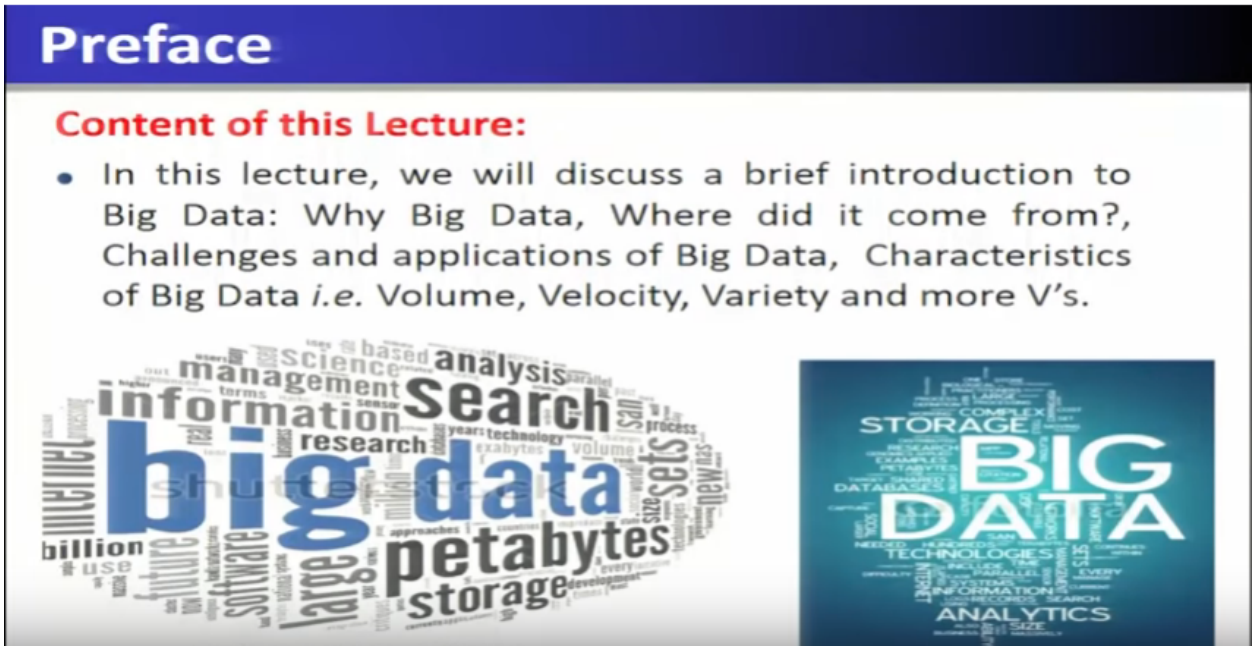
Lecture – 1

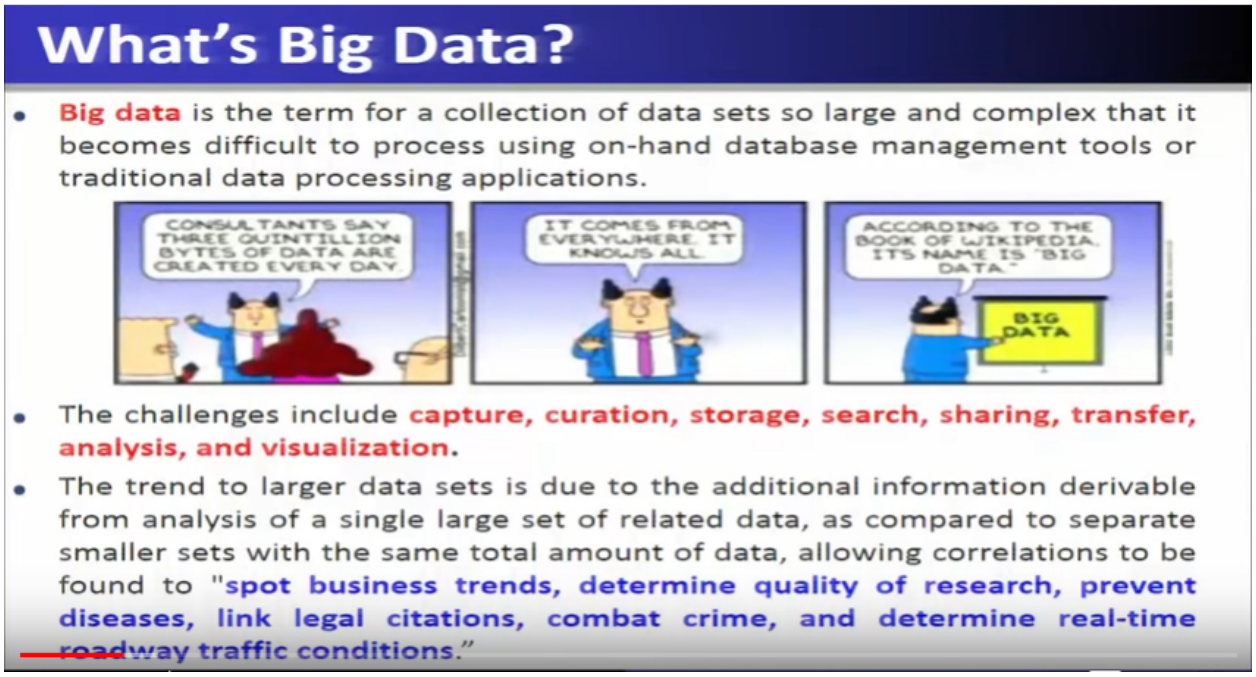
Introduction to Big Data

Refer slide time: (0:16)



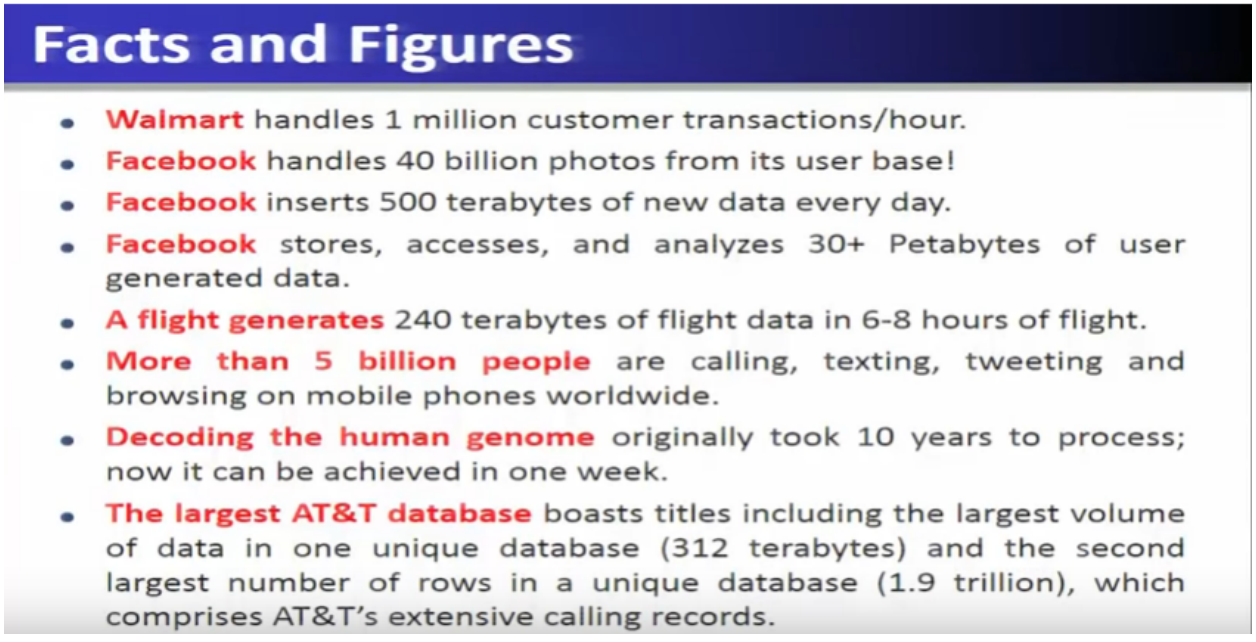
Preface, content of this lecture, in this lecture we will discuss, a brief introduction to the Big Data, why the big data, where the big data comes from? The challenges and the application of Big Data, the characteristics of the big data, that is in terms of volume, velocity, variety and many more V's. We are going to see in this part of the lecture.

Refer slide time: (0:43)



What, what is a big data? so ,big data is a term, for a collection of data sets, so large and complex that it becomes often difficult to process using traditional data processing applications .now, to see this particular picture here, a consultant is saying, that there are three continents byte of data, are being created every day in an organization. So, it comes from everywhere it knows all and according to the book of Wikipedia, its name is the big data. So, this is a simple way of explaining a big data using, this particular picture which represents only one aspect that is called volume or a size ,which is very big, we will see more such challenges in terms of big data in this particular lecture. So, such a huge volume of particular data, poses various challenges, which includes how to capture such a big amount of data how do you cure it? How do you store such big amount of data? How can you search? And how do you share this information? And how to perform the transfer of this huge volume of data? Also we will see, other further challenges like doing the analysis, analytics and its visualization, which is often Lee they are useful too for many applications, where this big data is going to be used. Now, the trend to this larger size of data sets, is due to this additional information, which are derived from the analysis of a single set of large related data as compared to the smaller sets, with the same total size of the data ,this large size will allow the correlations to be found ,to various opportunities to exploit in terms of spot business trends, to determine the quality of research, to prevent the diseases ,to link the legal citations, to come back the crimes and determine the real time roadway traffic conditions .so, given this particular a big data or a large size of data, and it will pose various opportunities and challenges and new, kind of applications, and also social service and is possible, that we are going to see ,that is why the big data is going to become popular.

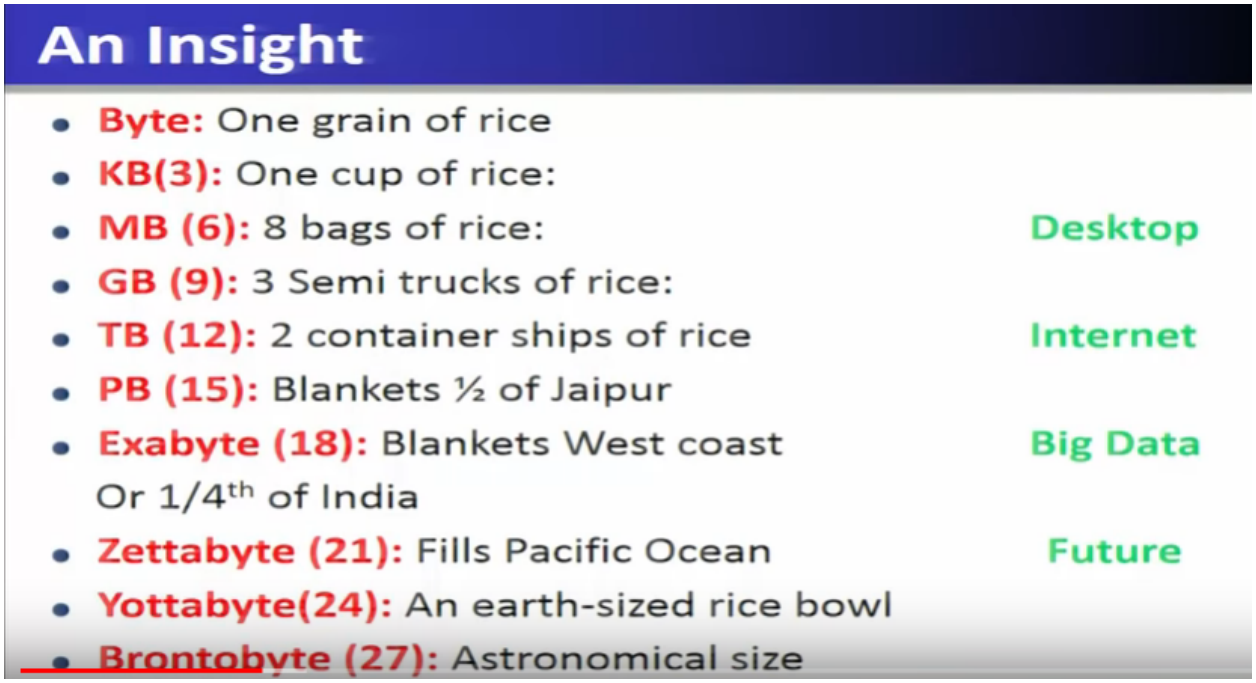
Refer slide time: (3:52)



Some of the facts and figures, of this particular size of the data or a big data. So, we are going to see into it for example Walmart is a company, which handles 1 million customer transactions per hour .so ,just see that here it deals with the volume or the rate, in which these transactions or the customers are handled, the second such company is which is called a Facebook ,which handles 40 million, 40 billion photos from its

user base. So, when you say photos that mean now here, the data is in different form and also of large size. So, to dimension of the complexity is being added now, Facebook here the inserts 500 terabytes of new data every day. So, this basically becomes the volume challenge, Facebook stores, accesses, analyzes, 30-plus petabytes of user-generated data, every day. Now, similarly a flight generates 240 terabytes of flight data in 6 to 8 hours of flight to make the customer safe, flight and also to basically ensure the, the comforts, during the flight journey. So, that is why this particular flight generates and uses this information for the analysis and providing the solutions, similarly more than 5 billion people, are calling, texting, tweeting, browsing, on their mobile phones, worldwide. so ,here the people are involved in generating the big data. Another thing is about, the decoding the human genome. so, originally it took 10 years to process it. Now, it can be achieved in one week .that means the computations of a big data is now becoming possible, to be completed within their time now, another company which is called AT&T databases, which boots, the titles, including the largest volumes of data, in one database that is 312 terabytes and the second largest number of rows, in a unique database that is one point nine trillion, which comprises eight ents extensive calling records. So, this particular company is this example, which uses the large size databases ,that is the data which are at the store and then it performs ,it has to perform the computations on this large sized data set, to gain the insight and basically drive the, the business of that company.

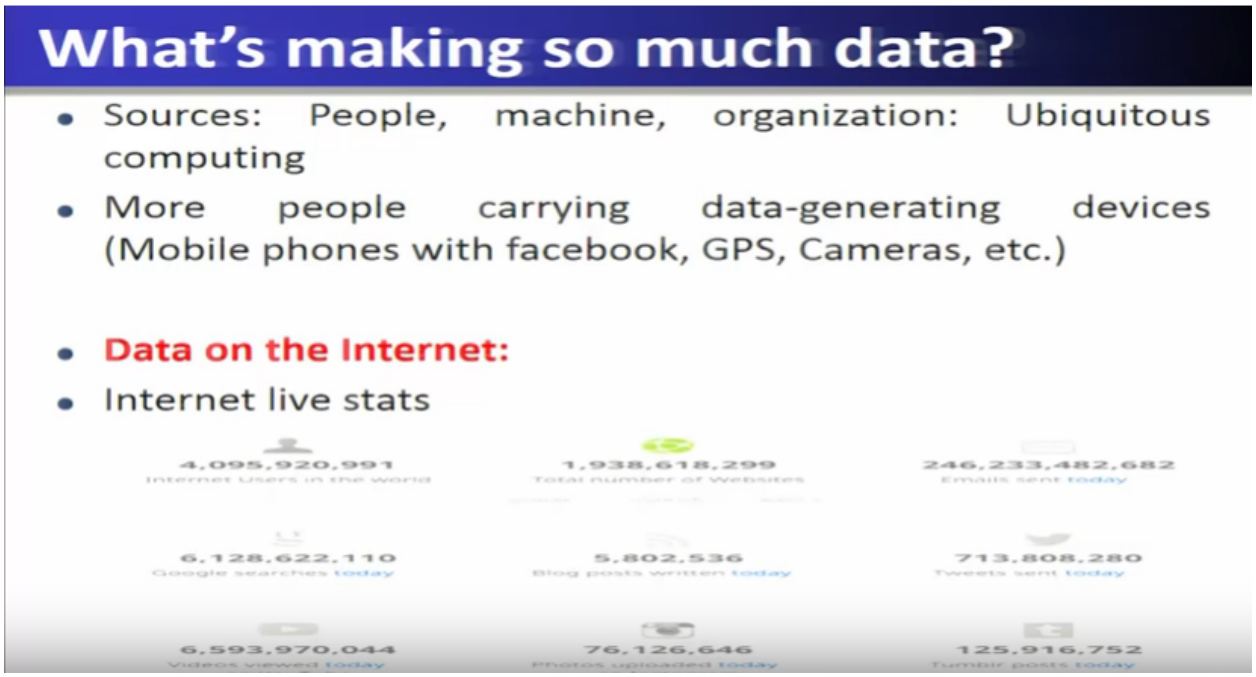
Refer slide time :(6:51)



Now, let us see the volume, with an insight that one, if we consider that the abite is a one grain of rice and a kilobyte, that is 10 to over 3, is a one cup of rice, then megabyte which is 10 to power 6 becomes 8 bags of rice and we can see the gigabyte 10 to power 9, which is nothing but, we can extend it and we can understand that 3 semi-trucks of rice is 1 gigabyte. So, 2 container ships of rice will become 1 terabyte that is 10 to the power 12 ,and it represents the, the amount of information which flows on the internet ,petabytes which is 10 to power 15 is the blankets half of our city Jaipur, except ID which is 10 to power 18 that size is called a big data and here it comprises of or we can visualize the size as 1/4 of this blanket,

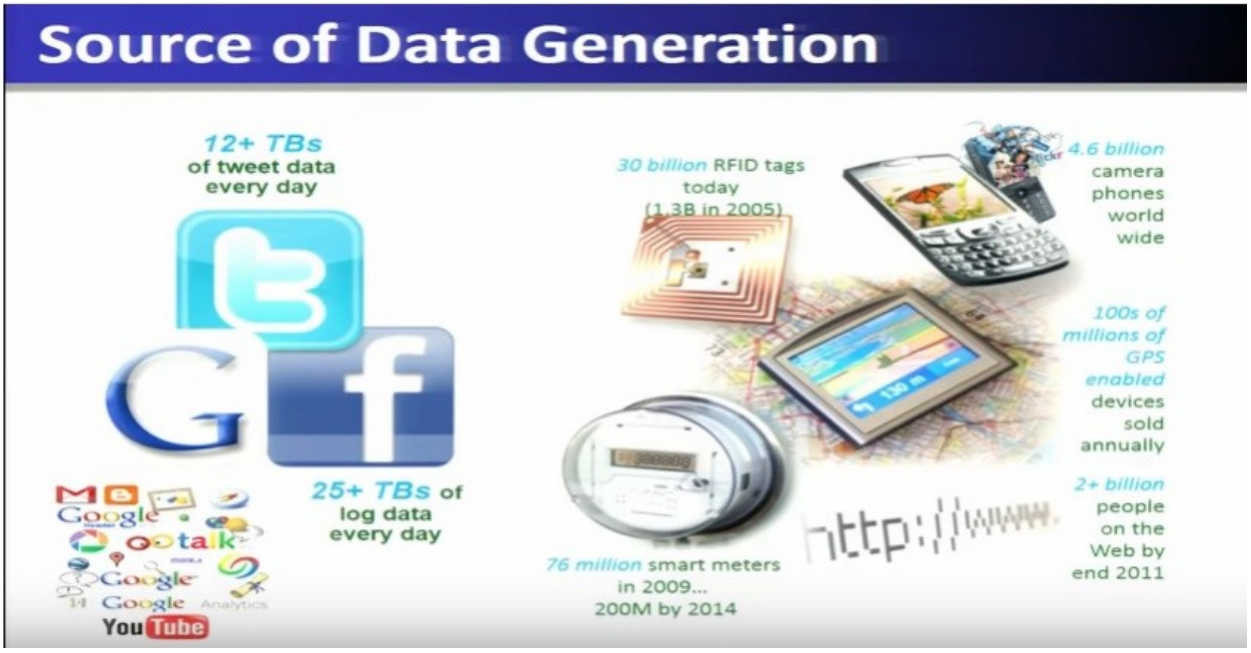
which are there in the country, and zettabyte which is 10 power 21 and this basically will fill the Pacific Ocean and that amount of data, which is called a zettabyte, is a future volume of the big data, similarly keep on extending zettabyte becomes, zettabyte that is 10 power 24 which becomes ,an earth-sized a rice bowl. And beyond that it's a brown to bide that is 10 power 27 that becomes an astronomical size of that particular data. So, we are going and moving towards this kind of huge volume of data, which is of astronomical size, how to handle this kind of data is called a big data computation. And we are going to see these particular entry cases in this part of the course.

Refer slide time: (8:35)



Now, what's make so much of data? Now, here we consider there are three different sources, which make or which contributes to this so much of data? the first one is called people, for example you might have seen the Facebook or a people carrying the mobile phone, all the time they are generating the data either in the form of a text, in a Facebook or a GPS ,when mobile is being carried or basically cameras taking pictures so ,these kind of data which is being generated ,by the people ,another type of data source ,which generates a huge volume of data is using sensors. So, sensors are normally deployed in smart city organizations or in the industries or in many places they keep on generating the time series data. And the third type of data, third source is called organizations .so, organizations normally do the transactions, of other services, which transactions also and the customers transactions all these will become the source of data out of this organization and this is all together will form a ubiquitous computing .so, the data on the internet basically sometimes requires, to basically do the analysis over the live statistics, that we are going to see here, in this part of the lecture.

Refer slide time: (10:07)

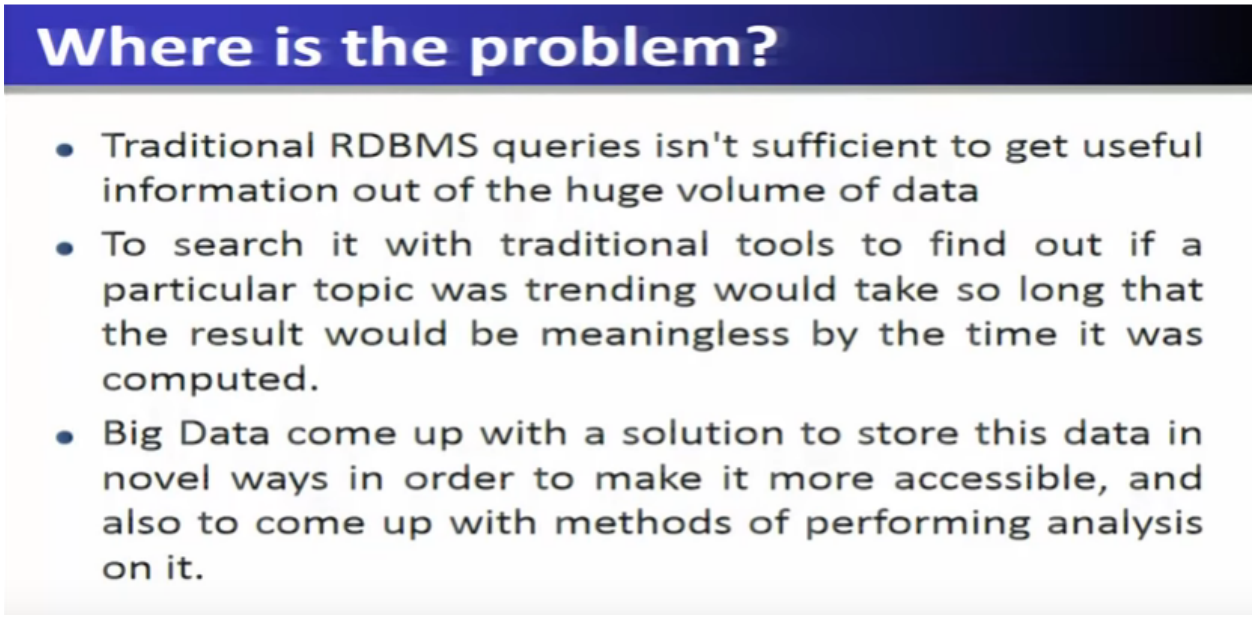


So, as I told you there are three different sources, one is the user. So, users which are basically using, the services or Facebook, Twitter, Google, you see that they are generating lot of data, and that basically is one of the sources of big data. The second kind of data source, which you see over here, is that the devices are, basically high devices with a sensor they are generating lot of data. For example, smart meters are generating data and RFID tags, in the objects, they are generating data and this camera which is there and sensors, which are there inside mobile phone, they are generating data. And also all the devices nowadays, equal with the devices which are called IOT devices and the sensors, they are continuously generating the data. And also two plus billion people on the web and this particular size of the web also are contributing enormous amount of data. So, there are all kind of sources which are now, generating the data and this becomes a big size of data.

Refer slide time: (11:25)

Another example of a big data at work is, our using crowdsourcing .so, using Crowdsourcing this particular data ,will be taken up or captured and perform the computation on it ,to basically gain the insight of the traffic congestion on the roads are doing the, the sensing ,where let us say an ambulance is moving and it requires a green path, and also it will perform or it will give a route, on the map and the routes are computed using this particular situation ,which is dynamically changing at a particular time.

Refer slide time: (12:11)



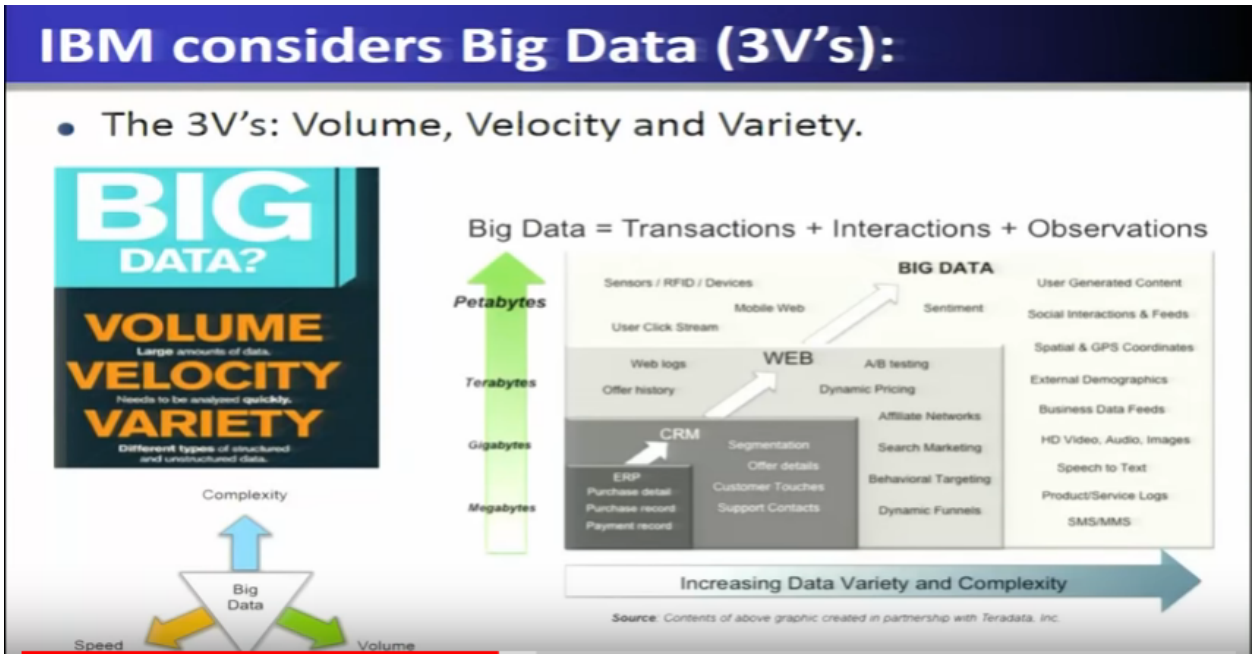
So, where is the problem In this particular entire landscape of a big data. now ,we see that in traditional RDBMS these queries isn't enough sufficient to gain useful information out of the huge volume of data .that means traditional RDBMS queries are insufficient to handle this kind of big data and to gain the

insight. To search it with traditional tools to find out if a particular topic was trending or take so long that the result would be meaningless by the time that means it requires a real-time computation or retrieval of that particular data, which is required at a particular point of time. And the traditional RDBMS operations that are that retrieval are slow and it is not useful for many of the applications. So, we will see in this particular part of the course, the remedy or the solutions which big data highs, regarding storing this information and providing the retrieval at a much faster speed and also newer methods which can perform the analysis on this kind of big data, at lightning speed.

Refer slide time: (13:34)

So, the challenges are summarized what Here, which are basically again around using the big data for different application is about capturing, storing, searching, sharing, analyzing, and visualizing.

Refer slide time: (13:49)

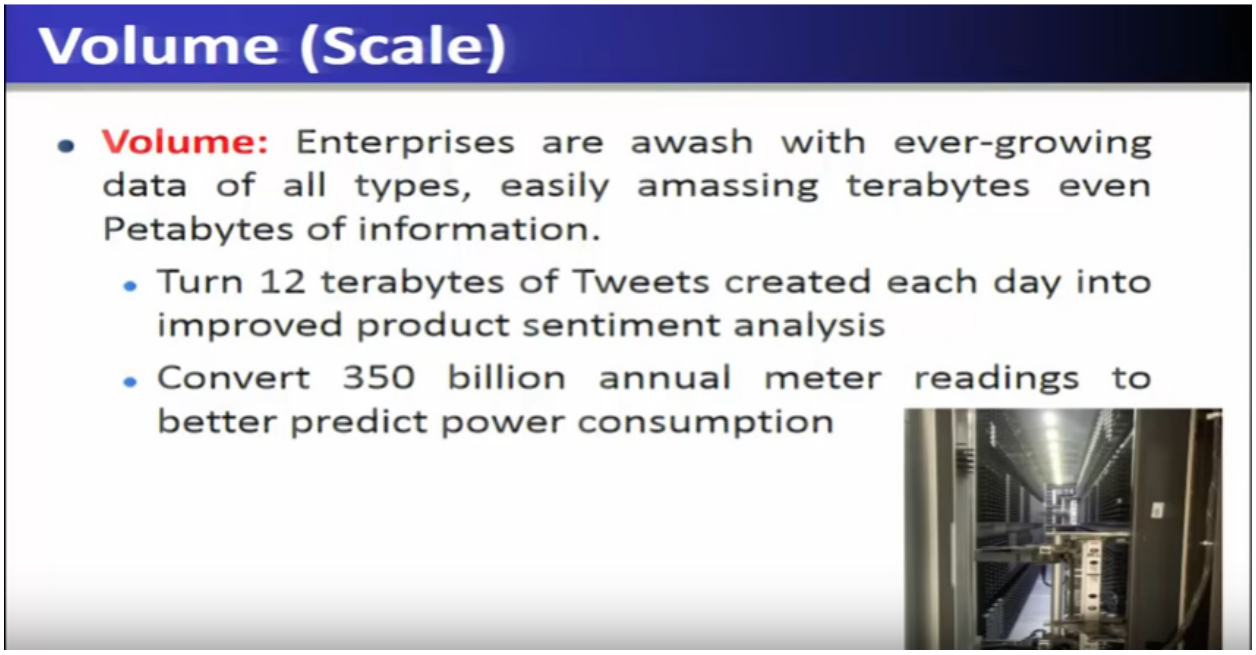


So, IBM and Gartner together, they consider the big data has three different 3Vs so the characteristic of a big data is given as these different 3V's. So here, gotten or explained or IBM also considered three, most important 3V's so, first three V's which characterize the big data are volume, velocity, variety. What are these? And how it is going to signify this particular data as the big data? So, the characteristics of these particular features, that is the volume, velocity and variety .will characterize the data as the big data .and we will see here, that when we say that it's the volume that is the size, that is beyond petabytes ,which will, which we have already seen ,if that size is there then basically it enters into the big data domain and ,and similarly if let us say the variety means, the data is not only the text data but it is in the form of the images, videos ,3d objects and so on, then basically there is a huge not only the size but also the data variety it's another dimension of the complexity .so, and another dimension which is called basically the velocity, that is the rate at which these data which is being generated, has to be tapped and a processed so, this becomes the speed or the velocity .

so ,the three means together, they basically will characterize the data as the big data. So, big data will be in the form of transactions, in the form of interactions or in the form of observations or together generating the large size of data that is data sets which need to be analyzed. so ,here we see that let us say that Big Data ,scenario using this big data they are doing sentiment analysis to understand more insight about the entire centric, customer centric businesses that is the sentiments, when it comes there are sentiments two types of sentiment ,individual sentiments, that means if the business is targeted to basically for a particular individual or basically the entire customer base it's not individual but it's the entire population and the businesses are trying to understand the sentiments and plan the new businesses, which are basically possible. So, and similarly sensors RFD, RFID and different devices, they also generate lot of different data, similarly user click streams also generate lot of data, these particular data will perform the mobile web and will be analyzed in the real time, and to ,to gain various insight for example, to understand the traffic condition situation in, in a smart city or, or to basically deal with the disasters. For example, if there is a fire at one place or is being triggered, triggering the fire so, it has to be controlled that is called disasters. So, all these that means the big data the mood inside has to be gained

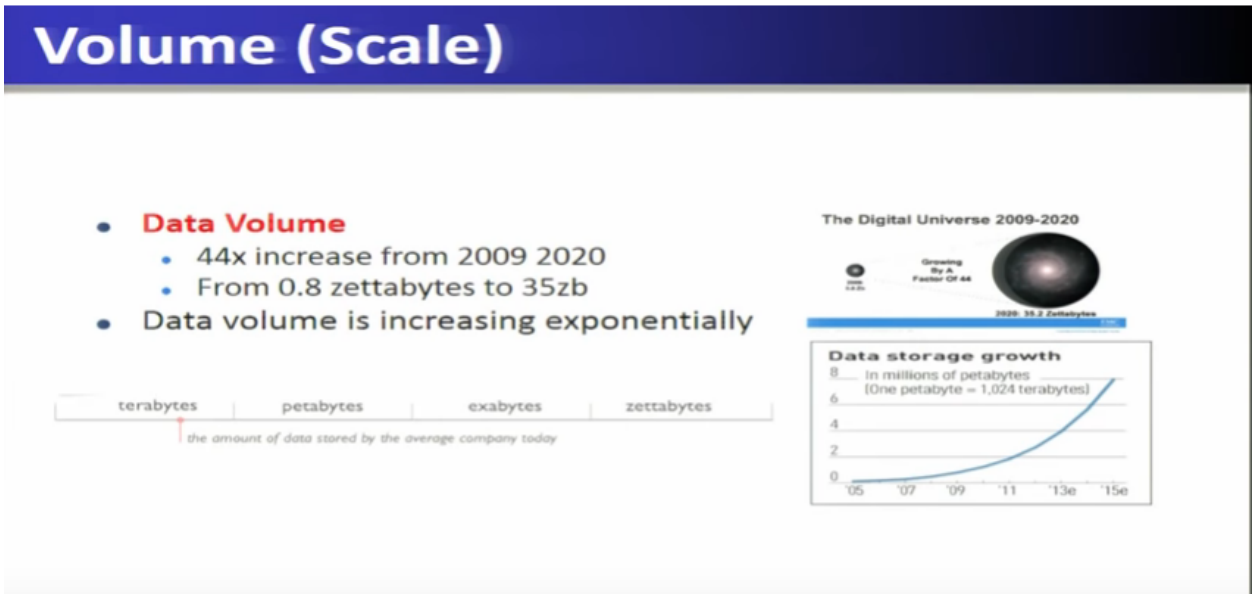
and it is better serving the community or basically the masses. So, that is why the big data is here, and it is becoming popular day by day.

Refer slide time: (17:34)



Now, let us see in more detail of these characteristics. So, the first characteristic of a big data is called volume and this is nothing but called scale. So, the enterprises are basically our growing and generating the data of all types and the size typically goes beyond terabytes then we'll be categorized as a huge volume and that basically is require different technology, which is called big data and for the computations. for example, if there are 12 terabytes of tweets, which are created every day, and which need to be analyzed, to for the sentiment analysis .so, this sometimes is a big data problem, similarly 350 billion, annual meter readings in a smart meter, to for the analysis, to predict the power consumption again becomes a big data problem, where the volume is involved to be undersold using big data problem.

Refer slide time: (18:45)

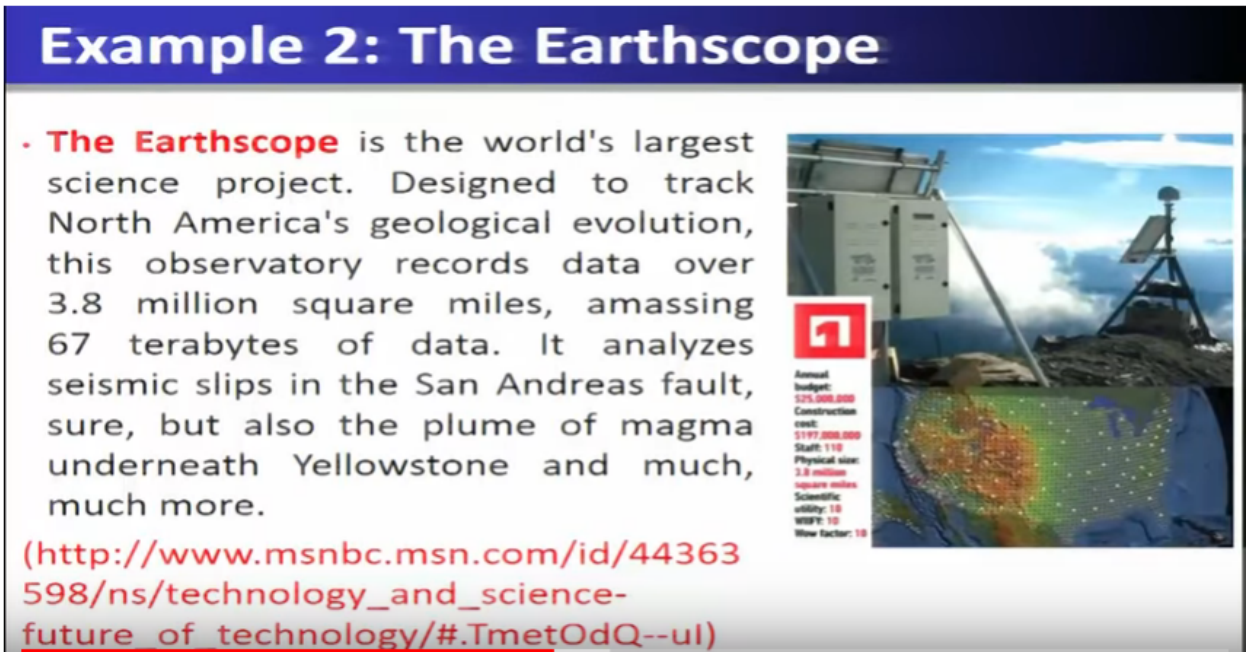


And Applications so, here we see that 44 times increase, in the volume of that particular big data is increasing from 2009, 2020 and also that is from 0.8 zettabytes to 235 Zeta bytes. So, data volume is increasing exponentially and it requires a Big Data Platform to be used in this particular considerations.

Refer slide time: (19:16)

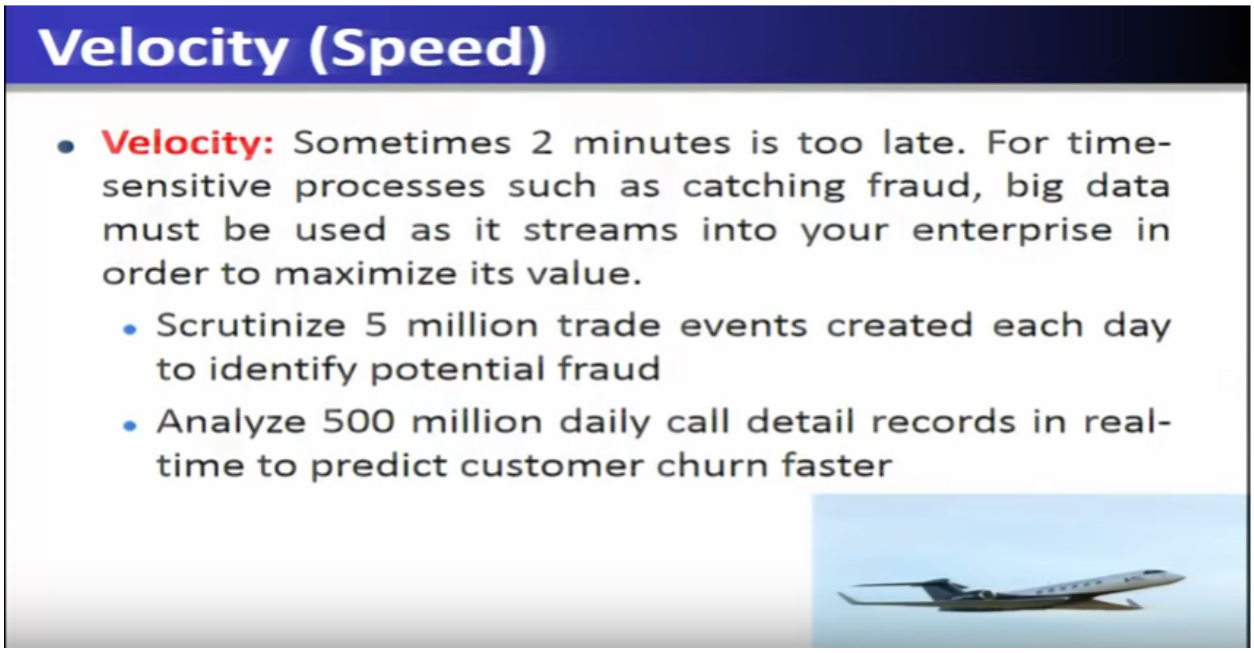
Another example, of generating the huge volume of data is the CERN's Large Hadron Collider, which generates 15 petabytes of data up here.

Refer slide time: (19:27)



Another, source of generating big data is the Aarthi scope, and here 67 terabytes of data is being generated and is being analyzed

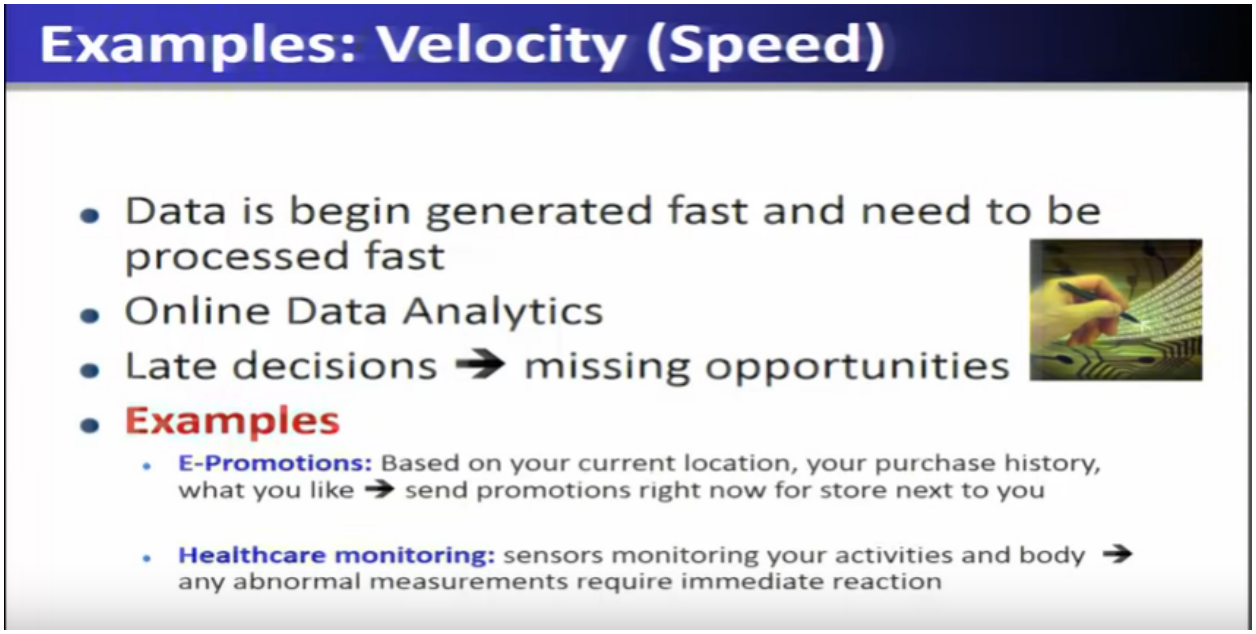
Refer slide time: (19:43)



Now, the next characteristics in this sequence is called velocity, that is called a speed, sometimes 2 minutes is too late. So, basically that shoes, that reflects, that here the time is a factor and everything has to be done within a particular time bound and this particular aspect of computation is called the velocity

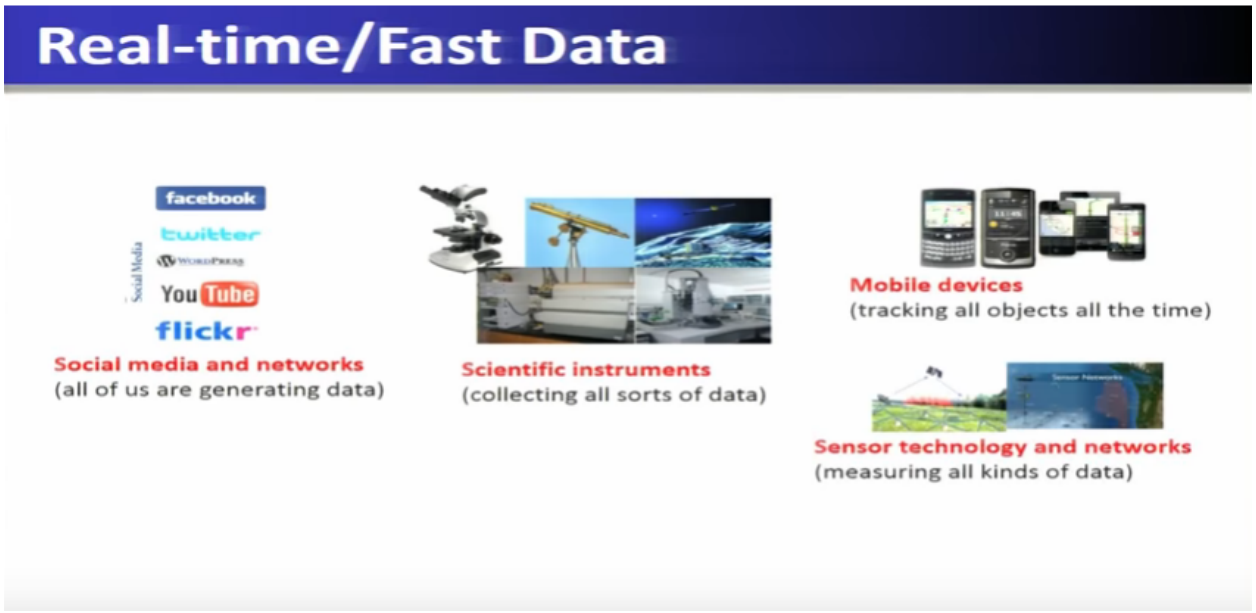
that means the data which is generated in real time it has to be analyzed and understood about various purposes. For example, if you want to catch a fraud, for an online transaction, then the entire transaction has to be analyzed and being detected whether it is a fraudulent transaction or it is a normal transaction, at that speed. So, big data must be using, the stream data in this particular scenario. so ,velocity is the stream of that particular data which is basically flowing, out of that the applications which need to be analyzed and being used in applications, similarly we have to scrutinize five million trade events which are created every day to identify the potential fraud ,similarly to analyze 500 million daily call details in a real-time to protect the customer churn at a faster or not. So, these kind of applications are now, driving the different companies or the organizations and to retain the customer and also to run the businesses in future. so ,obviously this aspect of a big data is also very much needed and volume, and velocity together, is creating the challenges in this big data computation.

Refer slide time: (21:39)



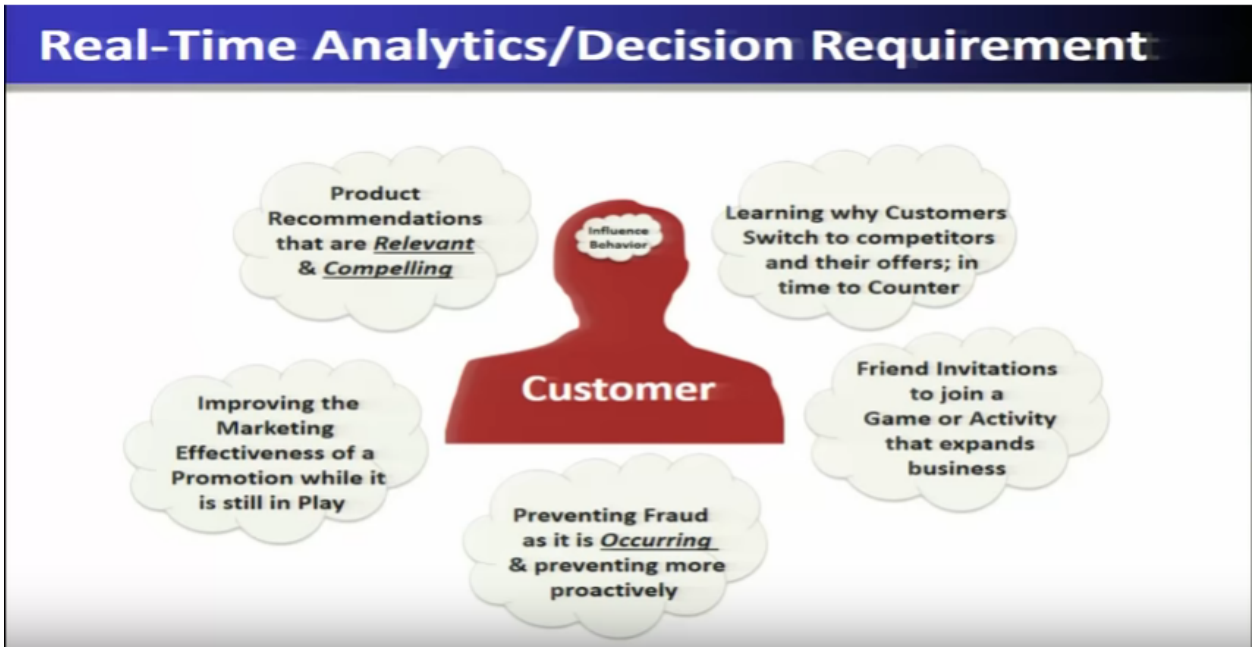
So, this we have already understood that Data, indicate in the velocity, means that I generated at a very fast pace and which need to be processed at that speed. Now, another use of this velocity is about online data analytics and here, if you miss or if you do this analysis, rate that means you will be missing the opportunity because, these operations are doing in real time and real time in some deadlines are there after that ,that decision is of no use. So, late decisions will employ the missing opportunities and this means that the velocity is in effect, at for that application. So, the examples of such cases are, like E promotions and healthcare monitoring, where the sensors are monitoring, your activities and the body and being alerting you for any abnormal measurements, which requires an immediate attention or a reaction.

Refer slide time: (22:48)



So, this will give the real-time or basically the first data and nowadays, the social media and the networks are contributing in this particular dimension and has to be not only captured but, need to be computed in real time all this data similarly for the scientific instruments mobile devices, sensor technology and networks .

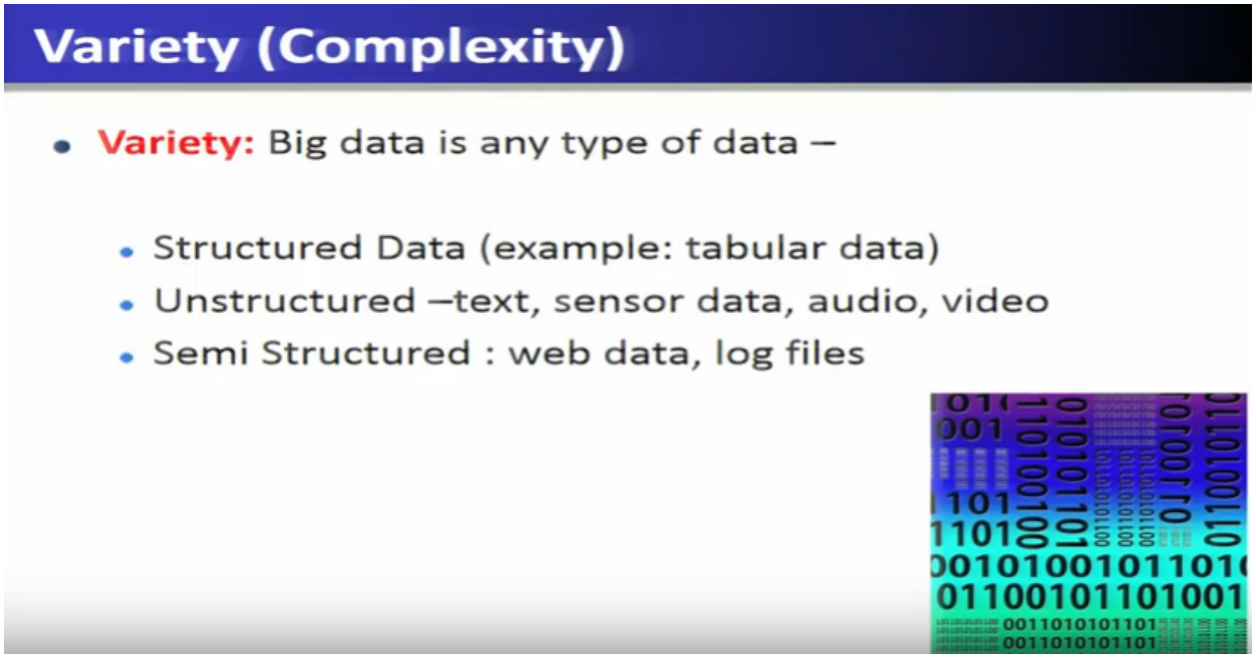
Refer slide time: (23:20)



So, this will also be very much requiring this kind of dimension that is the real-time, analysis or the decision-making. So, in most of the businesses, where customer centric decisions are to be taken, that means to give the product recommendations and to learn why the customers are basically making or turn out of that business or how the friend invitations are being sent to join together, which will basically be in the form of gaining, more businesses similarly, how to prevent the fraud? And, how to improve the

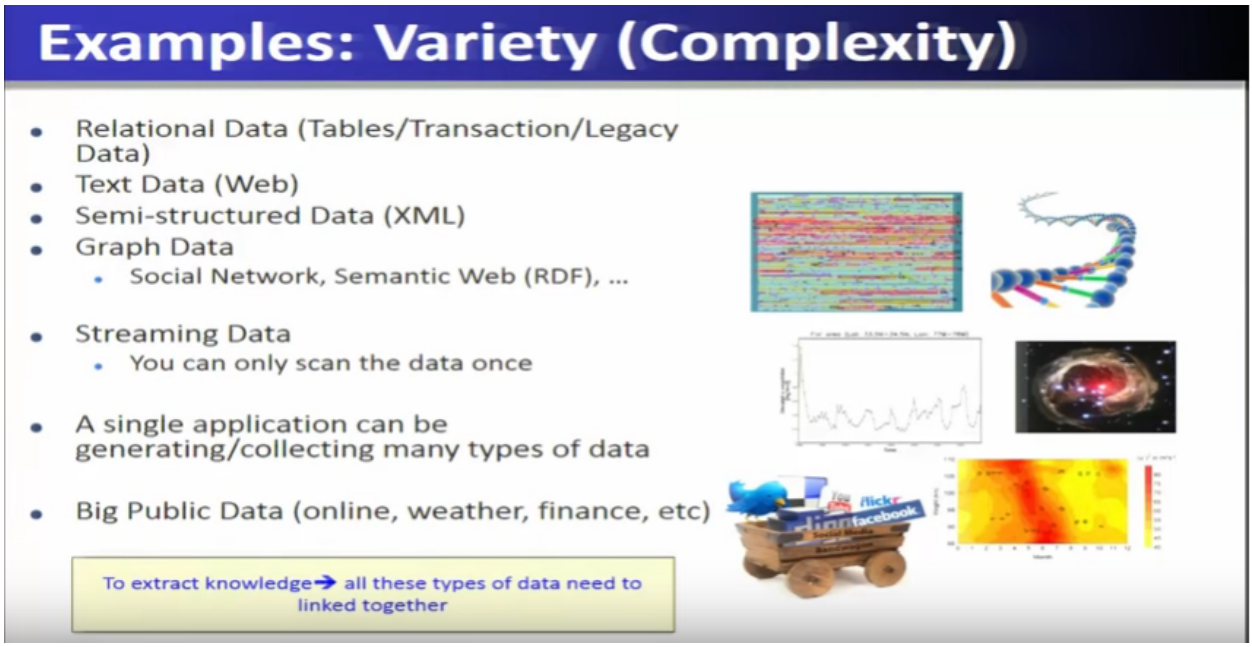
marketing? It is all customers centric to understand the behavior or sentiment of a customer and do the real-time analytics, and this will be a very good way of running the business and every business has to basically, be a customer centric to drive it further. So, real-time analytics is very much required and the decisions are being used.

Refer slide time: (24:33)



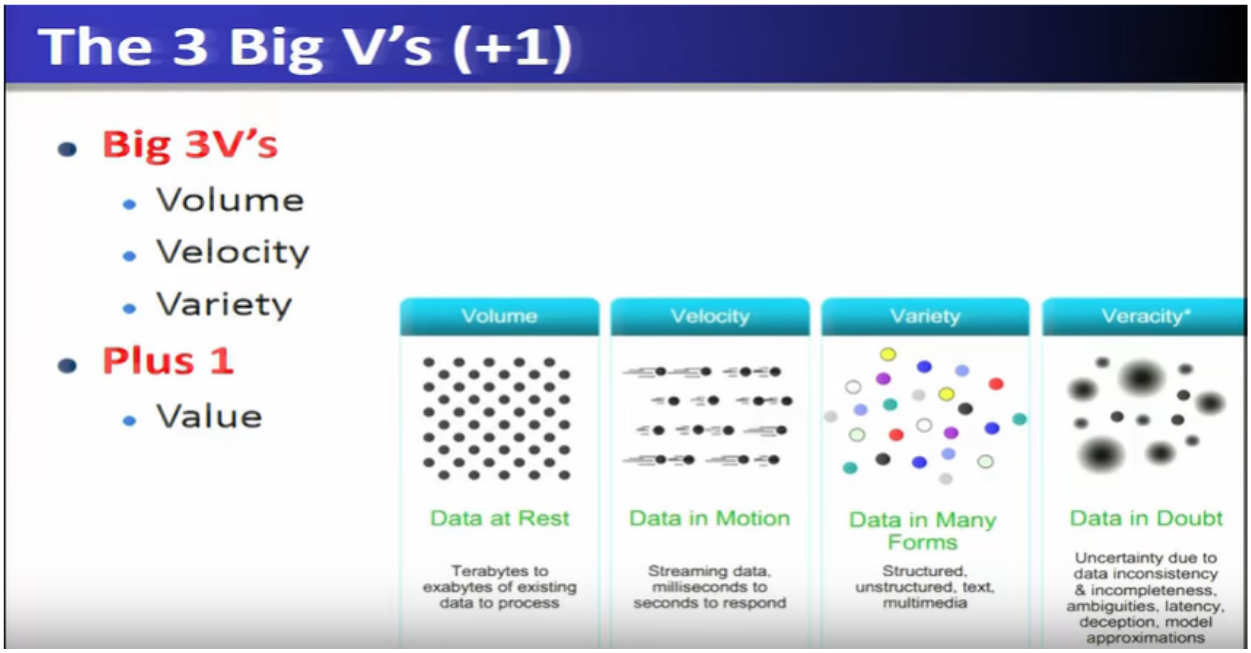
Next dimension is called a variety and which X to another level of a comp, which is called complexity. So, variety means the data ,big data is not of one form but, of several type of forms of big data, comprises all for example, the structured data when it calls, when it basically is perceived is, the data which is stored in a form of a tables. Unstructured data which cannot be stored completely in the table or form, which is called unstructured data .which, is basically the text sensor data audio and video, there are different type of data, which cannot be termed as the structured data that is lot of variety is there in the data becomes unstructured. Semi-structured is for example XML. So, web data, which is captured in the form of xml, forms a semi structured data, all these different variety, structured, unstructured and semi-structured data, will basically deals with a complexity to the big data, which is called the variety.

Refer slide time: (22:40)



Examples of this variety dimension into the big data is ,the data which is coming out, of the real time data, out of the transactions tables and legacy data, then the text data which is on the web and semi structured data that is the XML data ,which is being captured out of the web, graph data which is nothing but, a social network data Semantic Web, streaming data you can scan the data once and the public big, public data which is available as online or a weather data or a finance data and so on together, these different variety of data will add to different complexity in Big Data computation but, very much needed for decision making into an organization. So, extract the knowledge, out of these varieties of data means that, all these type of data need to be linked together or correlated together and gain the meaningful insight, out of these correlated events or activities.

Refer slide time: (26:53)

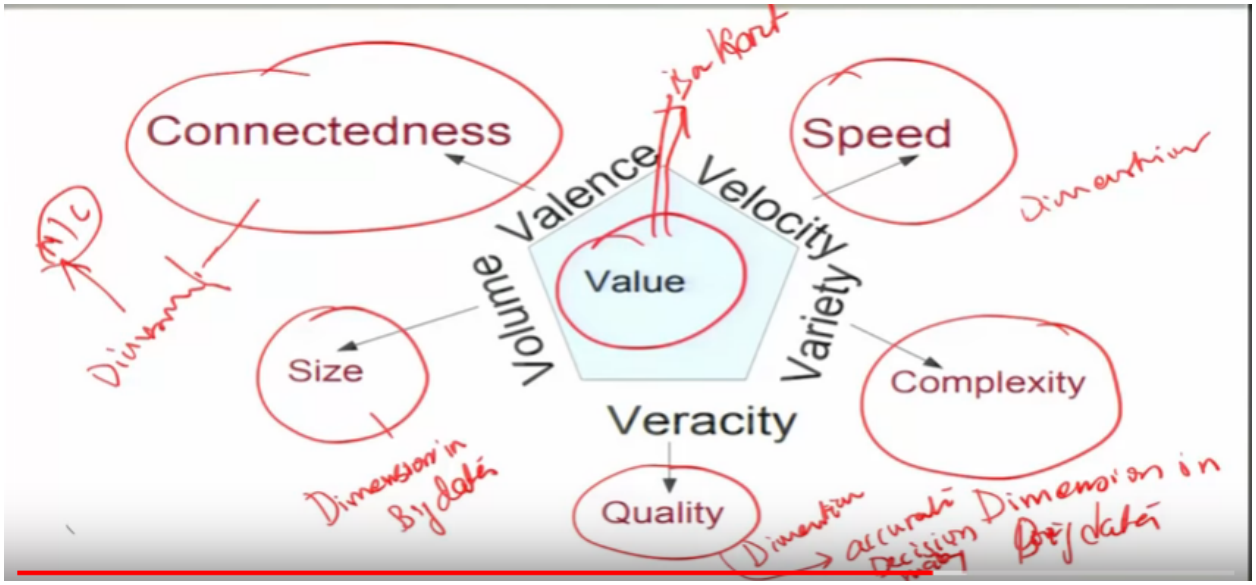


So, therefore we can summarize here ,the volume that is the data, at rest that means the terabytes or to the exabytes of the existing data need to be processed and this becomes the one V that is three V's of a Big Data, one of the three V's of the Big Data. the second one is called the volume or the velocity, velocity means the data which is there in the motion and this particular data is called a streaming data and this basically varies from milliseconds to the to the second to respond and this rate if it is the constraint and it becomes the velocity that data is called first data. Third type of data which is called, third type of characteristics which is called? the variety ,that means data is in many forms, that is structure, unstructured and semi-structured, that is the data is in the form of text multimedia and so on this becomes the variety of data. The fourth one, that means out of three, one more if we take this is called velocity. So, Velocity means, the data which is in Doubt that means the data which contains the uncertainties and this X to inconsistency, in completeness, latency, deception and this has to be curated before what is going to be used in the data. So, this veracity is basically that kind of errors, noise and uncertainties which are present in the data need to be handled.

Refer slide time: (28:30)

And there are many more V's and such as Valley DT and means so, so the time of that particular data, will indicate the validity ,variability and viscosity, volatility ,viability, when you vocabulary, vagueness, all these as, all these will add more V. so, it's not three V's, in big data but plus n movies are there .

Refer slide time: (29:04)

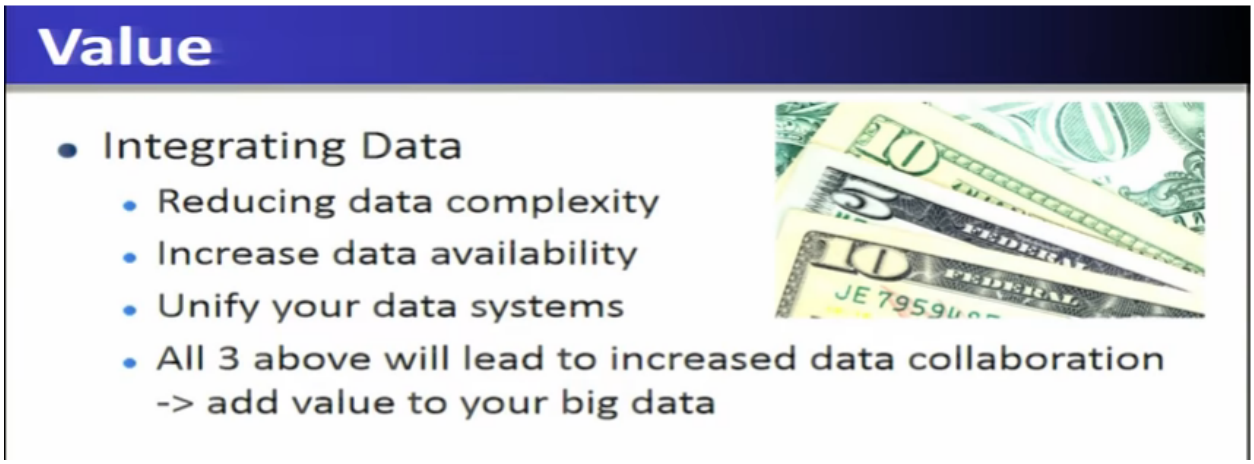


Now, let us summarize the most important V's and which we will be discussing in this part of the course, are as follows. So, that means the big data, first important characteristic of a big data is the volume which

will add the, the complexity in the terms of dimension in the terms of size, the second one is called variety, which will add. So, the first one which is going to add the dimension in the big data is called the volume is going to add this dimension which is called a size in a big data. The second dimension is called the complexity; this is also a dimension in the big data, in the terms of variety. So, what I do will keep on adding the complexity and so, variety is another dimension. So, this complexity is coming out of due to the different variety of data. Third dimension is given in the terms of the speed if it is required and this is called the velocity. So, velocity that is in the terms of speed, will add another dimension and this particular dimension will add more complexity, in a big data computation. Finally the valence, valence means it is the autumn, which is taken out from the chemistry means, the more connected the data is higher violence it is. So, connectivity will go to add one more dimension to the big data. So, why this connectedness is important? Because, if you design the machine learning algorithm and if the data is less connected, than machine learning algorithm will work fine but, if the data is more connected ,then those machine learning algorithm has to be taken into a new way or revisit and a new machine learning algorithm is sometimes required.

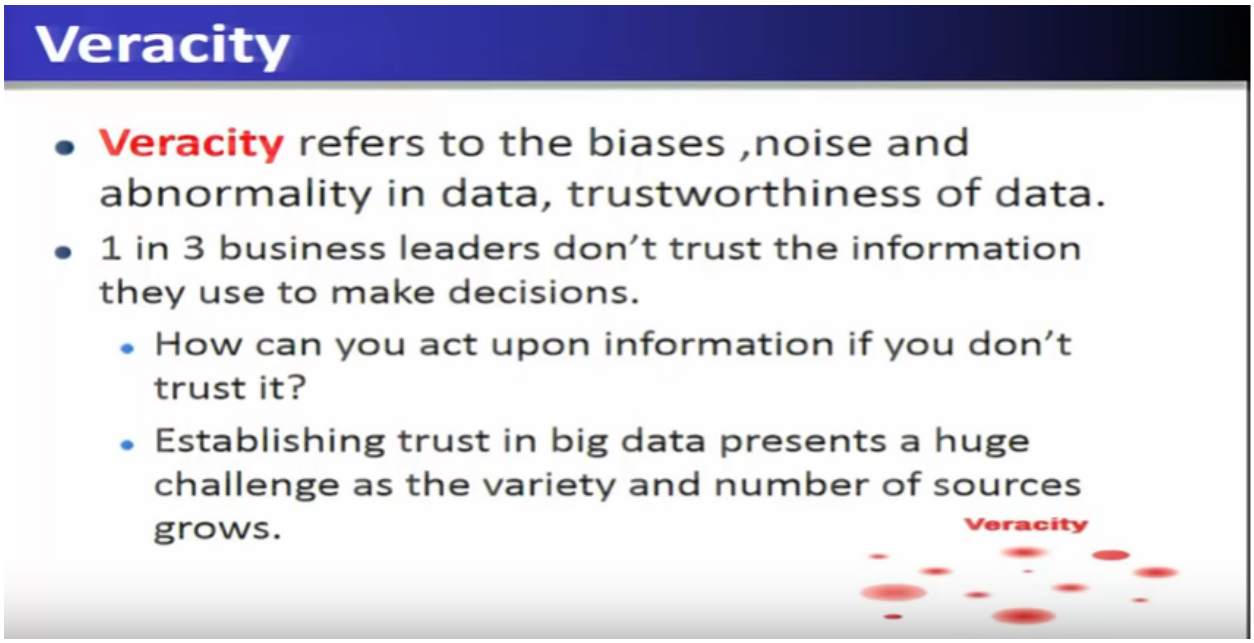
So, it basically depends upon these different characteristics of a Big Data, and how the analysis is to be done that is the techniques, need to be means revised again, that is why these complexities are so important in the processing of this big data. finally another porosity, as I told you that lot of noise is there so, with a lot of noise and incompleteness, inconsistencies, remains into the data, and this particular data if it is analyzed obviously the quality of the decisions will go down .so, this is the dimension, which needs that the data has to be cured, I had a quality data is required so, that the decisions also will be more accurate for, accurate decision making. So, this these are basically characteristics will add different dimension of complexity in computation or Anneli or analyzing the data so, hence the big data analytics has to deal with these complexities and we are going to see all these aspects in this part of the course and finally these at the heart of all these dimensions you see that, this is at the heart, meaning to say that finally using this particular different characteristics and their dimensions finally you have to gain some value for extract some value out of that particular big data and which is going to be useful for an application. So, this value it has to give a value otherwise why? This data big data is becoming so, important that we are going to see. So, value is going to be made finally and this will be used in various applications.

Refer slide time: (33:30)



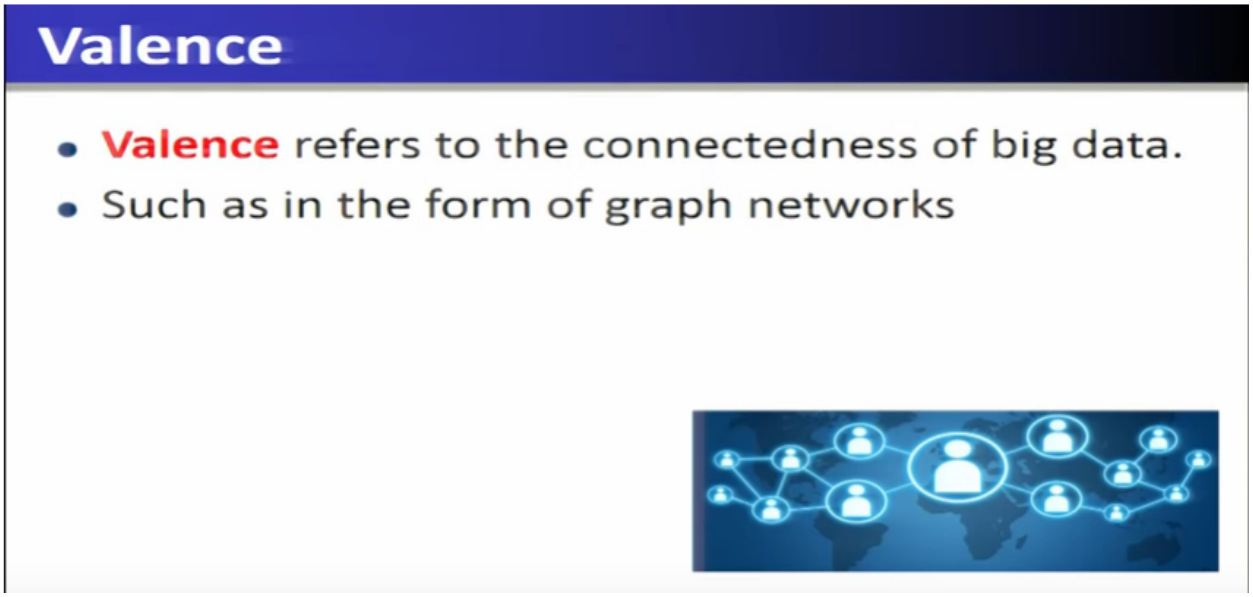
So, value is derived out of integrating these different dimension or characteristics of that particular data. For example, sometimes you can reduce the data complexity, increase the data availability, unify your data streams and all these above will lead to the increased data collaboration and also will add the value to your big data. So, value adding the value, are extracting the value out of the big data for a different application is going to be at the heart or at the center.

Refer slide time: (34:01)



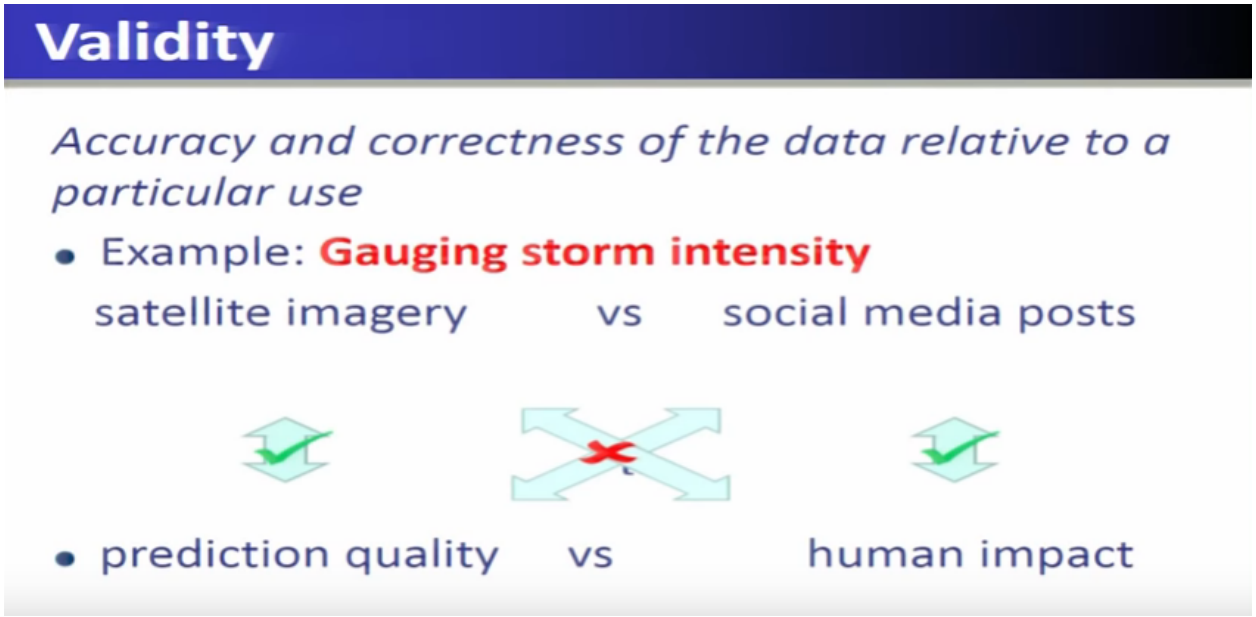
Now, we will we have briefly discussed let us see more detail about the characteristic, which is called the velocity. So, veracity refers to the biases or the noise or the abnormalities, which resides into the data and basically sometimes the doubt on the trustworthiness of the data. So, for example, one in three business leaders don't trust the information that they used to make the decision. and for example, if you let us say age is asked by a particular person and if that particular person is giving a wrong age, and so, basically this goes in the terms of noise or sometimes people don't specify their age and sometimes if the age is going to be important in making decisions. in a particular business and then basically this particular aspect is going to be touched upon as veracity. So, how can you act upon the information if you don't trust on it? For example if somebody gives wrong age information and if you are acting on it then the decisions are not going to be accurate. So, veracity is going to be an important factor and it will affect the decisions and so therefore, the quality of data veracity will ensure that way, so, establishing the trust in a big data presents a huge challenge as the variety and the number of these sources grows.

Refer slide time: (35:38)



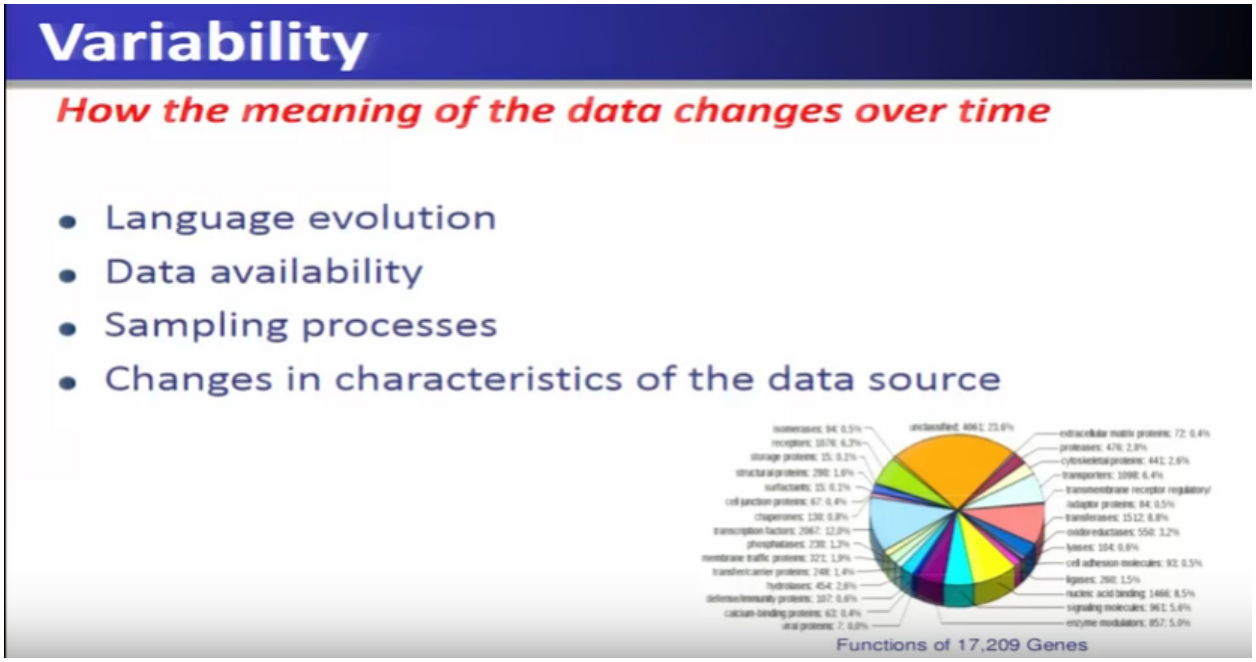
Another, characteristic is called valence or often refers to the connectedness of the big data. such as, in we see that the, the graphs of forms of a graph of the network, that means of the graph is dense sparse. So, there are different analysis in, in the algorithm which are to be applied in these different dynamic situations, varies and the valence, is going to be useful in that aspect.

Refer slide time: (36:13)



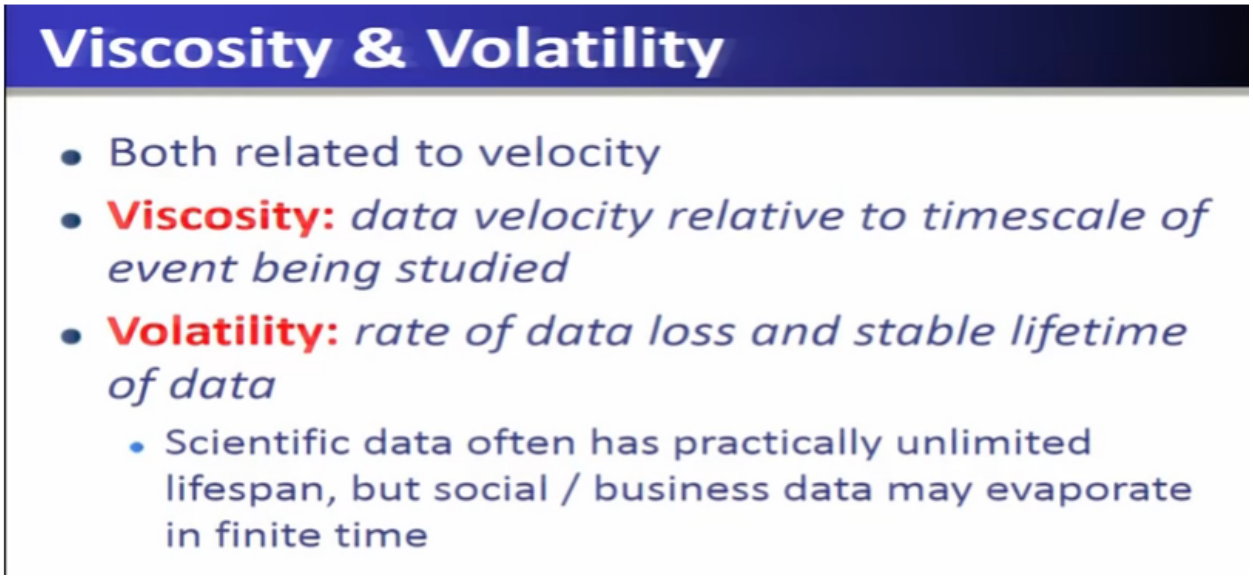
The next we is called validity, that is the accuracy and correctness of the data, relative to a particular use. So, it depends upon a particular use case this validity that is accuracy and correctness of the data is going to be useful. For example, in a satellite imaginary for predicting the quality versus social media post, where the human impact is going to be important part.

Refer slide time: (36:43)



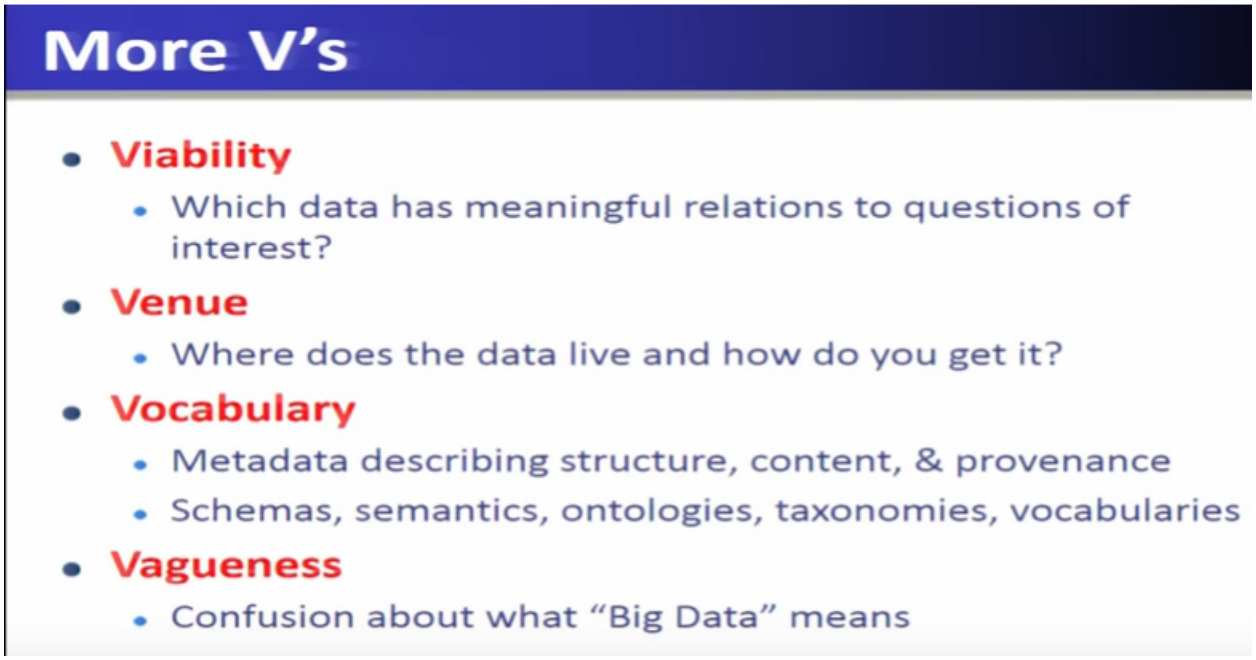
We'll see, more such use cases or examples of these characteristics of a big data, another characteristic is called variability that is, how the meaning of a data changes over time?

Refer slide time: (36:58)



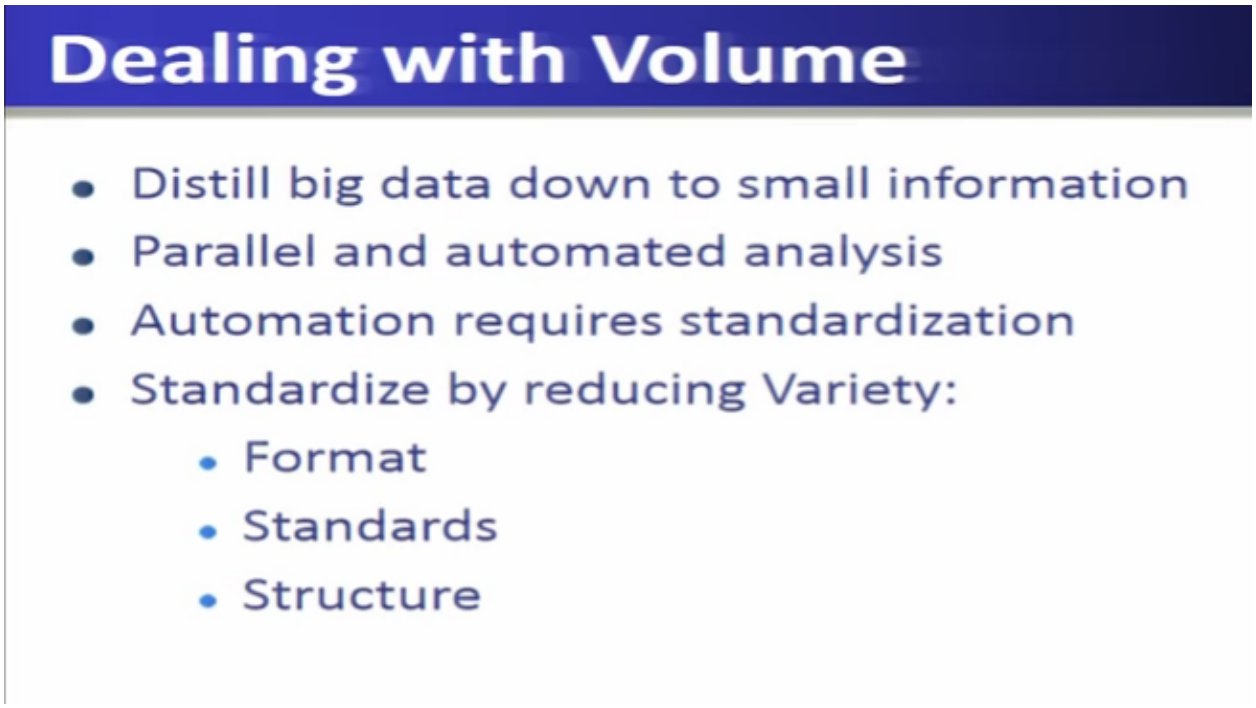
And, furthermore the characteristics are viscosity and volatility both, related to the velocity, viscosity is the data velocity relative to the time scale of event being studied and volatility means the rate of data loss and stable lifetime of the data.

Refer slide time: (37:16)



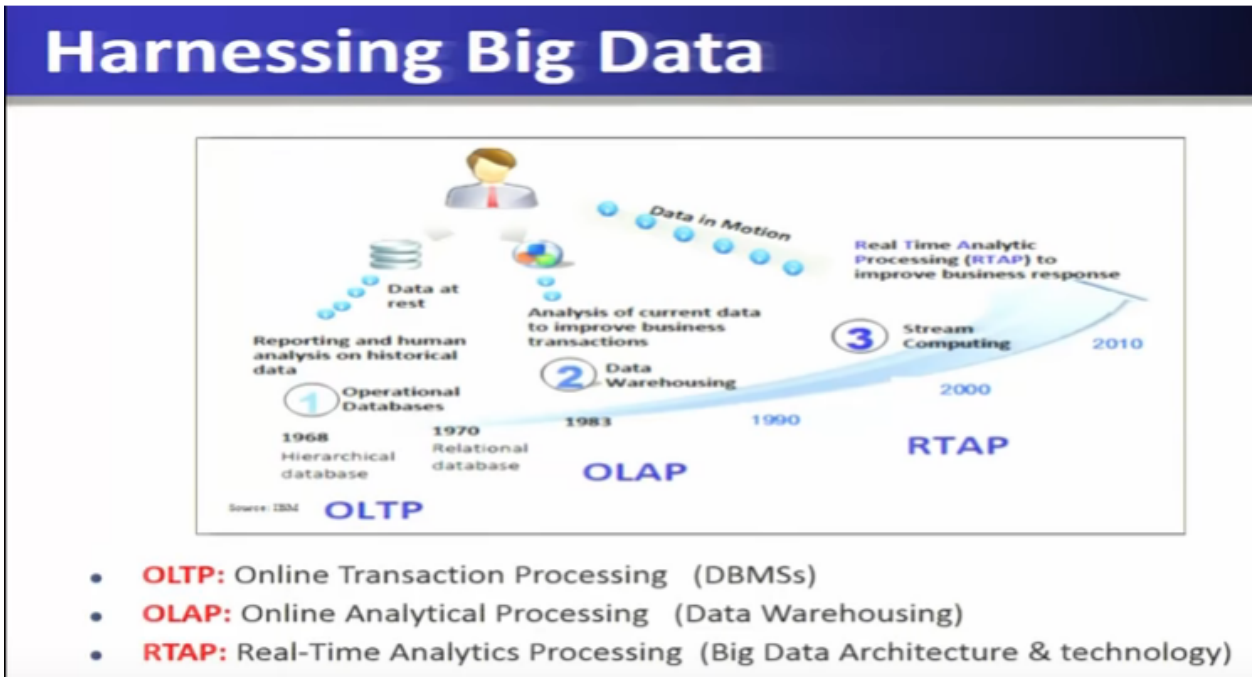
There is more visas and for example, vocabulary means, metadata describing the structure and vagueness is the confusion about what big data, means at that particular application, for that application.

Refer slide time: (37:32)



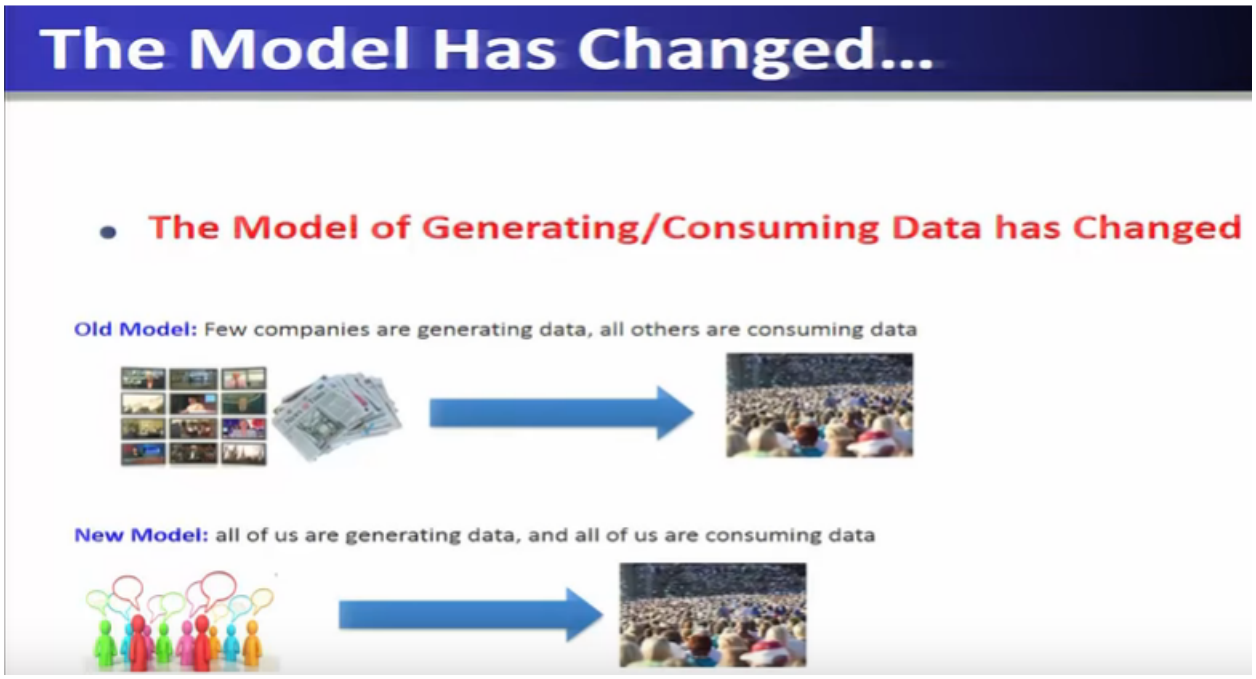
So, now coming, how are we going to address? these characteristics and the complexities around ,these different characteristics, which are there in the big data .so ,if if the volume is big then, we require to develop the method ,which can be computed in parallel the data and also perform the distill this particular big data to gain the summary of that information and how this particular data is to be handled? That is, what he will be the format, standard, instructure? And this all will be taught in the terms of dealing with this volume.

Refer slide time: (38:24)



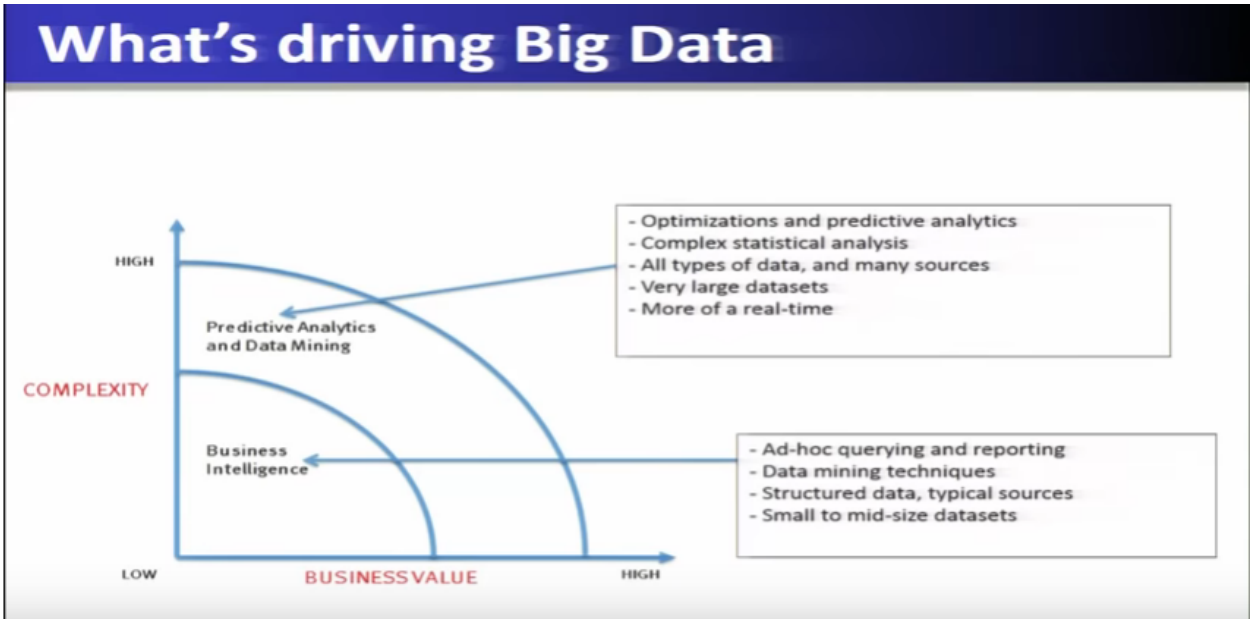
Similarly if we see, about the harnessing of a big data, one way means earlier, the traditional approach was using the operational databases every company was having the databases where the name of a customer and all the details were is stored. so ,that is how the relational database becomes very powerful in and the means the and has developed lot of classical techniques to handle and that is called OLTP, the next stage is again has been passed out, which is called OLAP. and this deals with the data warehouses that canes out of different databases it pulls the relevant information and forms the data warehouse for making the decisions finally nowadays, it is our tap and here the data which is in the form of a stream, that is data is in motion and it has to be called stream computation has to be applied on it to extract the meaningful insight and this RTAP is called, real-time analytic processing to improve the business ,response, and this is the latest trend and in the big data, we will see about the stream computation. So, OLTP means online transaction processing, which is related to the date DBMS and OlAP, stands for online analytical processing, which deals with the data warehousing and RT AP which is called real-time analytics, processing which handles the big data architecture and the technology.

Refer slide time: (40:11)



So, we see that the model of the competition is quite changing, that means the earlier model, if we see was based on the a DBMS, OLTP and OLAP, this was the old model and new model is based on the real-time data, that means all of us are generating the data and all of us are consuming the data is not only the companies which are generating the data and they are consuming it.

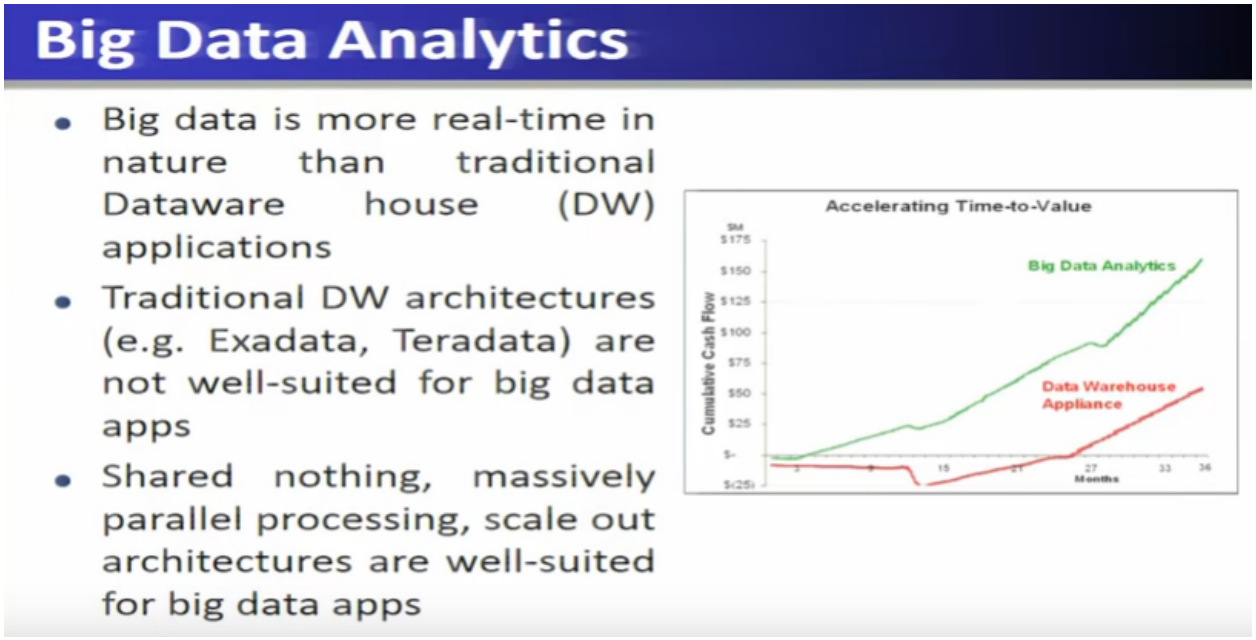
Refer slide time: (40:42)



So, the new Model, is required and to be integrated into the different business, decision-making and the solutions and therefore, if we see this particular picture which will, which is? What is the driving the big

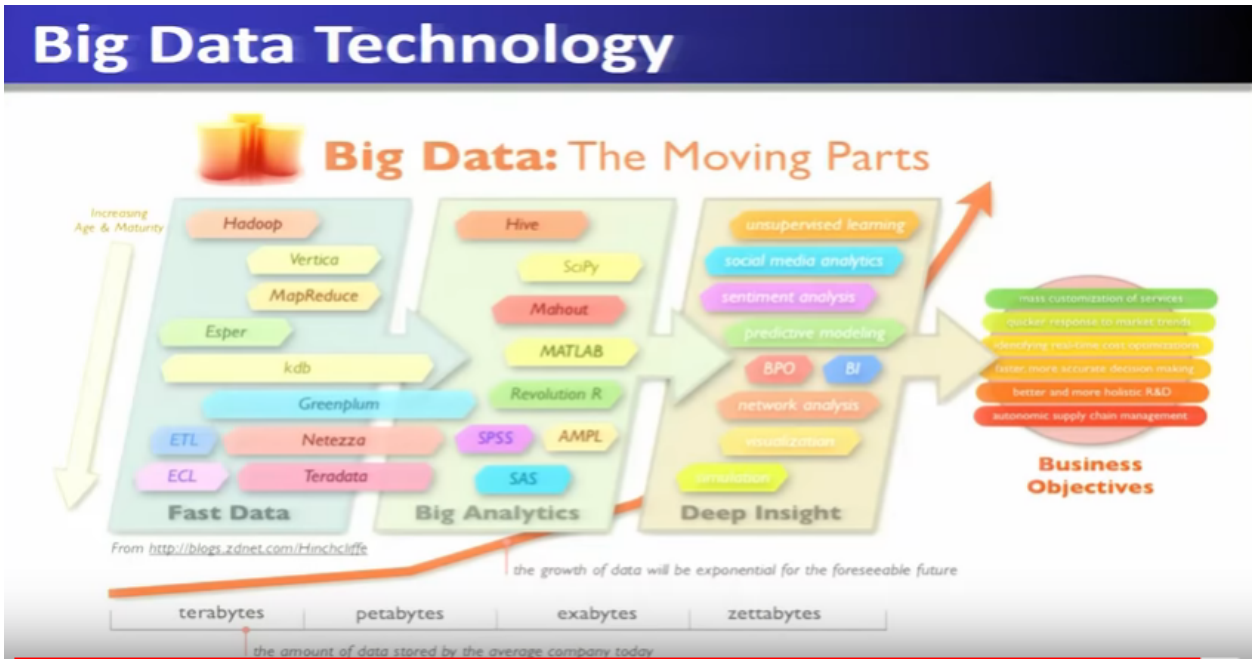
data? further development and its research and its use case is shown here in this particular picture that means earlier it was a business intelligence, we are the value of smotret and the complexity also was moderate but, nowadays it comes predictive analytics and data mining here, the optimizations and predictive analytics are not easy and requires a computation of the big data. And, and also has to be done in the real time. So, all the complexities are now there and the analytics becomes now, called predictive analytics, and earlier analytics ,in the business analytics, business analytic ,intelligence uses the, the prescriptive and descriptive analytics ,but nowadays, predictive analytics is very much of use which needs a real-time or a stream computation processing of the large data sets.

Refer slide time: (41:57)



So, big data analytics is driving, these different businesses and requires the insights out of the big data computations, that we are going to cover up in this part the course.

Refer slide time: (42:17)



So, as far as if you see the big data, which is moving? So, big data is first we see that it is the first data and ETL and all these things, which we have already seen, and then comes the big analytics big data analytics, which are different tools? Which are available? Which is based on the Big Data technology? And nowadays to gain deeper insight, different machine learning and predictive analytics are applied on the big data that we are going to see. So, these kind of I mean now, now those techniques, for analysis, analytics, requires, computing in terms of terabytes, petabytes, exabytes and zettabyte that is the huge size of volume.

Refer slide time: (43:02)

So, conclusion in this lecture, we have defined Big Data and discuss the challenges and various applications of big data, we have also described in more detail about, the characteristics of big, big data and the three most important characteristics of a Big Data, that three V's we have covered in quite detail that is the volume velocity and variety. Furthermore we have also seen other V's, which are evolving

around that big data as the, the time progresses and matures this particular big data area furthermore. So, big data analytics we have also seen a little bit about that and also, about the big data landscape and various terminologies and technologies we have already just touched upon. Thank you.