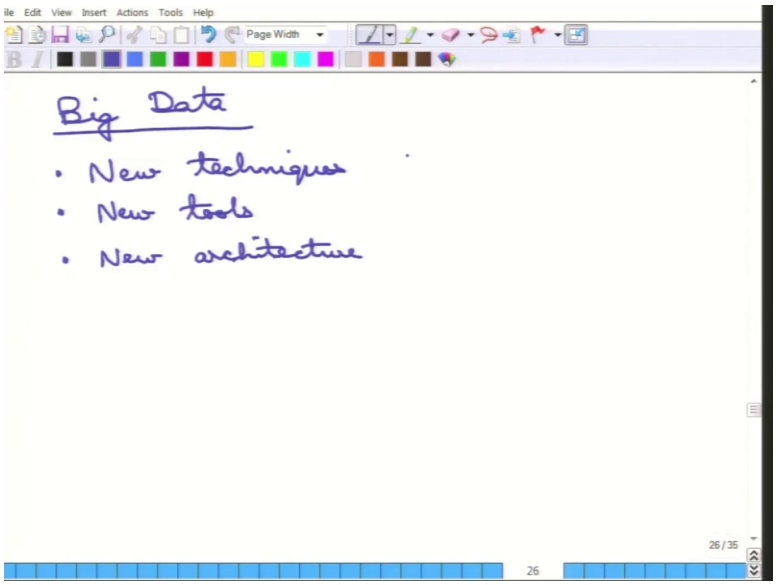
Fundamentals of Database Systems
Prof. Arnab Bhattacharya
Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

Lecture - 48
Big Data

We will start the last module which is on Big Data. So, big data is the term that I am sure all of you must have heard it.

(Refer Slide Time: 00:21)



So, what is big data? So, we will cover very briefly about big data. The first question that everybody has in mind is, what is big data, now big data very simply, data which is big. Now the question of course, is what does it mean, how big is big. So, for example, if you go to sociology, sociologists will do experiments, field experiments, will ask people questions and will take their answers and so on and so forth. For sociology, may be even 10 percents data is big, because it is very hard to get a complete set of 10 percents, we will answer all their questions honestly and reliably and so on and so forth.

So, for them 10 is big data, on the other hand you go to physicist, who takes astronomy. So, people who gets these images from astronomical telescopes and so on and so forth, for them terabytes of data is what they get in a day, single day they get terabytes of data. So, terabytes of data is, even terabytes of data is not big for them, so beta bytes and so on so forth is probably big for them.

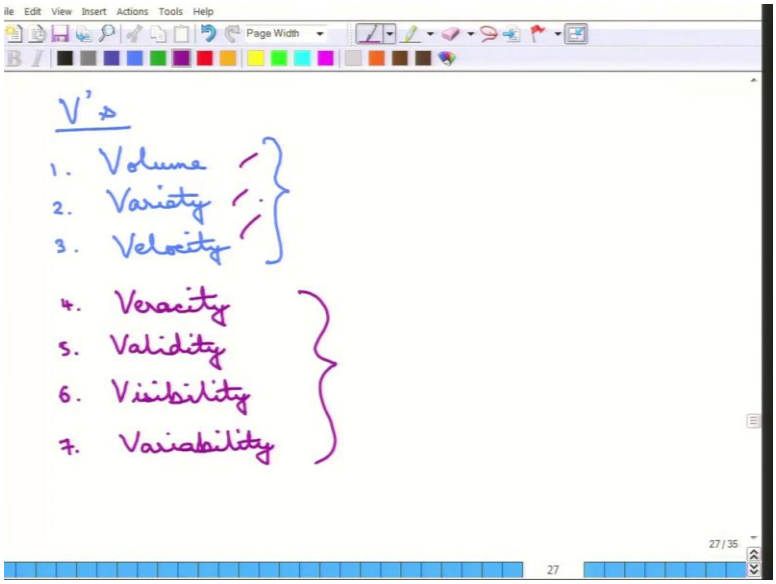
So, there is no threshold of, there is no very nice threshold of saying, oh after this, this is called big data and below this, this is not big data, there is nothing like. It critically depends on the application that you are talking about. So, once more there is nothing concrete about what big data is, it critically depends on the application. Generally, what is being done is that, what is considered big data is that, when the data or the algorithms that will run on the data, this cannot be handled by a single machine that is typically considered to be big data.

So, when that data is too big, that cannot be stored in a single machine or that cannot be handled by traditional algorithm, because it probably requires too much time, impractical amounts of time or too much resources that is being considered as big data. But, do note that this is all are very fussy definition, there is no strict crisp definition of what big data is, it depends very much critically on the application that it is talking about.

So, that is big data that, but and let us see a little bit of what the characterization of big data is essentially, now this large volumes of data as I am saying, the traditional algorithms may not be able to handle it. So, essentially you may require new techniques to just to store the data, just to query, handle the data. So, new tools, because the traditional tools such as the traditional RDBMS systems may not be able to handle and many new architectures of computing systems.

So, because single machines may not store it, so you will require probably cloud computers or super computers or distributed setups and so on and so forth. So, that may be necessitated and so the big data may require all of these things. Now, for big data what are the, there are certain properties of big data that is being done.

(Refer Slide Time: 03:19)



So, the three most important properties of big data are the... So, the big data is typical this properties which are called the v's, v properties. So, the three important v's are the first three important v's is volume. So, big data generally has a very large volume, so when it is very large, even how to load even all the data into one machine or into whatever, I mean there are system to load it, index it, query it etcetera that is a problem, this volume. Then it is variety, because big data may not always be pertaining to data that is of a single types, so it is not one schema or one cascade of things, then we will various kinds of things mixed in a big data setup, so that is the thing.

And it can be the data can be structured like RDBMS or semi structured like key value stores or completely unstructured like text, written memory I mean there is no structure almost no structure, so that is the thing and the velocity. Now, what does it mean to say velocity? Velocity is essentially the big data may be arriving at real time, so it may be streaming data, so it may be as the algorithm process it, more and more data keep coming. So, and it can be so big that the entire data cannot be stored and then it processed upon, it needs to be processed a faster, it need to be processed in real time in a streaming manner, so as data is coming in, there is some processing going on to it and that is being processed.

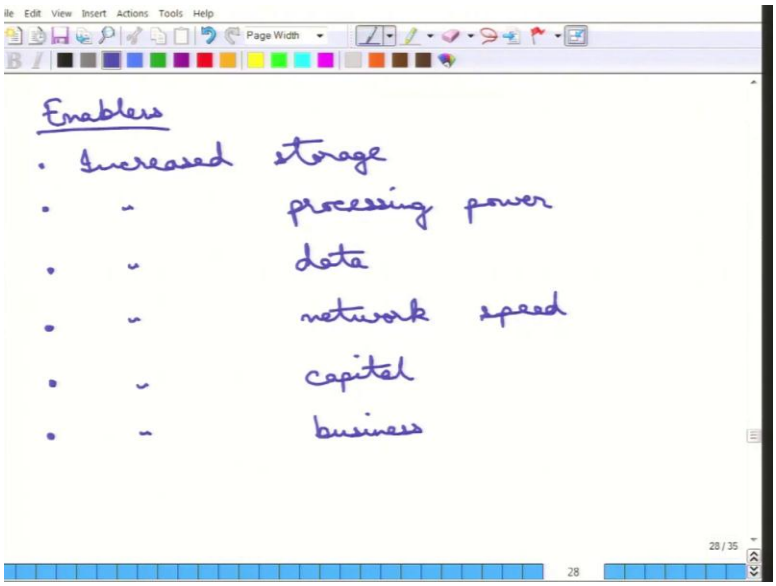
So, these are the three initial v's of big data, then as happens with any hypes transfer big data is very much hyped and there have been some other terms that have been associated

with this. So, veracity, veracity means whether the data that is coming or the authenticity of the data, the truthfulness of the data whether it is correct or not, the validity same thing whether the data is still valid or the data that one is processing big data has expired.

So, by the time I write a post into the face book and I delete it that post is no longer valid, so that is the validity. Then visibility, how does one visualize the entire data and whether it is visible to all parts of the system can everybody. For example, seeing my face book post and so on and so forth, it can even if all the machines want it and so on and so forth, the visibility is another issue and the seventh one is the variability.

So, how much variety can this big data handle in the single setup? So, how much can it be anything or can it does it need to pertain to certain kinds of structure at least, certain kinds of rules at least. So, these are the issues with big data, so these are the three I mean the volume, variety and velocity are the three initial v's and then there are many other v's that has been talked about.

(Refer Slide Time: 05:59)

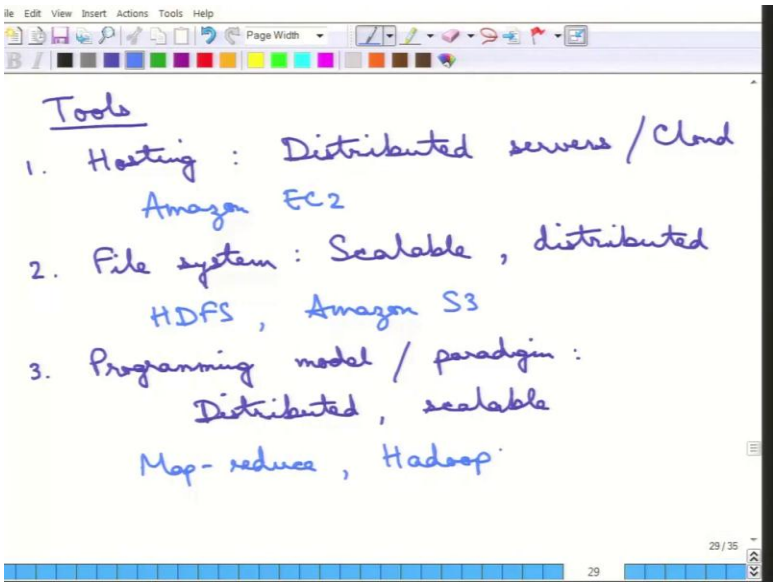


Now, to for big data systems to happen we require certain enablers, enablers meaning certain things have to fall in place, so the big data systems can be talked about. So, first thing is that it requires increased storage space, so suppose terabytes of data are coming in a day. So, you will require beta bytes storage that so 1000 terabytes is one peta byte. So, you will require that kind of storage, so there must be those hard drives or those systems must be made, so that is the thing, so this must be available.

So, increased storage volume and increased type of storage, then increased processing power. So, we will require faster and faster machines, more and more better and better CPU's, etcetera just it is very understandable and of course, increased data. So, big data is all about this data, so more and more data is being produced and that is why it is called big data, then many times the data is available only over network. So, increased network, speed or network capabilities, whatever you want to say, because data may be streaming in from different networks and you may send it to multiple database, because it is a distributed system, so it must be sent to different places, etcetera.

And of course, capital or I mean increased capital essentially money, we require money to store all of these things. So, all this is required to do big data things and well why will you do big data, there must be increased business. So, the last two things are essentially trying to say that if you wants to do a big data something on big data some company wants to do big data. So, it must get the profit out of this to... So, that it will be it is worthwhile to engage in this big data principle, so that is the thing about big data and these are the enablers of big data and then there are certain tools for big data that one has to talk about and that is it.

(Refer Slide Time: 07:52)



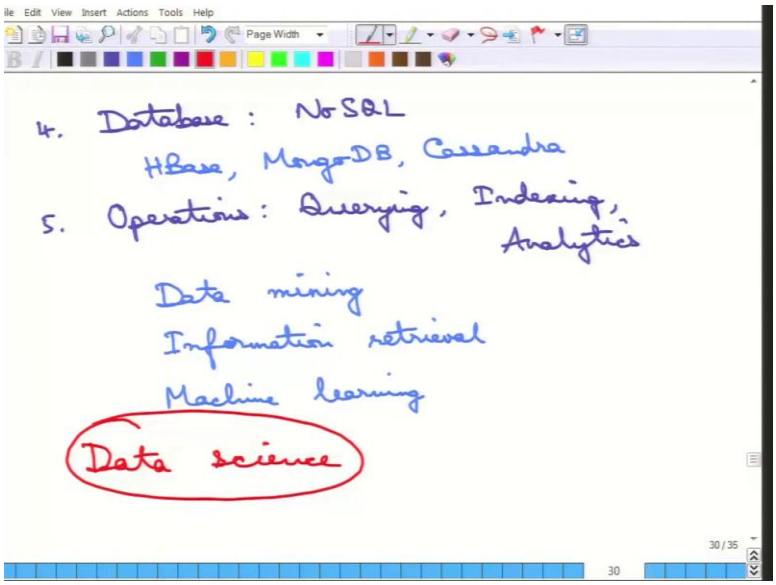
So, there are certain tools that are already out already available, so tools for big data. So, the first thing is that for hosting big data there are this clouds. So, the cloud the distributed servers or the cloud that are being... So, these are already available, so this is

the famous cloud. So, all of you must have heard this term cloud computing and again that is not a very crisp term, but essentially cloud it is a distributed server. So, one can store data and there are examples is that your Amazon easy to that let us you store lots of data and of course, you have to pay by it is not free, but at least let us you do that.

Then the file system, there are certain tools, there are certain file systems available which will let you do this thing. So, in the file system the properties of this file system is that this must be scalable and they must be distributed, the file system itself must support this data, a traditional file systems may not be useful. So, again HDFS the file system this thing that supports this is already available and then there is a Amazon has it is own file system called S 3 that is there then there is a programming model.

So, the programming paradigm itself changes, because big data cannot be handled by traditional algorithms may be. So, that is why the programming model or the programming paradigm needs to change, so this has to be more scalable and distributed. So, this is again distributed and scalable, so this the new programming module itself has to have this constructs built into it, it is not it must support this naturally and the map reduce frame work, as I have been talking about map reduce frame work will handle this and Hadoop is another tool that let us one do all of these things.

(Refer Slide Time: 09:51)



Then it may require of course, it will require database support. So, these databases may need to go beyond RDBMS and as we have been saying this essentially this is all

clubbed under this NoSQL term as we saw earlier. And these are examples are of course, your H base then mongo DB and Cassandra and all of these things and all of these are enablers for big data. So, these are tools for big data then this will let you do certain analysis on big data that is all and then there are finally, this is operations.

So, operations are your, what are the kinds of things? So, it must let you do... So, tools for querying, tools for indexing, tools for doing analytics. So, one has to do run analytics and this O lap etcetera these are all coming from that querying indexing and analytics and then there are... So, for this there are this data mining, data mining there are different data mining tools etcetera there are different...

So, essentially the entire data mining is ((Refer Time: 10:58)) because of this big data things, because it has to have this operations doing for this thing. So, data mining then information retrieval these... So, these are all generic terms of course, and these are not particular tools that I am mentioning, because there are many, many tools in these spaces. But, all of them essentially can handle as enablers as tools for doing big data.

And machine learning of, so machine learning. So, one particular example of machine learning which is interesting is that there is a mahout tool which is built on top of Hadoop. So, it can do certain machine learning algorithms directly with Hadoop database, so with the Hadoop this thing. So, there are many open source tools, the many of them are open source and are free and some of them are not everything is free cloud for example, ((Refer Time: 11:51)) etcetera, etcetera are not, free quite costly, so we have to store this.

So, one has to take look at the application to see whether actually big data is suitable term for that whether it makes sense to have big data for that the application must clearly define whether what it is big data or not and just saying that my application has large amounts of data. So, I will use big data tools or big data application that is not fair sometimes that is not correctly actually sometimes traditional algorithms can do quite well.

So, one has to clearly define and see whether big data is doing all of that to handle all of these there is a new paradigm that is coming out which is there is the new term that is coming out which is called the data science and there is the, there are data science

courses and there are data science things that is there. So, essentially to handle all of this big data cloud computing all of this things weather, so that is the emerging term.

So, again currently just like NoSQL big data is not clearly understood, what is big data it depends critically on the application and it is way to hype, but these are certain enablers there are certain tools that can be done for that. So, that ends the thing on the this course that ends all the lecture modules on the course. So, I hope all of you have enjoyed this course, I hope all of you have learnt something from this course, you have done the assignments honestly and correctly and have learnt something from the assignments and the slides have been put up and the assignments solutions have been given the final certification exam, if you are taking do study for it, it is not going to be easy.

And, but I hope you all pass and we will get to know and we will have a rough idea at least of what can be there from the assignments and at the end of this course and at the end of all certification exams, you will have a quieter good grip on what database systems are all about the UG database systems. And you will be very much comfortable and confident about handling issues for databases.