques, or different kind of techniques through which you can optimise the processing over the big data technology.

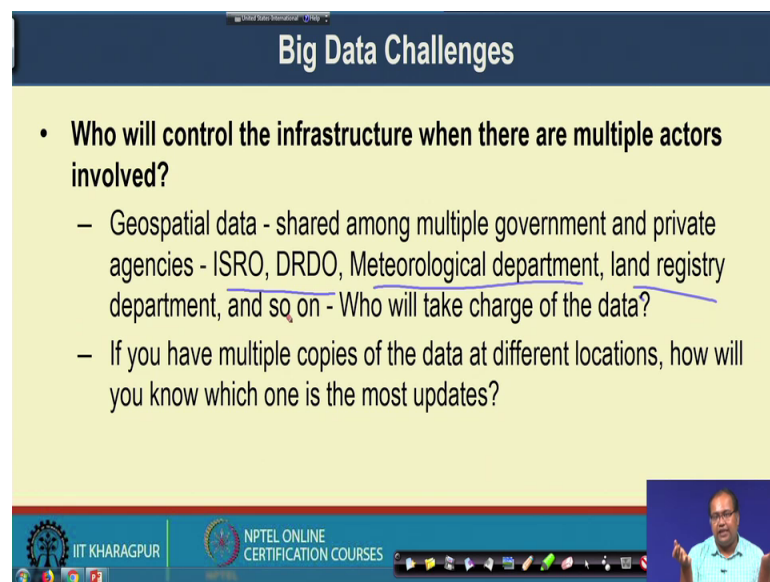
Then came different kind of open source project like this Hadoop distributed file systems; which was there for handling which was there for storing big data in a decentralized platform. So, storing big data is again challenge. Because you have this

huge amount of data how will you effectively store the data in a platform and how will you access it.

So, we require some kind of specialised file system to make distributed storage of data, because a central storage or a single storage is not sufficient to store the entire component of this data which is available. So, that is why people designed this kind of distributed storage architecture; where data is stored into multiple storage environment multiple storage platforms through network attached storage NAS or storage area networks SAN this different kind of distributed storage platform.

And then this Hadoop distributed file system, it helps in effective access of the data over this kind of distributed platform. And then came this Hadoop MapReduce. So, which is again an open resource project, and Hadoop MapReduce say it helped in executing query effectively on a on a HDFS kind of platform by optimizing the query processing time by combining the data by clustering; the data into multiple groups, and then processing the query on the target groups which are interest for the end user; so, that way we witness different such technologies for processing big data.

(Refer Slide Time: 09:05)



The slide is titled "Big Data Challenges" and is presented in a yellow box with a dark blue header. It contains two main bullet points. The first bullet point asks "Who will control the infrastructure when there are multiple actors involved?" and lists several agencies: ISRO, DRDO, Meteorological department, and land registry department. A blue line underlines the text "and sq on - Who will take charge of the data?". The second bullet point asks "If you have multiple copies of the data at different locations, how will you know which one is the most updates?". At the bottom of the slide, there is a small inset video of a man in a light blue shirt speaking. The footer of the slide includes the IIT Kharagpur logo and the text "NPTEL ONLINE CERTIFICATION COURSES".

Big Data Challenges

- **Who will control the infrastructure when there are multiple actors involved?**
 - Geospatial data - shared among multiple government and private agencies - ISRO, DRDO, Meteorological department, land registry department, and sq on - Who will take charge of the data?
 - If you have multiple copies of the data at different locations, how will you know which one is the most updates?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

But, then comes that what are our challenges for handling big data in a large scale platform.

So, the first challenge that we bother about is that who will control the infrastructure when there are multiple actors involved. So, just think of the example of geospatial data that I have mentioned. So, this geospatial data if you want to execute certain geospatial query, the data is shared among multiple government and private agencies. There are multiple stakeholders who hold the data of different types and all those different type of data is required to process geospatial query. Say for example, you want to find out for a specific location say, what is the possibility of having a flood in Mumbai say during April 2, 2018.

If you want to have this kind of query over a big data platform, then you require multiple data. First of all, you require the metrological data, you require the weather data, you require the land usage data, you possibly required data from the marine department, the c level and all these things. So, you require data from multiple such sources and then executive query on top of the data to find out the result. Now interestingly these different type of data are having in the hand of different agencies or different take holder.

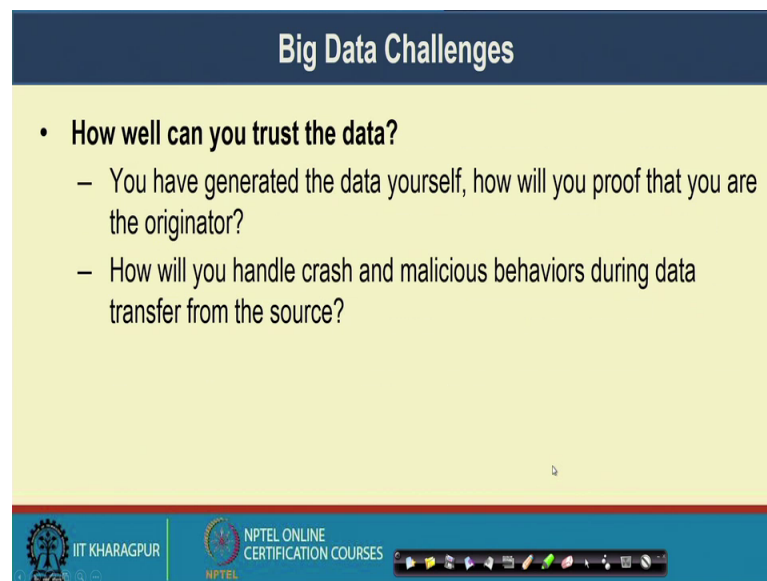
For example, the weather data is derived the weather department the metrological department contents their set of data, the marine department contents constant c level data, the land register contains the land registry data. Then the municipality contains the data of this water disposal and all this individual things. So, that way multiple such stakeholders contain different piece of data and for this effective processing you required the data from multiple stakeholders.

So, the question comes that with this kind of geospatial data; where the data is shared among multiple government and private agencies like say ISRO DRDO metrological department, land registry department, and so on who will take charge of the data. Say, whenever I am taking data from meteorological department and I am trying to combine it with some imaging data from ISRO. Then if there is a data loss or if there is some kind of fraudulent behaviour from my side, then who will take charge of that? Whether that that is because of certain problems, certain security loophole for the data sharing platform of the metrological department, or certain loophole for the data sharing platform from the ISRO.

So, that is a major challenge. So, the first thing the second thing is that if you have multiple copies of data at different locations. How will you know which one is the most

updated data? So, for example, from the ISRO imaging data, you can get information about the land usage. From the land register record you can also get the information about the land usage. How will you know which data is most trustworthy or most recent data? So, that is the challenge related to developing data sharing platform or data sharing infrastructure; that who will have the control over the infrastructure when there are such multiple stakeholders.

(Refer Slide Time: 13:01)



The slide is titled "Big Data Challenges" in a dark blue header. The main content is on a light yellow background and lists two main points under the heading "How well can you trust the data?". The first point is "You have generated the data yourself, how will you proof that you are the originator?". The second point is "How will you handle crash and malicious behaviors during data transfer from the source?". At the bottom of the slide, there is a blue footer containing the IIT KHARAGPUR logo, the text "NPTEL ONLINE CERTIFICATION COURSES", and the NPTEL logo. A navigation bar with various icons is also visible at the bottom right.

Big Data Challenges

- **How well can you trust the data?**
 - You have generated the data yourself, how will you proof that you are the originator?
 - How will you handle crash and malicious behaviors during data transfer from the source?

The second problem that we have is that how will can you trust the data. So, whatever data you are getting whether the data is trustworthy or not.

So, you have possibly generated the data yourself, how will you prove that you are the originator of the data, or you are authorised person to share the data. Or for example, if you are getting certain data from some agencies or some stakeholders, how will you ensure or how will you verify that he is the authorised person to share the data with you are not. Say, these are the kind of important questions, otherwise there can be serious copyright valuation and you can be in a big travel for that.

So, to solve this problem to handle this problem, you have to think of a mechanism to maintained this kind of cooperative relationship among multiple stakeholder or multiple actors. Another problem comes that how will you handle crash and malicious behaviour during data transfer from the source.

Say when you are getting the data from ISRO, it is coming over the network. If certain malicious behaviour happens or if someone make some changes in the data, how will you ensure that that changes have not been done by ISRO or have been done by some malicious agents, or it may also happened that possibly ISRO has share the data with another agency. And they have authorised that agency to share the data further with you, and you are getting the data, how will you detect that that intermediate agency who is sharing the ISRO data with you, maybe with the consent of ISRO, but they have not done any modification on the data. So, the trustworthiness of the data is a major requirement which we need to ensure.

(Refer Slide Time: 14:57)

The slide is titled "Big Data Challenges" in a dark blue header. The main content area is light yellow and contains a bulleted list:

- **How do you monetize the data?**
 - How do you transfer the rights of the data?
 - Can we develop a universal data marketplace? - Look data like electricity or Internet

Below the text is a circular diagram with "BIG DATA" in the center, surrounded by various icons representing data analysis, security, and connectivity. At the bottom left of the slide, it says "Image Source: <http://www.narolainfotech.com>". The footer of the slide is dark blue and contains the logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES, along with a navigation bar.

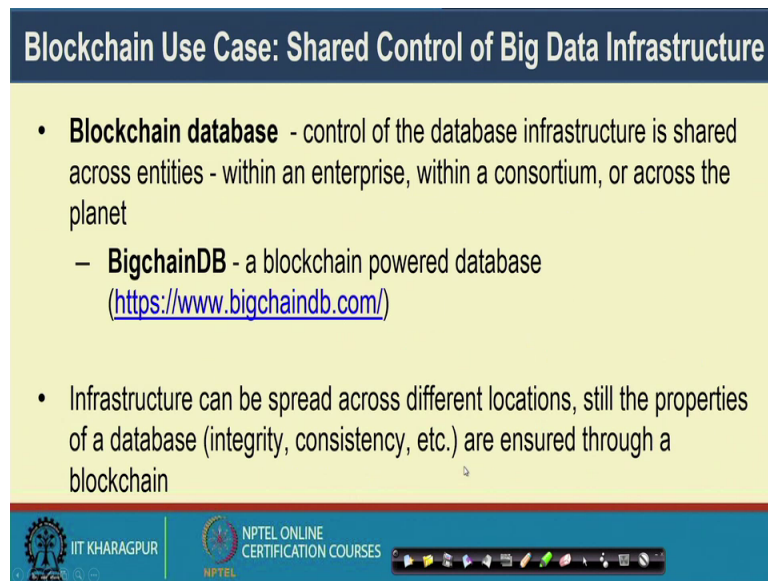
The third requirement is that, how will you monetize the data, how do you monetize the data.

So, the question comes that how do you transfer the rights of the data. Say for example, whenever you are sharing your data with a third party, how will you transfer the right to that third party, how will you ensure that that third party we should not share the data with anyone else or they should share the data under some say control policies. So, that is one important question, and another question which is actually currently floating in the big data community, that can we develop a universal data market place. Just like data like electricity or internet.

So, everyone can buy electricity or buy internet service and then can participate that. So, data can also be a kind of service, because the way we are generating huge amount the data. And if we get certain data certain share of the data, we can possibly developed lots of applications and we can possibly utilise it for developing a better nation.

So, the question comes that how will you handle this kind of universal data market place without having kind of fraud fraudulent or malicious behaviour in the environment.

(Refer Slide Time: 16:27)



Blockchain Use Case: Shared Control of Big Data Infrastructure

- **Blockchain database** - control of the database infrastructure is shared across entities - within an enterprise, within a consortium, or across the planet
 - **BigchainDB** - a blockchain powered database (<https://www.bigchaindb.com/>)
- Infrastructure can be spread across different locations, still the properties of a database (integrity, consistency, etc.) are ensured through a blockchain

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

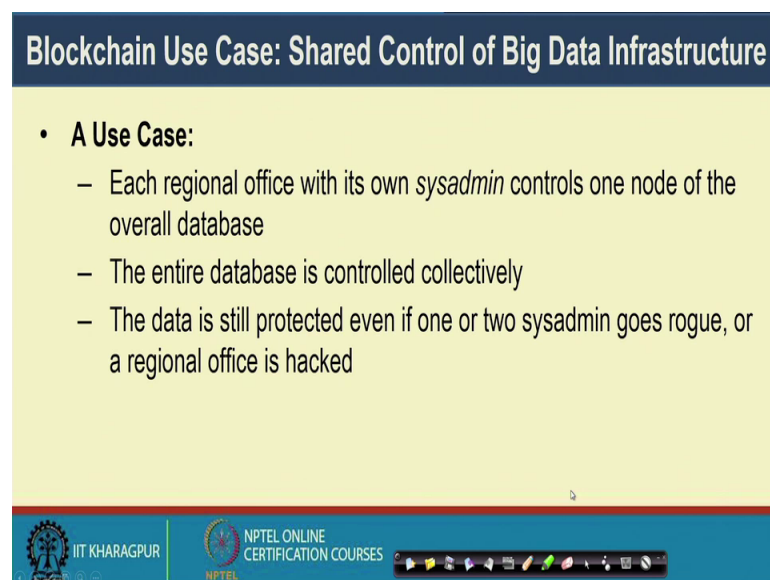
Well, let us look into certain use cases and see that how can we effectively solve this kind of problems. So, the first use case that we are going to discuss is the share control of big data infrastructure. So, just think of a blockchain database, rather than a traditional database that we have. So, if you have a blockchain database. That means, you have say the multiple copies or multiple infrastructural component; which are owned by say different zonal offices and whatever access is being done on these infrastructure, that is maintained the log is been maintain inside the blockchain. So, that we are calling as a kind of blockchain database.

So, the control of the database infrastructure it is shared across entities, the entities can be within an enterprises within an consortium or across the planning. So, there can be a consortium where the where multiple private companies can join together and they agree to share data among themselves. So, they have their own infrastructure and those infrastructure need to be connected in a transude way. So, we have an example of this

kind of blockchain database which is called BigchainDB. We will look into BigchainDB in little details. So, the advantage here is that that the infrastructure can be spread across different locations still the properties of a database like integrity consistency all these things or ensure with the help of blockchain.

So, the blockchain will help that whatever access is being done on top of that data base the transactions are being made; their consistent and they are tamper proof. So, if you access the data through that blockchain, everyone will be able to validate who has access the database, who has access that particular say data storage and what type of data has been accessed.

(Refer Slide Time: 18:27)



Blockchain Use Case: Shared Control of Big Data Infrastructure

- **A Use Case:**
 - Each regional office with its own *sysadmin* controls one node of the overall database
 - The entire database is controlled collectively
 - The data is still protected even if one or two *sysadmin* goes rogue, or a regional office is hacked

The slide footer includes the IIT KHARAGPUR logo, NPTEL ONLINE CERTIFICATION COURSES logo, and a navigation bar with various icons.

Well, let us see specific use case of this. So, each regional office with it is own *sysadmin*, they control one node of the overall database. So now, we are distributing the database among multiple places.

So, the entire database is control collectively, but every regional office has their own *sysadmin* control. And in this architecture data is till protected even 1 or 2 *sysadmin* goes rogue or a regional office is hacked, because even if a regional office being hacked that things are through blockchain. It will not be able to access the infrastructure at the other regional offices.

(Refer Slide Time: 19:11)

Blockchain Use Case: Audit Trails on Data

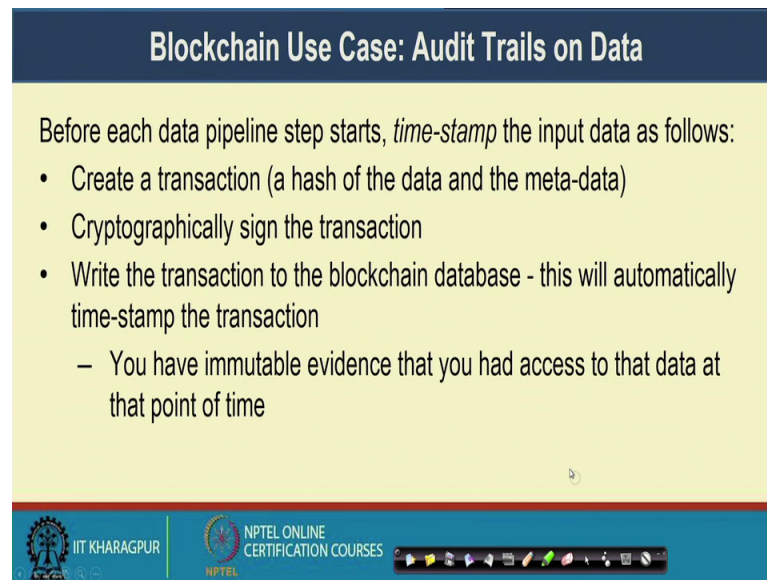
- Consider a data pipeline:
 - IoT Sensors → Kinesis/Event Hub + Stream analysis → HDFS storage → Spark data cleaning → Spark normalization → MongoDB storage → Tableau analytics

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let us look in to a second use case audit trails on data. So, consider a data pipe line the data is being generated and finally, the data is been utilized. So, that data been generated from the IoT sensors. Then you have some kinesis or event hub and stream analysis which is taking that data from the IoT sensors and doing some initial streaming analysis on top of that, then the data is put inside and HDFS storage.

Then you are applying some says spark data cleaning mechanism on top of the IoT sensor data, then you are applying certain normalisations, spark normalisation on the data. Finally, you are transferring the data to a MongoDB storage a specific database storage and then doing the tableau analytics showing the query on top of that data. So, if you consider this data pipeline with the audit trails. So, you mean that you should have an information that how the data how a particular piece of data has passed into this individual pipeline, who has made that passing and on what time it has passed from one stage of the pipeline to the next stage.

(Refer Slide Time: 20:27)



Blockchain Use Case: Audit Trails on Data

Before each data pipeline step starts, *time-stamp* the input data as follows:

- Create a transaction (a hash of the data and the meta-data)
- Cryptographically sign the transaction
- Write the transaction to the blockchain database - this will automatically time-stamp the transaction
 - You have immutable evidence that you had access to that data at that point of time

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

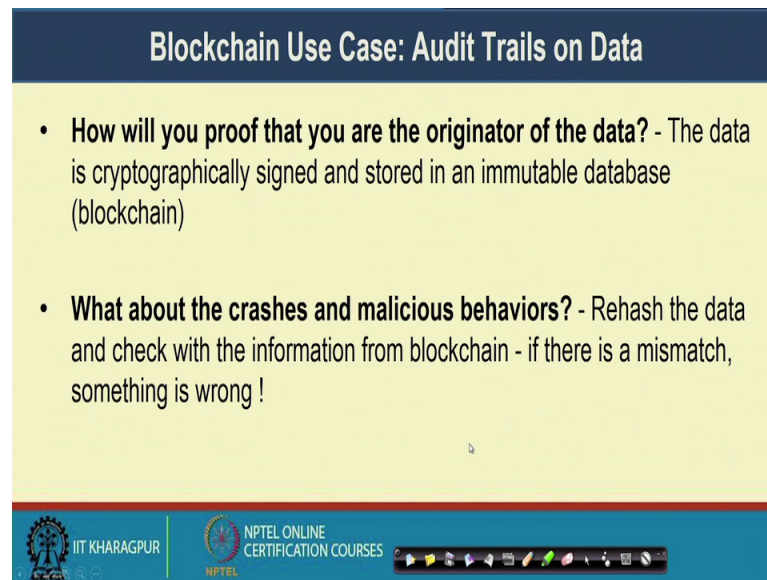
So, we can again solve this with the help of a blockchain. So, think of a solution of something like this; that before each data pipeline step starts you timestamp the input data as follows.

So, you create a transaction so that the transaction is an hash of the data and the corresponding meta data or any other information that you want to add the in the transaction. Then the cryptographically sign of the transaction, and finally you write the transaction to a blockchain database. So, whenever you are writing the transaction to a blockchain database, it will automatically timestamp the transaction.

So, whenever you are writing the transaction in the blockchain data base, you have immutable evidence that you had accessed to the data at that point of time. So, you are ensuring that you have got the data from the IoT sensors, then at that point of time you have done a stream analytics, and then at the next point of time you have put the data in a HDFS database and so on.

So, that way every event which is being happened on top of the data, that is being logged in the form of a transaction inside the blockchain. And with that blockchain the things are tamper proof and you can also do certain verification.

(Refer Slide Time: 21:49)



Blockchain Use Case: Audit Trails on Data

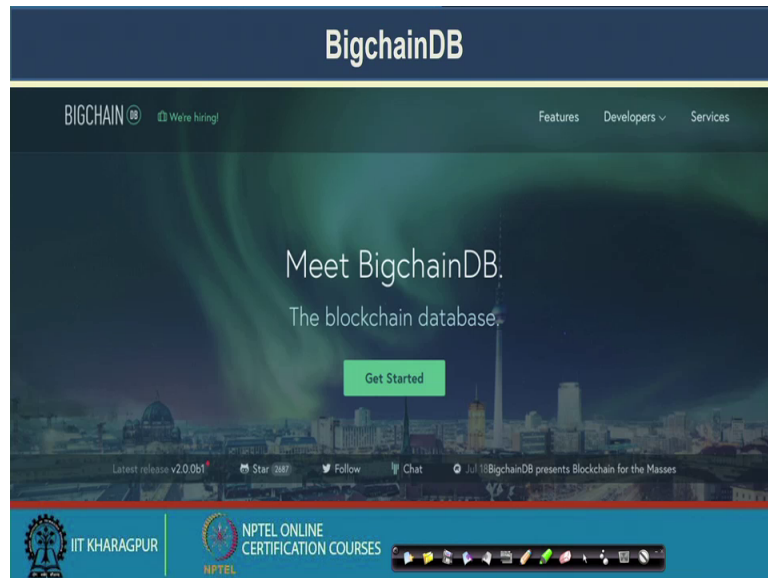
- **How will you prove that you are the originator of the data?** - The data is cryptographically signed and stored in an immutable database (blockchain)
- **What about the crashes and malicious behaviors?** - Rehash the data and check with the information from blockchain - if there is a mismatch, something is wrong !

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this can actually answer few of our questions that we asked before. So, first of all how we will you prove that you are the originator of the data. So, that was one of our original questions. So, to prove that you are the originator of the data say, your data is cryptographically signed by you and stored in the immutable data base. So, that way anyone can use your public key to verify that you are the originator of the data. Because that information is already there in the blockchain and the blockchain is temper proof.

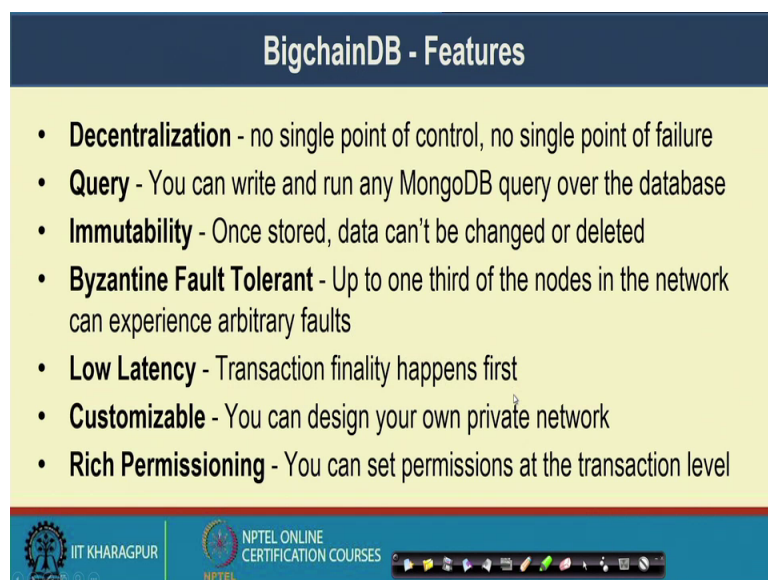
Then the second thing is that, what about crashes and the malicious behaviours that we are talking about. So, if there is certain kind of crash on malicious behaviour, what you can do whatever data you are getting some from the source from data source you re hash the data and check with the information from the blockchain that that is there. So, if there is a match; that means you got the correct copy of the data you can do a validation if there is a miss match; that means, something is wrong somewhere. So that means, the data that you have got possibly it is not the correct copy of the data or someone else has tried to forge that data, ok.

(Refer Slide Time: 23:01)



So, with this 2 use case let us look into a practical blockchain database a open source blockchain database that are I will say which is there in place and it is gaining popularity for big data application. So, it is called BigchainDB. So, I suggest all of you to explore BigchainDB in more details and a do a analysis of it is content it is code and how you can write application on top of BigchainDB.

(Refer Slide Time: 23:35)



So, BigchainDB has the following features, first of all decentralization. So, you do not have any single point of control. So, there is no single point of failure. The second is the

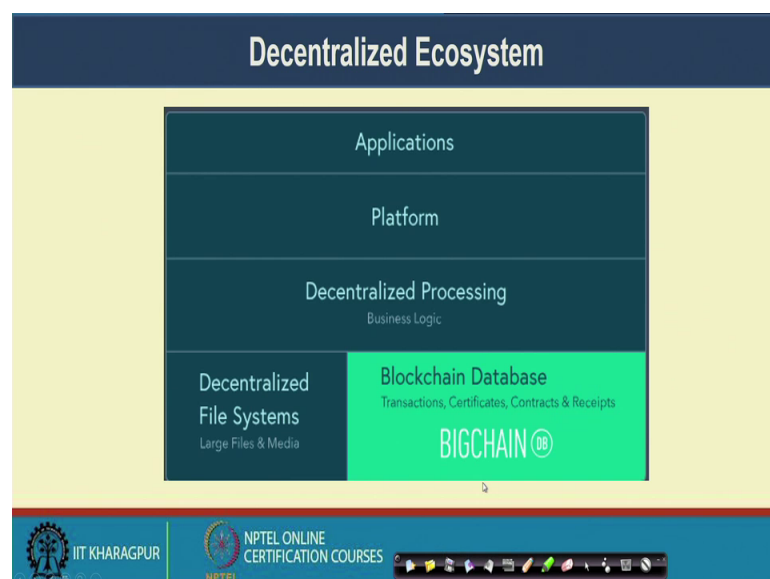
query so, you can write and run any MongoDB query over the database. So, it supports the MongoDB query format, the third properties immutability. So, once the data is stored in the database, that data cannot be changed or deleted. So, that data is immutable, the 4th is it is byzantine fault tolerant.

That is an important property of BigchainDB. So, it is support for and these byzantine fault tolerant properties to support this infrastructure decentralization. So, based on the bfd principal as we discussed earlier it can support up to one third of the nodes one third of the nodes in the network that can experience arbitrary failures. So, even if there is one third of the node at most one third of the nodes experience arbitrary failures.

The system will still be able to recover and we will be able to give you the correct result. The 4th one is the fifth one is the low latency. So, the transactions finally, it happens fast because of this byzantine fault tolerant consensus algorithm which is their. The next one is the customizable. So, you can design your own private network with the help of certain infrastructure and then installing BigchainDB on top of this multiple machines and then connecting them with each other.

And then it has a rich permission in support that you can set permissions at the transaction level. So, you can specify that this particular transaction should be accessible by this person or this group of persons and no one sells in the universe. So, it is supports the good support of access control list or access control mechanism.

(Refer Slide Time: 25:33)



So, as I mentioned that this BigchainDB provides decentralized ecosystem for big data applications. So, you can write the application of your standard big data processing using MongoDB, then you have this platform which is again decentralized platform, you can have multiple data server. So, at this BigchainDB is installed and then can connect it with each other. Just like a peer to peer architecture it supports the decentralized processing of the business logic.

Like, whenever you are throwing the query, it will take care of that in which particular data source the data is there and it will access it and perform the query. And that the lower end in the stack, you have the decentralized file system to store the file in a decentralization environment. And the bigchain data base BigchainDB data base the blockchain data base that we are talking about.

(Refer Slide Time: 26:35)

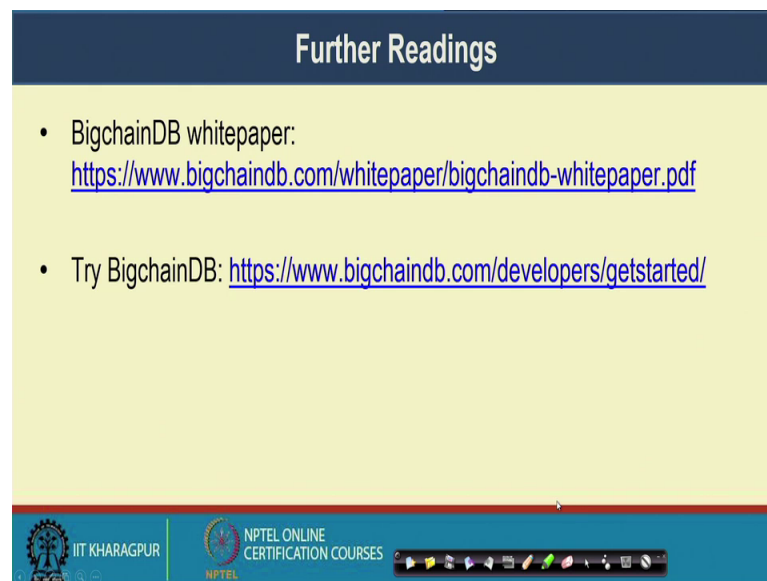
	Typical Blockchain	Typical Distributed Database	BigchainDB
Decentralization	✓ / ✗		✓
Byzantine Fault Tolerance	✓ / ✗		✓
Immutability	✓ / ✗		✓
Owner-Controlled Assets	✓ / ✗		✓
High Transaction Rate		✓ / ✗	✓
Low Latency		✓ / ✗	✓
Indexing & Querying of Structured Data		✓ / ✗	✓

So, here is a comparison between a typical blockchain architecture and typical this data base distributed data base architecture. Interestingly BigchainDB combines the properties of blockchain allocate distributed data base. So, the blockchain supports decentralization byzantine fault tolerance immutability and owner controlled assets. Whereas, a typical distributed data base it supports high transaction rate, low latency and indexing and querying unstructured data.

BigchainDB actually combines all these features all together. So, it takes the power of blockchain technology to make a decentralized platform with byzantine fault tolerance,

and immutable architecture and this (Refer Time: 27:26) controlled assets for a which a permission in of access control. And it combines it with the features of typical a big data base, distributed data base like a high transaction rate low latency and indexing and querying on big data through map reduce or this kind of architecture; which is supported by MongoDB kind of data base. And combines and having central controlled platform. So, that is something a description of a BigchainDB.

(Refer Slide Time: 28:05)



Further Readings

- BigchainDB whitepaper: <https://www.bigchaindb.com/whitepaper/bigchaindb-whitepaper.pdf>
- Try BigchainDB: <https://www.bigchaindb.com/developers/getstarted/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, here are some further readings that you can explore.

So, there is a whitepaper from BigchainDB. So, the link is given here so, I suggest all of you to explore it and find out different properties of BigchainDB, and how BigchainDB actually implements it at the code level combines the feature of database as well as the blockchain technology all together. And you can also try BigchainDB. So, you can go to their developer platform and here in this particular link. All this things are under this dot dot dot BigchainDB dot com website. So, you can install it, you can create a local cluster of a few machines, and start implementing your own database and running your own queries. So, just try it out hopefully it will be enjoyable for you.

So, thank you all for attending the class. So, in the next last we will come with another application with certain research aspects of blockchain for data science or data processing applications.

Thank you all.