

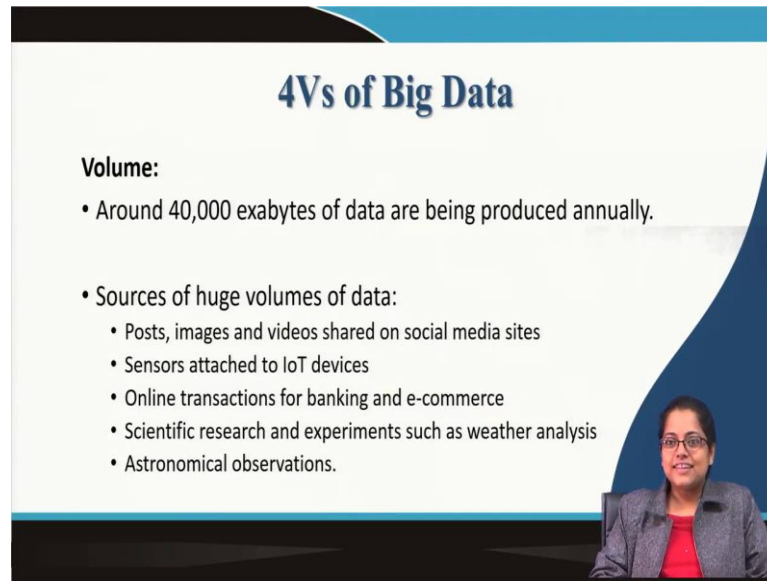
So, the first property is the volume of data, which means the data volume is extremely high. The second represents the rate at which the data flows or the rate at which the data is available, which is again very high and the third refers to the variety of data. So, the data set that we have or the data set in question has a lot of variety, which means it comprises unstructured semi structured as well as structured data. So, what are these different types of data in terms of the structure?

What is structured data? What is unstructured data and so on? We will be discussing soon. Prior to that, you know the key enablers for the growth of big data. Because earlier, you know maybe 10 years back we had a data set, but we could not term that data set as big data, why so? Because in those days, maybe 20 years back or 15 years back or even 10 years back storage capacity was not so huge as it is today. Moreover, processing power was not enough and also the availability of data.

So, data was collected all the time, but the availability of data was not as much as it is today. So, because in today's context if you are wearing a wrist smart wrist band, you would be producing huge amounts of data every second, right. Departmental store such as a Big Bazaar, would be producing huge amount of transactional data every day.

So, if we consider the availability of data, it has become huge and the arrival of internet of things, which we had discussed in the previous lectures has facilitated the collection of data to a large extent. So, because of IoT huge amount of data is available today. Therefore, the key enablers for the growth of big data in today's context are increase in storage capacities, increase of processing power and availability of data.

(Refer Slide Time: 03:42)



4Vs of Big Data

Volume:

- Around 40,000 exabytes of data are being produced annually.
- Sources of huge volumes of data:
 - Posts, images and videos shared on social media sites
 - Sensors attached to IoT devices
 - Online transactions for banking and e-commerce
 - Scientific research and experiments such as weather analysis
 - Astronomical observations.

So, we had mentioned that there are three primary characteristics that a data set should possess in order to be qualified as big data. But, in fact today there are not just three we have four. So, let us discuss those. The first V stands for volume of data so big data has huge volume. For example, around 40,000 exabytes of data are being produced annually and an exabyte is 10 to the power 18 bytes.

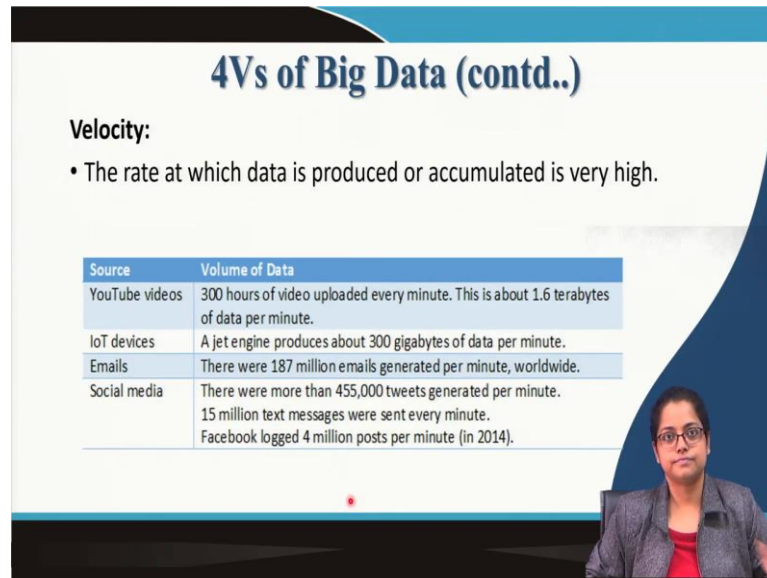
So, that much of data huge volume of data; 40,000 exabytes of data being produced annually. Sources of huge volumes of data, because today data is collected from multiple sources and those data sources have the capacity of generating huge volumes of data. So, for example, posts, images, videos shared on social media, sensors attached to IoT devices as we have just mentioned, online transactions for banking and e commerce.

So, e commerce firms we were just talking about say a D Mart or you know any departmental store or say an e-commerce company like Amazon, which produces huge amount of transactional data every day. Scientific research and experiments. Scientific researches and research and experiments such as weather analysis, produces again huge amount of data.

Astronomical observations so huge huge amount of data; so, all of these are only a few examples. There could be a lot of other examples of data sources which produce huge volumes of data.

So, volume of data is the first essential property of the data set, which has to be met in order to be for the data set to be qualified as big data.

(Refer Slide Time: 05:43)



4Vs of Big Data (contd.)

Velocity:

- The rate at which data is produced or accumulated is very high.

Source	Volume of Data
YouTube videos	300 hours of video uploaded every minute. This is about 1.6 terabytes of data per minute.
IoT devices	A jet engine produces about 300 gigabytes of data per minute.
Emails	There were 187 million emails generated per minute, worldwide.
Social media	There were more than 455,000 tweets generated per minute. 15 million text messages were sent every minute. Facebook logged 4 million posts per minute (in 2014).

The slide features a presenter in the bottom right corner, a red cursor arrow at the bottom center, and a blue decorative wave on the right side.

So, moving on. The second V of the 4 Vs is velocity, by velocity which we mean the rate at which the data is produced. So, for a data set to be qualified as big data; the rate at which it is produced or accumulated should be very high. Let us take a few examples, YouTube videos. So, if YouTube video is considered to be a source of data YouTube 300 hours of video are uploaded every minute, can you imagine? So, 300 hours of YouTube videos are uploaded every minute.

This is about 1.6 terabytes of data per minute which is again huge. IoT devices, a jet engine produces about 300 gigabytes of data per minute. So, 300 gigabytes of data is also very huge. Emails, there are around 180, so this is a fact that there were around 187 million emails generated per minute worldwide 187 million emails per minute and we are tired writing drafting just one email. So, there were around 187 million emails generated per minute worldwide. So, that is huge-huge amount of data again.

Now, social media, again a, you know, one of the primary sources of big data; so, there were more than 455,000 tweets generated per minute. So, Twitter, a very popular social media generated around 455,000 tweets per minute, which is again very huge; 15 million text messages were sent every minute; 15 million.

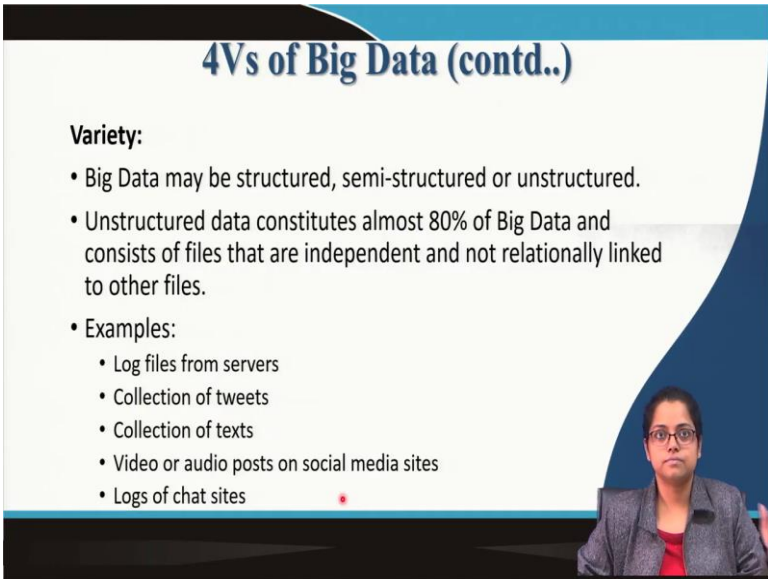
So, we can only imagine because we have maybe sent 2, 3 texts per minute, but here 15 million text messages were sent every minute and this is all of these pertain to data in 2014.

Today, possibly the volume or the velocity has become much more. So, Facebook logged 4 million posts per minute in 2014. So, 4 million post so all of these you know point towards the fact that the data sources produce data at a very rapid rate, the pace of production of data or availability of data is immense. So, YouTube videos 300 hours of video every minute; so, you can imagine the rate at which the videos are being produced. Emails: the rate at which the email data is being generated.

Social media, also you know I will give you another example, telephonic conversations. So, telephonic conversations form a very important example of data having a huge velocity. So, you imagine the speed at which we talk. So, we are generating huge amount of data every minute or every second. So, all of these you know data sets are generated at a very rapid rate. Therefore, the velocity of data here is very high, so we can consider that this kind of data pertains to has one of the features of big data.

So, moving on, after 'volume' and 'velocity', we will talk about the third V of big data that is the third property of big data which is 'variety' of data.


(Refer Slide Time: 09:18)



4Vs of Big Data (contd.)

Variety:

- Big Data may be structured, semi-structured or unstructured.
- Unstructured data constitutes almost 80% of Big Data and consists of files that are independent and not relationally linked to other files.
- Examples:
 - Log files from servers
 - Collection of tweets
 - Collection of texts
 - Video or audio posts on social media sites
 - Logs of chat sites



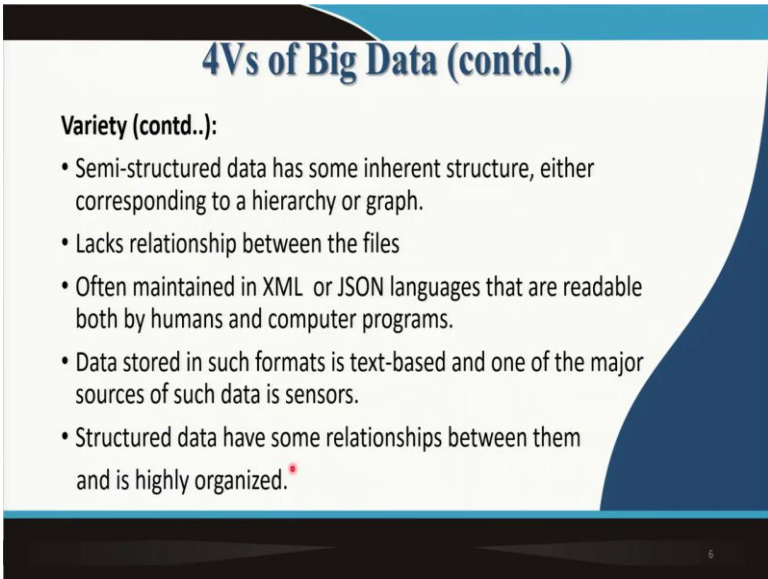
So, by variety of data, we mean data may be structured, semi structured or even unstructured. So, what do we mean by these three categories of data? When do we say data is structured? When would we consider data to be unstructured and when we would consider it to be semi structured?

So, a big unstructured data constitutes almost 80 percent of big data and consists of files that are independent and not relationally linked to other files. So, in one of the previous lectures, where we were talking about relational databases we understood the concept of relations between data.

So, how data is related? With the help of entities and with the help of their cardinality we have tried to understand how data could be related to each other. But unstructured data, has absolutely no relationship with between it. So, a data files which contain unstructured data are not relationally linked to other data files. Examples, logs from servers so such data are not linked to other data sources.

Collection of tweets, collection of texts. A video or audio posts on social media they are usually not related, logs of chat sites. So, all of these; all of these here pertain to unstructured data and unstructured data come you know forms the chunk the major chunk of big data around almost 80 percent of big data is made up of unstructured data.

(Refer Slide Time: 11:02)



4Vs of Big Data (contd.)

Variety (contd.):

- Semi-structured data has some inherent structure, either corresponding to a hierarchy or graph.
- Lacks relationship between the files
- Often maintained in XML or JSON languages that are readable both by humans and computer programs.
- Data stored in such formats is text-based and one of the major sources of such data is sensors.
- Structured data have some relationships between them and is highly organized.

6

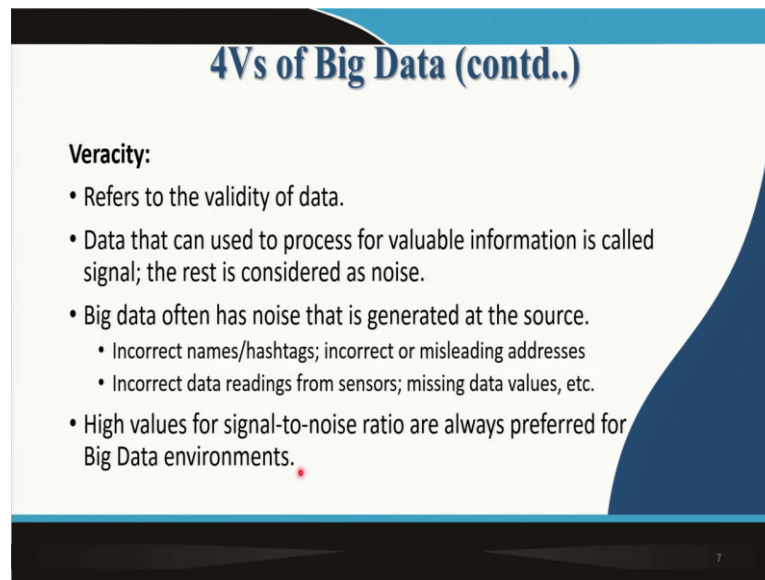
Moving on, the other two you know types of variety of data are you know data can also be not neither structured nor unstructured so it could be semi structured. Semi structured data has some inherent structure, either corresponding to a hierarchy or a graph. It lacks relationship between files.

So, semi structured data is often maintained in XML or JSON languages that are readable both by humans and computer programs. So, XML stands for extensible markup language and JSON stands for JavaScript Object Notation languages. So, these languages are written in the form of texts. Data stored in such formats is text based and one of the major sources of such data is sensors. So, because more and more data is being generated through IoT systems with the help of sensors.

So, data generated by sensors is often in the XML or JSON format. So, such data is usually semi structured. So, there is no relationship between the data, but the data could be you know in the form of a hierarchy or in form of a graph. Finally, structured data have some relationships between them and they are highly organized. So, by structured data we mean, we had discussed again about in the previous lecture we had spoken about you know data being relations between tables.

So, that kind of a data is absolutely structured. Structured data have some relationships between them and is highly organized. Therefore, the third V pertains of big data pertains to variety. So, here we have three we have already discussed three Vs of big data volume, velocity and variety.

(Refer Slide Time: 13:07)



4Vs of Big Data (contd.)

Veracity:

- Refers to the validity of data.
- Data that can be used to process for valuable information is called signal; the rest is considered as noise.
- Big data often has noise that is generated at the source.
 - Incorrect names/hashtags; incorrect or misleading addresses
 - Incorrect data readings from sensors; missing data values, etc.
- High values for signal-to-noise ratio are always preferred for Big Data environments.

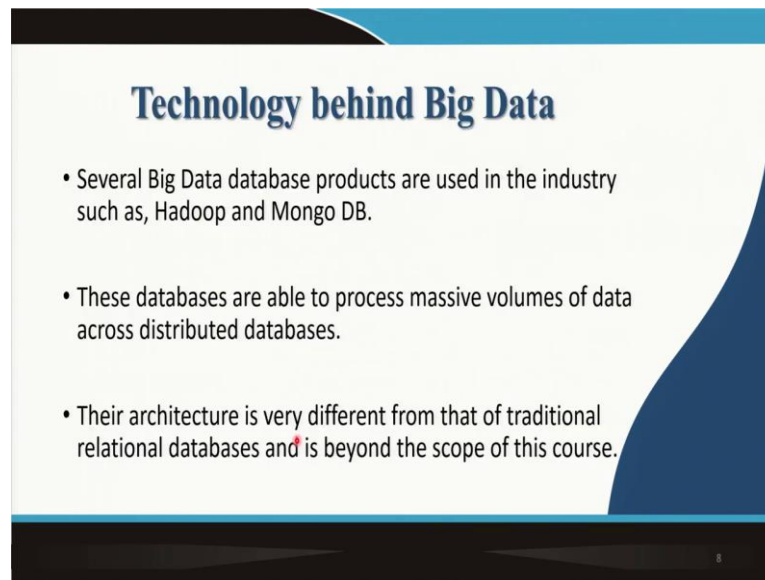
Now, moving on to the fourth V, the fourth V refers to veracity of data. By veracity we mean validity of data. So, validity means that you know how relevant your data is. So, a data that can be used to process for valuable information is called signal, the rest is considered as noise.

So, if you see here data that can be used to process for valuable information is called signal and the rest is considered as noise. Big data often has a lot of noise that is generated at the source. Incorrect names, hashtags or you know incorrect or misleading addresses, incorrect data readings from sensors, which happens all the time it could happen anytime, missing data values.

So, all of these are considered as noise. They cannot be used in further data processing. Therefore, they are considered as noise, but data that is relevant or valid and can be used for further processing is considered as signal. Now, high values for signal to noise ratio are always preferred for big data environments. So, a data that has got higher veracity should have a higher signal to noise ratio, which means noise should be as minimal as possible compared to valid data.

So, for big data environments, data should have high veracity which means it should have high signal to noise ratio. So, here we have discussed the 4 essential Vs of big data, which characterize a data set as big data volume, variety, veracity and velocity of data; right; so the 4 Vs.

(Refer Slide Time: 15:07)



Technology behind Big Data

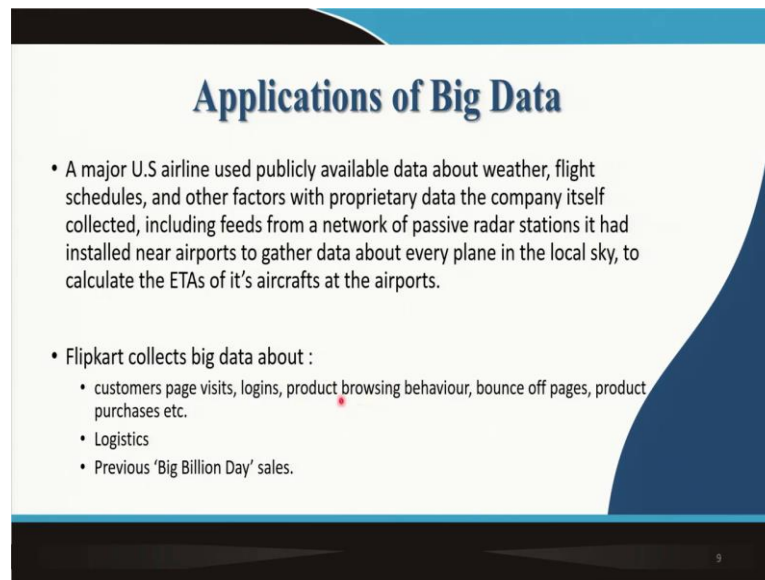
- Several Big Data database products are used in the industry such as, Hadoop and Mongo DB.
- These databases are able to process massive volumes of data across distributed databases.
- Their architecture is very different from that of traditional relational databases and is beyond the scope of this course.

Now, moving on technology behind big data; so, the ordinary or traditional database management systems that we had discussed in the second module are not sufficient for big data environments. Because in big data, data pertains to huge volumes there could be a structured, semi structured as well as unstructured data and also data could have a lot of you know very high velocity. Therefore, traditional databases, relational databases that we had discussed are not capable of handling a big data processing.

So, several big data database products are used in the industry such as, Hadoop and Mongo DB. So, these are very popular infrastructure or technology behind processing of big data. Now these databases are able to process massive volumes of data across distributed databases. Their architecture is very different; we have already mentioned this that their architecture is very different from that of traditional relational databases and you know, discussion of this is actually beyond the scope of this course.

Because this is an introductory course, in management information system. So, discussing about Hadoop Mongo DB and other big data technologies is actually beyond the scope of this course, but if you want to learn more about these you should take up a course in big data technology. So, with this let us move on to some applications of big data.

(Refer Slide Time: 16:50)



Applications of Big Data

- A major U.S airline used publicly available data about weather, flight schedules, and other factors with proprietary data the company itself collected, including feeds from a network of passive radar stations it had installed near airports to gather data about every plane in the local sky, to calculate the ETAs of it's aircrafts at the airports.

- Flipkart collects big data about :
 - customers page visits, logins, product browsing behaviour, bounce off pages, product purchases etc.
 - Logistics
 - Previous 'Big Billion Day' sales.

There are certain very important applications, because big data with the arrival of IoT is becoming more and more and more relevant form of technology to handle, huge amounts of data sets that flow with rapid velocity and have a huge variety. Let us take an example, a massive US airline used publicly available data.

So, let me give you a background before going into the example. You know, in airlines there is the concept of predicting the estimated time of arrival of flights and this is often compared with the actual time of arrival.

So, estimated time of arrival is very extremely important, because it determines you know at what time the passengers would board the flight? It also determines how ready the airport is to handle the traffic at that point in time. Now, so estimated time of arrival of airlines is very important and in general there should be absolutely minimal deviation of the estimated time of arrival from the actual time of arrival, but what happens in you know in case of let us take this example of a major US airline used which used publicly.

So, in general the estimated time of arrival is given by pilots, but pilots do have a lot of other things to focus on. So, in general this particular US airline observed that there was a deviation of around 10 minutes or more between the estimated time of arrival and the actual time of arrival of flights.

So, there was a problem there as you can understand. So, now this particular airline used big data so a data available from about weather flight schedules and other factors which are publicly available data. Along with some proprietary data the company itself collected, including feeds from a network of passive radar stations it had installed near the airports to gather data about every plane in the local sky, to calculate the ETAs of its aircrafts at the airports.

And by doing so it observed that gradually, by you know collecting data from all of these through all of these resources and many more. They were able to predict the time of arrival of aircrafts much more accurately. So, then the deviation between the ETA estimated time of arrival and the actual time of arrival gradually reduced, which is what is expected and required.

So, and of course, this particular data pertains to a big data because it has got all the 3 Vs, as its properties data has huge volume because it is collecting data about every plane in the local sky.

So, it has got huge volume, it has got lot of velocity because data is constantly flowing in more and more aircrafts are coming in every day has you know so many aircrafts landing and so many aircrafts actually taking off, landing is more important in this context. And at the same time the data has a lot of variety also, because you are using proprietary data, you are using data publicly available and data from multiple sources also.

So, this is from radar station, this is publicly available data and there could be data pertaining to you know about pilots, there could be data about aircraft characteristics and so on, which are used to predict the ETA. So, with that we will move on to the next example. Flipkart, which is a very popular e commerce giant in India, collects big data about a lot of you know lot of activities daily.

So, let us take some examples. Customer's page visits, logins, product browsing behaviour, bounce off pages, product purchases etcetera. So, data is collected from all customer behaviour that happens during the entire day and this is huge, has a lot of variety at the same time it has a lot of high huge very high velocity also.

So, data collected in this through about these various parameters can be actually utilized to find out about customers' behavior on websites, products that they are in general

purchasing and bounce off pages which means that which are the pages on the website which are not performing well.

So, which are the pages which are actually you know putting off the customers so that they are bouncing off from the website. So, these all of these could give you know good pointers to the company to actually improve certain pages, to stock in certain products, if customers are purchasing certain products and so on. Logistics, so Flipkart collects again collects a lot of data about logistics.

So, for example, it collects different pin codes through which customers generally of order products. So, if the pin codes are available Flipkart would be able to streamline or optimize its delivery schedule. Previous, Big Billion Day sales. So, those of you who are not familiar Big Billion Day sale is a very very popular flash sale that happens in the Indian context by the e commerce giant Flipkart.

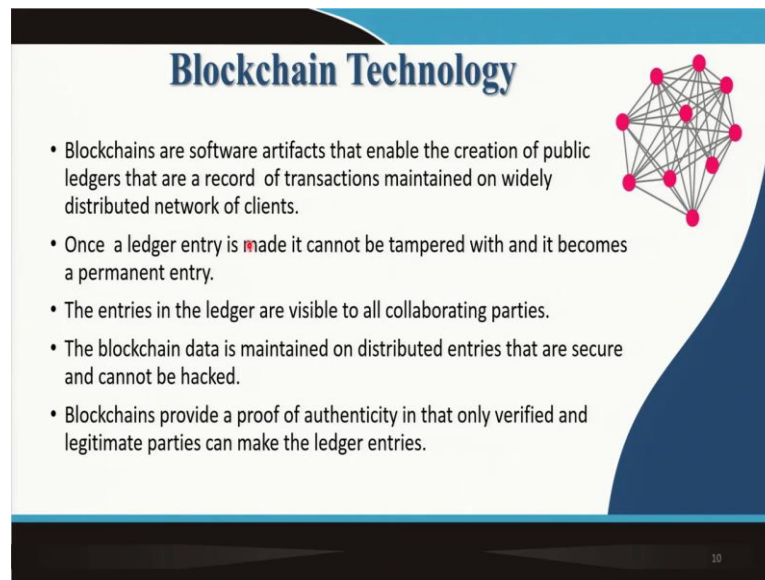
So, every year there are around one or two Big Billion Day sales in general there is one. So, what Flipkart does is, it collects a lot of information about all its prior Big Billion Day sales. So, that it can use that information to strengthen or prepare itself for the next Big Billion Day sale.

So, for example, which are the products that people are offering more, which are the products that are being sold more during are ordering sorry not offering ordering more, which are the products that are being purchased more during the you know short period of time that is the flash sale.

So, all of these could actually give lot of pointers at as to what product should be stocked and which should not be. This data also is pertains to big data, because data is flowing at a very rapid speed within a short period of time you know say a few hours all the sales are happening and data pertains to multiple to a lot of has a lot of variety and of course, data is huge in volume. So, these are certain applications of big data.

So, moving on again let me just you know you know in the session on IoT. We had you know, we had discussed about smart cities smart cities use lot of applications of big data for a lot of you know coming up with a large number of applications. So, from traffic management system to water management in a smart city everything uses both these technologies so big data and IoT together in combination.

(Refer Slide Time: 24:11)

A presentation slide titled "Blockchain Technology" with a blue and black background. The title is in a bold, blue font. To the right of the title is a network diagram consisting of several red nodes connected by thin grey lines. Below the title is a bulleted list of five points. The slide has a dark blue curved shape on the right side and a small number "10" in the bottom right corner.

Blockchain Technology

- Blockchains are software artifacts that enable the creation of public ledgers that are a record of transactions maintained on widely distributed network of clients.
- Once a ledger entry is made it cannot be tampered with and it becomes a permanent entry.
- The entries in the ledger are visible to all collaborating parties.
- The blockchain data is maintained on distributed entries that are secure and cannot be hacked.
- Blockchains provide a proof of authenticity in that only verified and legitimate parties can make the ledger entries.

Now, we have discussed a lot about big data moving on to the next technology, emerging technology, of course, which is very-very promising and has a lot of potential in the future block chains. So, block chains are software artifacts that enable the creation of public ledgers that are a record of transactions, maintained on widely distributed network of clients.

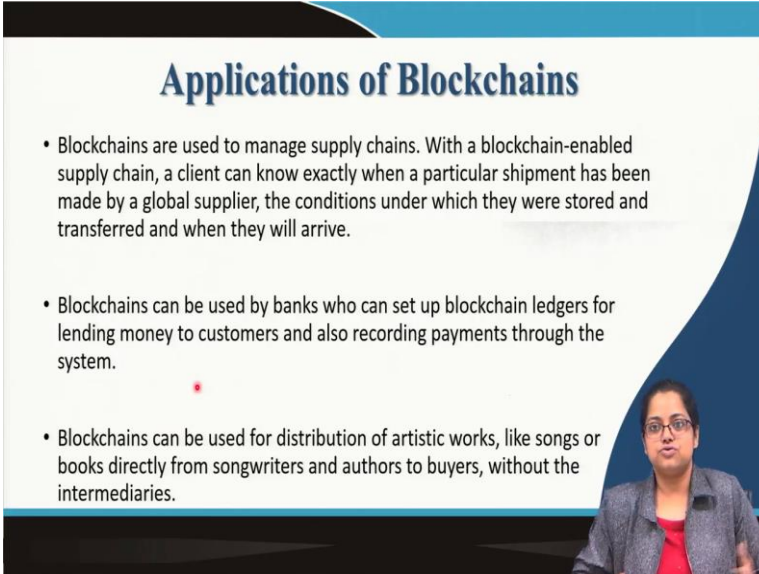
So, there are widely distributed network of clients and public ledgers are distributed these are records of transactions, on these widely distributed network of clients. Once a ledger, so a block chain has very you know certain very important features, which make it very useful for large number of applications.

So, once a ledger entry is made, it cannot be tampered with and it becomes a permanent entry and the entries in the ledger are visible to all collaborating parties. So, these properties make it very useful for being used in banking applications, wherein say you are giving out a loan and they you are giving out a loan and you know a note of that transaction is available on every system that is a part of the network.

So, it is maintained on every ledger, that is a part of the network. So, you cannot tamper with it, so you cannot later on go back and deny taking the loan and the entries are visible to all the collaborating parties. So, of course, there is no question of denial, there is no question of going back and everybody is a party to it.

Block chain data is maintained on distributed entries that are secure and cannot be hacked. Block chains provide a proof of authenticity in that only verified and legitimate parties can make the ledger entries.

(Refer Slide Time: 26:14)



Applications of Blockchains

- Blockchains are used to manage supply chains. With a blockchain-enabled supply chain, a client can know exactly when a particular shipment has been made by a global supplier, the conditions under which they were stored and transferred and when they will arrive.
- Blockchains can be used by banks who can set up blockchain ledgers for lending money to customers and also recording payments through the system.
- Blockchains can be used for distribution of artistic works, like songs or books directly from songwriters and authors to buyers, without the intermediaries.

The slide features a blue and white color scheme with a curved design on the right side. A small red dot is visible on the slide. In the bottom right corner, there is a video feed of a woman with glasses, wearing a grey jacket over a red top, who appears to be presenting the slide.

So, moving on, let us see some applications of block chains in business. Block chains are used to manage supply chain, so they have a huge application in supply chains especially in global supply chains. So, with a block chain enabled supply chain a client can know exactly when a particular shipment has been made by a global supplier.

Because if you are a company dealing with multiple global suppliers, it might be very difficult for you to keep track of you know when the supplier has shipped a particular product or part, the conditions on under which the products the parts were actually stored and transferred and when they will arrive?

So, but with block chain, since a ledger entry is made at every stage and that ledger entry is available to all parties to the network. A client can easily know when a particular shipment has been made by a global supplier, the conditions under which they were stored and transferred and exactly when they will arrive.

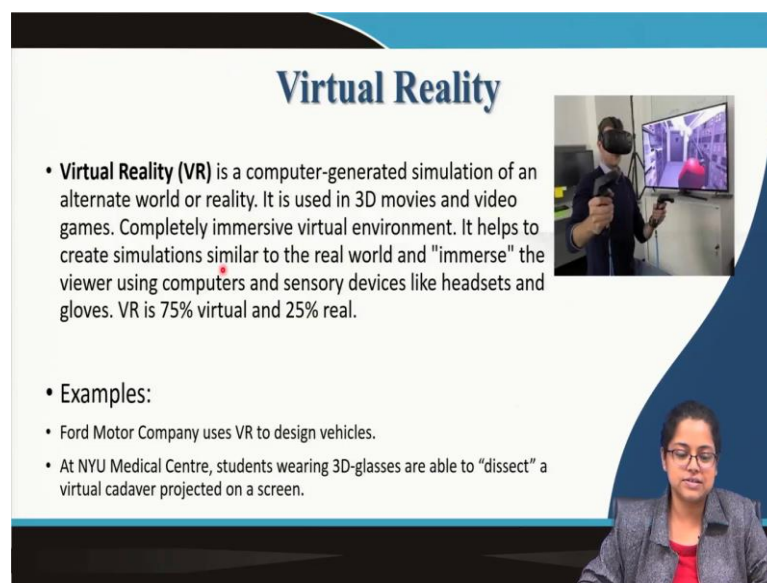
So, block chains have a very important role in supply chains. Block chains can be used by banks, who can set up block chain lend ledgers for lending money to customers and also recording payments through the system, we have already discussed this. Now, block

chains also have a huge application or potential in distribution of artistic works like songs or books directly from songwriters or authors to buyers, without the involvement of any intermediary.

So, artistic work can be directly you know that, you know if there are intermediaries involved in distribution network they there is a lot of information asymmetry and there are certain problems associated with it, certain other problems. So, in order to eliminate those block chains can actually get rid of intermediaries in between and the books or the songs can directly be you know distributed from the songwriters or the buyers or the authors to the buyers, without the involvement of any intermediary.



So, these are only few applications of block chains that we have discussed. There are of course, a lot of other applications of block chains in business. Now, let us move on to the last but very important emerging technology in the world of you know information systems, virtual reality and of course, along with it augmented reality.

(Refer Slide Time: 28:32)



Virtual Reality

- **Virtual Reality (VR)** is a computer-generated simulation of an alternate world or reality. It is used in 3D movies and video games. Completely immersive virtual environment. It helps to create simulations similar to the real world and "immerse" the viewer using computers and sensory devices like headsets and gloves. VR is 75% virtual and 25% real.
- **Examples:**
 - Ford Motor Company uses VR to design vehicles.
 - At NYU Medical Centre, students wearing 3D-glasses are able to "dissect" a virtual cadaver projected on a screen.



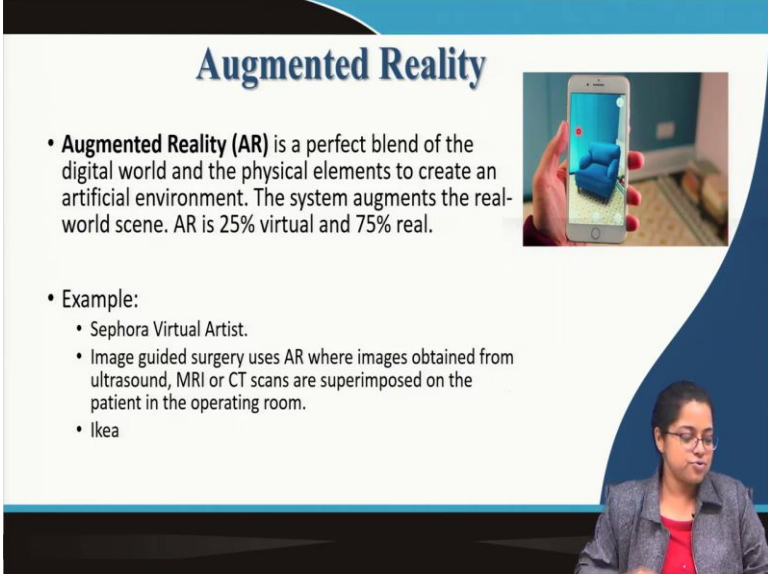
So, a virtual reality is a computer generated simulation of an alternate world or reality. So, these are immersion technologies, using which the customer would be immersed into a different world altogether. It is used in 3D movies most of you must have watched certain 3D movies and video games. A completely immersive virtual environment, it helps to create simulations, which are similar to real world and immerse the viewer using computers and sensory devices like, headsets as you see here and gloves.

Virtual reality is 75 percent virtual and only 25 percent real. So, here you see you know this person has been using virtual reality through a headgear and is immersed in a different world altogether. So, there are a lot of examples of its application in the real world. Ford motor company uses virtual reality to design vehicles.

Wherein, the designer is exposed to a virtual world through VR, wherein he can actually find out you know in the virtual world without actually beings without actually sitting in a car would be able to figure out the distance between the dashboard and the steering and would be able to plan accordingly in the real life vehicle. Another example is at NYU medical center students wearing 3D glasses are able to dissect a virtual cadaver projected on a screen.

So, it is another very-very interesting and noteworthy application of VR. VR has a lot of other applications, but we would like to discuss you know only a few here you can take up you know further studies on VR to understand its role in today's world.

(Refer Slide Time: 30:30)



Augmented Reality

- **Augmented Reality (AR)** is a perfect blend of the digital world and the physical elements to create an artificial environment. The system augments the real-world scene. AR is 25% virtual and 75% real.
- Example:
 - Sephora Virtual Artist.
 - Image guided surgery uses AR where images obtained from ultrasound, MRI or CT scans are superimposed on the patient in the operating room.
 - Ikea

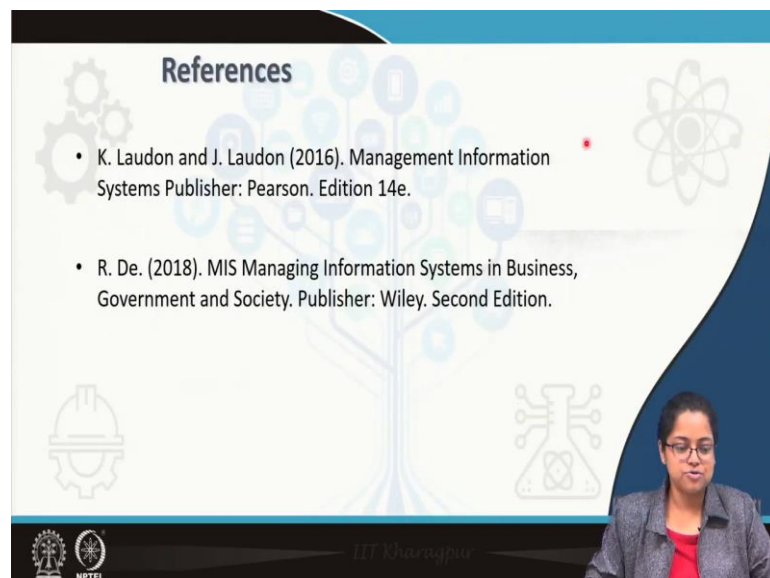
Now, coming to the last you know emerging technology, which is a deviant of VR is augmented reality or AR. Augmented reality is a perfect blend of the digital world and the physical elements to create an artificial environment, the system augments the real world scene. So, contrary to VR which is 75 percent virtual and only 25 percent real. AR is 25 percent virtual and 75 percent real. So, the real component is much more in the context of augmented reality.

But nonetheless both of these are very very you know important tools, in all applications ranging from e commerce to the world of you know medicine everywhere VR and AR have a role to play. So, here we see Sephora virtual artist, Sephora is a very renowned you know cosmetics giant. So, Sephora has this virtual artist application, using which uses augmented reality through which a customer can actually see how a particular shade of a cosmetic looks on him or her.

So, it uses augmented reality here. Image guided surgery very important application of AR, where images obtained from ultrasound MRI or CT scans are superimposed on the patient in the operating room. So, again a very useful application of AR. Ikea, which is a you know very renowned furniture retailer. So, Ikea so for example, here we see that using augmented reality you can see how a particular piece of furniture looks in a corner of your house.

So, you can superimpose the piece of furniture on your, you know, in your houses' settings and see, how it looks using an augmented reality application. So, these are two very interesting technologies which have a lot of potential in the future but we have just tried to introduce you to these two technologies in this particular lecture.

(Refer Slide Time: 32:48)



So, here we come to the end of our session on 'Emerging Technologies' and here we have discussed from cloud computing to internet of things to big data, block chains, AR and VR, which are all emerging technologies, which play a very important role in the

world of information system and they have actually transformed the world of information system today.

Thank you, see you around!