**Practitioners Course in Descriptive, Predictive and Prescriptive Analytics**

**Prof. Deepu Philip**
**Dr. Amandeep Singh Oberoi**
**Department of Industrial & Management Engineering**
**Indian Institute of Technology, Kanpur**
**National Institute of Technology, Jalandhar**

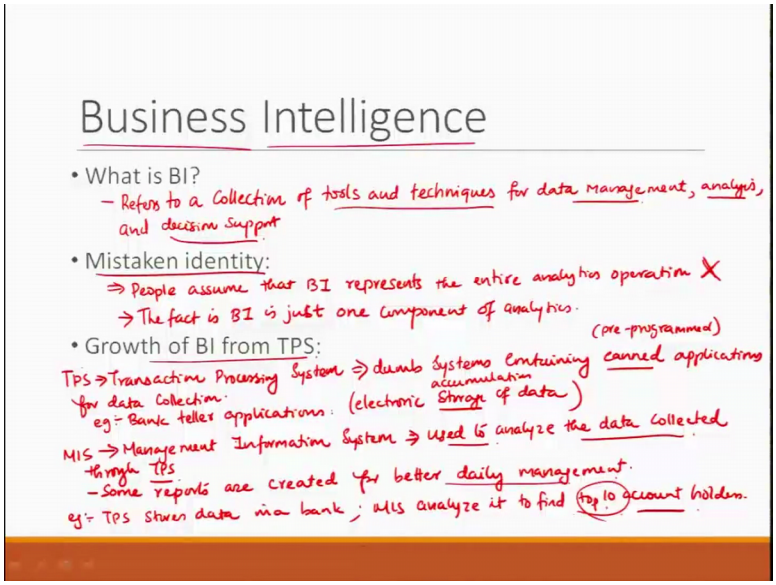
**Lecture – 13**
**Business Intelligence & Analytics**

Good evening students welcome to one more lecture of the Practitioners Approach and Business Analytics on Descriptive, Prescriptive and Predictive Analytics. And today we are getting into one of the ones will take a small deviation to a concept called business intelligence and analytics.

So, the title is Business Intelligence and Analytics Intelligence and analytics today and why we get into this course today is because from here after we are getting into a lot of high end analytic models like customer lifecycle, value, then market basket analysis or risk analytics and those kind of stuff and which requires large data sets and playing with the data that is a pretty much allowing coming from business intelligence system or a data warehouse or something like that.

So, in that regard we require we need to know you know what are the what are some of the major aspects of business intelligence; you know why is it an important component of business analytics and I am Dr Philip from IIT Kanpur.

(Refer Slide Time: 01:17)



So, getting into the topics the business intelligence everybody hears about this topic called Bi and we all talk about Bi Bi Bi. So, what is business intelligence? And let us define business intelligence for this course as it refers to a collection of tools it is another set of tools; collection of tools and techniques; this is tools a collection of tools and techniques for data management, for managing the data and analysis of the data analysis and decision support ok.

So, it help us to manage the data, help us to analysis analyze the data and help us to support the decision support using the data. Many a times people take business intelligence as a mistaken component of the concept of the business intelligence is mistaken by people ok. People assume that; people assume and they think that Bi represents the entire analytic operation; the entire analytics operation, it is not true ok this is wrong not correct.

The fact is Bi is just one component of analytics, it is not analytic it is just a component of analytics and say collection of tools and techniques in the simplest sense is a collection of tools and techniques to for managing the data and then using this to do the analysis so, how did the b business intelligence systems or business intelligence grew from the basic version of TPS.

So, the TPS stands for Transaction Processing Systems processing system which are in a way they are dumb systems containing canned applications; canned means pre programmed , canned applications for data collection ok. A classic example of this is the bank teller operation; bank teller applications. So, when you transfer money from one account to another or when you put deposit a money to an account and other things it just basically stores the data.

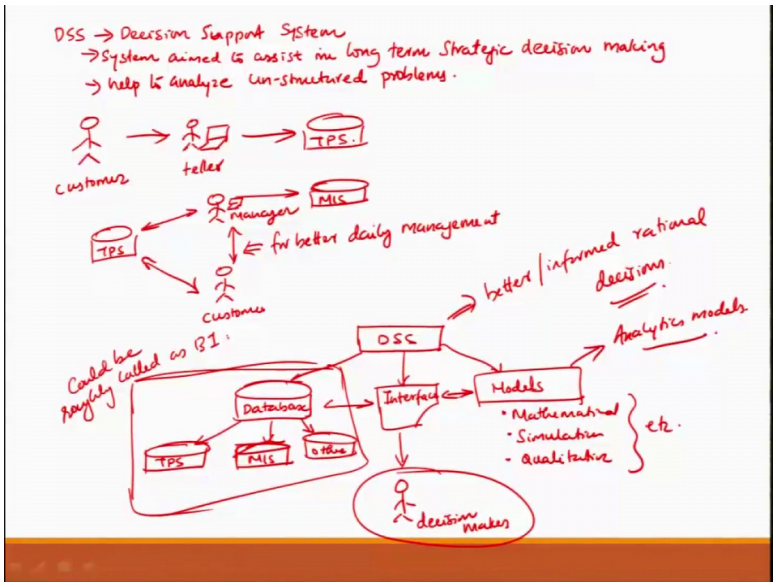
So, this is the electronic storage of data ok; instead of when just storing let us call the word accumulation that is what happens in a transaction processing system. Then after the transaction processing system then comes the system called as MIS ok; which is called as the Management Information System.

In the management information system these are like you know used to analyze the data collected through TPS. So, whatever data we are collecting through TPS; MIS basically analyzes this and why do we do this? Some reports are created; created for better daily management. So, one example of this is the TPS will store all the data like an example is

like TPS stores data in a bank ok; MIS analyzes it to find top 10 account holders who have the largest balance in it.

So, then when these people come into the bank that; the bank manager might like to talk to them personally interact with them stuff like that so, that they deposit more money into it. So, the manager is using this information to basically do better daily management so, that he can get more funds into the into his bank.

(Refer Slide Time: 06:59)




So, the third part of it is what we call as the Decision Support System instead of that let us write it as DSS; colloquially known as DSS which is called as the Decision Support System ok. What is the big thing about it is; this is the system aimed to assist in long term strategic decision making also help to analyze unstructured problems ok. The major difference between MIS and DSS is that the DSS helps to deal with the unstructured problem.

So, let us I will explain to this in a minute, but if you look at the transaction processing system here is a customer and the customer comes to let us say here is a computer. And here is your bank teller and the data gets stored someplace this becomes the TPS ok

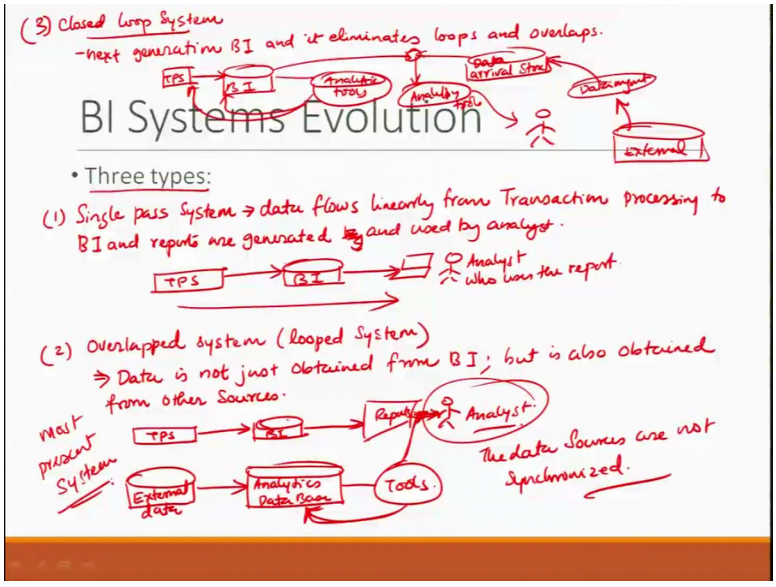
Now, none other thing about this if you think about the same system the TPS data is coming in; this is TPS and here you have your manager ok. And manager interacts with the TPS and in that process he has a MIS data and he is looking into his computer and

here is a customer, he is anyway interacting with the TPS through the teller as we said earlier. And that manager looks at the MIS and then he interacts with the customer for better daily management ok. This is the MIS followed by the manager; on the other hand the DSS is more like DSS is a system where you have something called as a database, it could have underneath this TPS data; it will also have MIS data, it will also have other data sources also. Then there is what we call as a interface for people to interact with the system ok

So, on which the people will interact the decision maker will interact n and it also has different models these models might include mathematical, then could be simulation then qualitative it could be multiple models. And so, in this case what happens is using the interface; the DSS allows the decision maker to interact with the data's, the models and then resulting in better or informed rational decisions; rational decisions ok.

So, DSS to a large extent these models this could also include analytics model ok; this is also a part of the model. So, DSS is very close to what we can call it as a what think about it I say analytic system. So, one aspect of it is this part where you are collecting data and storing of this; this could be could be roughly called as Bi business intelligence, where you are collecting the data and keeping it; it is not just the database it is a little bit more than that the tools and techniques as I said earlier.

(Refer Slide Time: 11:48)



So, the Bi evolution when you think about it there are three types of evolutions with which was with the Bi and they are similar to what I just drew earlier, but there is some difference to it. The first evolution is what we call as the single path system and the single path system is data flows data flows; linearly from transaction processing to Bi; Business Intelligence and reports are generated and reports are generated by generated and used by analysts ok.

So, if you think about it the idea is like this, you have TPS coming in TPS and from there the data linearly goes to the business intelligence system; the Bi system from there various reports are generated is the computer reports are generated. And here is the analyst who uses the report ok. So, it is a single pass; it goes in one direction; now the second pass of the second type of the evolution in this case is overlap the system; overlapped system.

Some people also called this as the loop the system it is also it is an open loop the system actually it is not really a closed loop system because the next evolution is a closed loop system. So, what here it is that data is not just obtained from Bi, but is also obtained from other sources ok. So, an example of this would be you have a system the typical system; the TPS to Bi to the reports ok; let us think about this as the reports and the analyst is looking into this.

The other part of it is also there is somewhere called as a external data and this data from there ok; you have like something like a analytics database ok. And from there are various tools of analytics using this; the reports are it this tools also provides reports which also goes into the analyst and this is the analyst ok.

So, it is not just that is the Bi data, but also external data are used, but here the difference is that there is the biggest disadvantage of the system is ok; you can also think about somewhere during the tools you can probably modify the analytics database also, but the data sources are not synchronized ok. So, you are kind of still analyzing the Bi and as well as analytics independently and the analyst is the one who is taking a look into this ok.

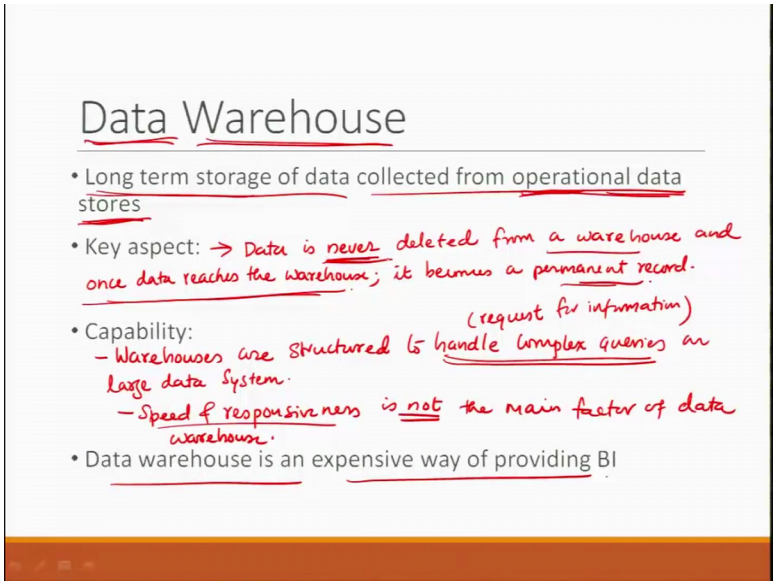
So, it is kind of an overlap the system, but the overlapping is to a large extent is the decision of the analyst. The third one let me write it here; the third system is what we call as the closed loop system ok. And the closed system which is what is called as the next

generation Bi; next generation business intelligence; Bi and it eliminates; it eliminates loops and overlaps.

So, what happens in this system is basically you have your TPS and you have your business intelligence and we have our Bi ok. From Bi there are analytic tools let us think about this as analytic tools there is some synchronization that goes back to TPS also. And then similarly there is external source external source; from the external source there are data management applications ok. And from there they create something called as a you know data arrival storage and then these two data's are combined think about it that way. And then the this is where the analytics tools comes into picture and then that is looked up by the analyst.

So, a system where the data is synchronized and these loops the open loops are eliminated that kind of a system is actually called as a closed loop system next generation; it is also an expensive system. So, currently the most of the systems are here, most present state systems are in the overlap the mode very rarely the closed loop systems are used alright.

(Refer Slide Time: 18:43)




Now, let us talk about things like a data ware think the concept of data warehouse ok. People talk about this quite a lot and there are lot of misconceptions about this. So, what data warehouse is a very simple concept; it is a long term storage of data collected from the operational data stores ok.

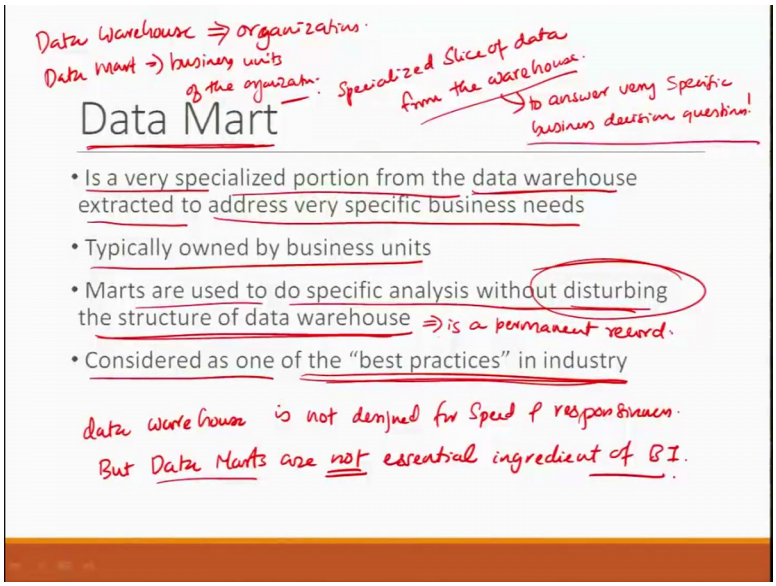
So, when you have operational data source where day to day operational data or data that is related to the day to day functioning is getting stored. From their data is collected and taken for long term storage stored for a long term. The key aspect is the key aspect of data warehouse is that data is never as the key word is never deleted; never deleted from a warehouse ware house. And once it is once it reaches the warehouse once data reaches the warehouse reaches the warehouse then what happens? It becomes a permanent record a permanent record.

So, you can think about it as in a data warehouse where data is collected from different operational side stores and it is there for long term storage which in other words means it never gets deleted from a warehouse; that means, once the data reaches if the data reaches the warehouse then it becomes a permanent record of the sort. So, what are the capability of the data warehouse? What capabilities do it require to achieve this function ok? So, what happens is warehouses data warehouses are structured they are structured to handle complex queries on large data systems ok.

So, they are designed into basically for handling large complex queries or this is you can think about it as request for information that what is called as a complex queries ok. But one thing also you should remember is that speed and responsiveness is not it is not the main factor of data warehouse of data warehouse ok.

So, data house warehouse is not really designed for speed and responsiveness; its main aim is to basically store the data permanently and also handle complex queries and provide the appropriate information ok. But most important one other important product is data warehouse is an expensive ways of providing business intelligence. So, Bi is a set of tools and techniques that allows us for the storage of data for future usage ok. So, if you follow the data warehouse route ok; then it is a very expensive way because data gets never deleted. So, as you make it as a record in the data warehouse, it keeps on the volume keeps on increasing and the storage cost increases ok.

(Refer Slide Time: 22:17)




Then the next concept is called as the data mart this is another aspect and people sometimes confuse between data warehouse and data mart there is a relationship between them, but both are different. What data mart? It is a very specialized portion of the data warehouse or it is a you can think about it as a specialized slice of data; from the warehouse ok.

You are taking a very specialized portion or a slice of the data from the warehouse; you are extracting it to address very specific business needs to answer very specific business decision questions, you want to find out what are the very specific business needs and you want to typically address them ok.

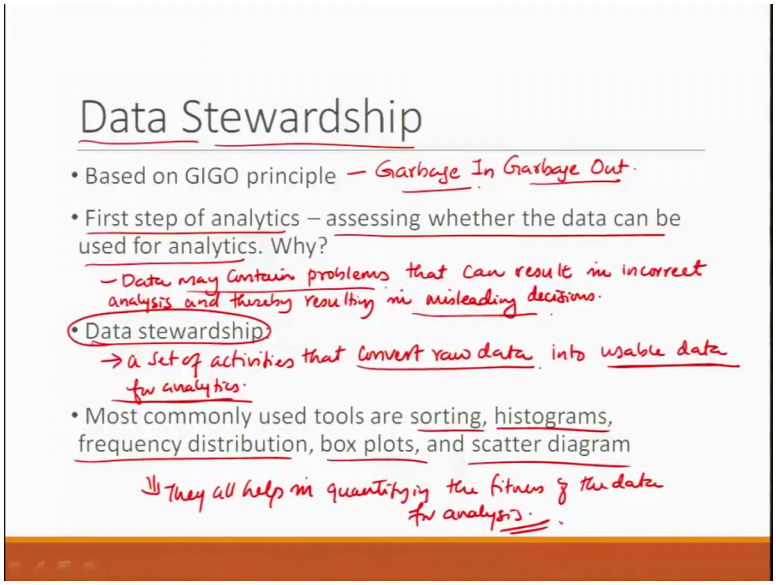
The while the data warehouse ownership is with the organization; data marts ownership is by the business units of the organization. So, you can think about it this way data warehouse the ownership is organization the organization owns the warehouse whereas, data mart is the business units of the organization ok; so, that is the idea here right. So, why do we use data marts? Data marts are specifically they are used to do specific analysis without disturbing the structure of data warehouse. Because remember data warehouse data warehouse is not designed for speed and responsiveness.

Whereas, in the data mart you are anyway doing specific analysis and you do not want to disturb the structure of the data warehouse. So, then you can take your very specialized cut from the data warehouse and then use it and data mart is considered as one of the best

practices in the industry, but data marts are not; they are not essential ingredients or ingredient of Bi to provide business intelligence you need not half data mark ok.

Data mart is an optional thing if you have data mart which means you are taking a very specialized apportion of data from the warehouse; to answer your specific business question. Then that becomes a best practice in the industry because you are not disturbing the structure of the data warehouse because data warehouse is a permanent record. So, I hope you guys understand the concept of data mart clearly because this is one concept that a lot of people get confused with ok.

(Refer Slide Time: 25:43)




Now, we get into the next concept which is called as the data stewardship ok; this is another interesting concept which is developed based on the GIGO principle; GIGO stands for Garbage In Garbage Out; that means, if you put the garbage into the analysis system; the data is fouled or wrong another things then you will get a garbage analysis out of the system ok

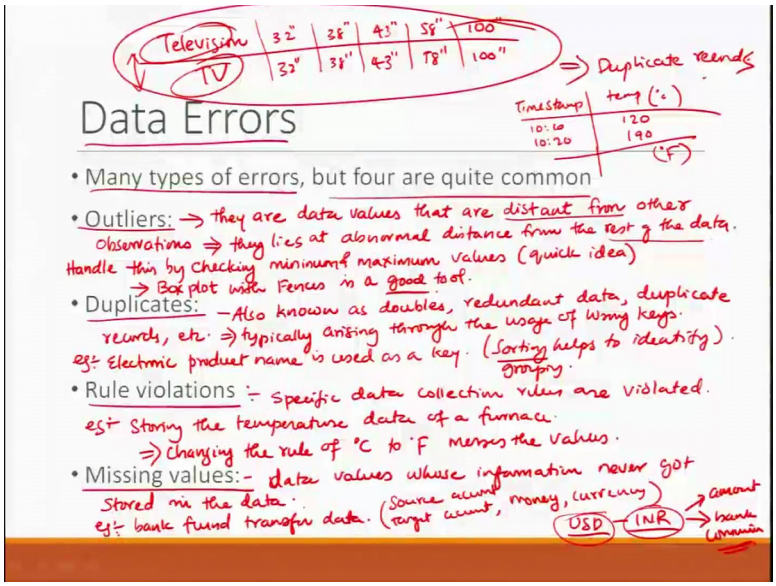
So, the first step of analytics is any before any analytics decide whether the data can be used for analytics; see whether the data is fit for doing the analytics why do you need to do that? Because data may contain; data may contain problems that can result in incorrect analysis that can result in incorrect analysis and thereby resulting in misleading decisions ok. if you want to do proper decisions not to take wrong decisions then you

how to ensure that the data does not have any problems that can result in incorrect analysis ok.

So, data stewardship it is a set of activities that convert raw data into useable data for analytics. So, the aim here is that you first your aim is to convert the raw data which could have problems and convert that by eliminating the problems and other things so, that the data can is usable for analytics. Once you do that process is called as the data stewardship and the most common tools that are used before are sorting; we already seen histograms frequency distribution box plots, scatter diagrams etcetera ok

So, we have seen many almost all of these tools and see how these tools can be used to decide whether the fitness. So, they all help in quantifying the fitness of the data of the data for analysis where we are checking whether we can use this data to do analysis that is the part of data stewardship.

(Refer Slide Time: 28:47)




Moving on we get into the next the important portion of this which is what is called as the errors in the data. This is a very large field and in a course like this our aim is to introduce you briefly into the main errors in the data systems and what are the errors that you need to be worried about if you are conducting analytics.

So, practitioners viewpoint we need to be aware of some of the errors and we need to find out a ways to tackle them ok. There are many type of errors if you think about it

many of them are available, but there are four which are quite common and we will be dealing with these we will be talking about these four as part of this lecture.

And we have tried to find out how to deal with these four type of data errors; the first one we talk about is the outliers ok. Outliers they are defined as they are data values data values that are that are distant long distance distant from other observations ok; they are distant from other observations, which means they lie they lie at of they lies at abnormal distance; abnormal distance from the rest of the data ok

So, here the trick here is that from the rest of the data these data values are too far away they lay a much distant very distant from the other data values ok. So, couple of ways to handle is; handle this by checking minimum and maximum minimum maximum values ok. And this will provide a quick idea what it is right and another tool is a box plot with fences is a good tool. So, if you use box plot with fences we are already showing you that you can use the upper inner fence lower inner fence upper outer fence lower order fence to identify potential outliers and as well as definite outliers in this regard.

So, outliers you can be handled by dealing with the can be quickly identified by checking the min and max values and see whether it is within the range. And can be definitely taken care of by the little bit of the tools that we studied little earlier called as the box plots and that can be used to handle the outliers.

The next form of issue that we are going to talk about is called as the duplicates it is also known as there are many names for this also known as doubles, redundant data redundant data, duplicate records etcetera. There are many ways to talk about this ok, but the this is typically arising through the usage of wrong keys ok

So, for example, I will try demonstrate an example to you ah; assume that we are using the assumed that electronic product name is used as a key ok. So, one electronic product is called as television this is an electronic product and somebody says let us say we are storing the data of the available television models in the LED television models in common.

And let us say we say that we have 32 inch 38 inch then 43 inches 58 inches and 100 inches let us say what it is. And somebody instead of the television uses the concept called TV and stores the same data 58 inch and 100 inches then these two these two can

be called as the duplicate records. The same exact data, but the only reason is because of the television and TV have been used to represent the same concept, but on a database systems these are two different values.

And hence duplicate value gets into picture and this is typically a nightmare to deal with and to a large extent the one way to do it is sorting is a way sorting helps to identify or sorting or other option is grouping sorting and grouping tools allows us to identify duplicates very well that is why we studied quite a lot of grouping tools earlier in this stuff ok

So, then the third one we can talked about it as rule violations where we can call it as specific data collection rules are violated; rules are violated. It can be due to many reasons sometimes instrumental error on other things, but think about a situation where I will give you an example is you are storing the temperature data of a furnace ok.

So, you have if you think about the data you can think about it as that is a time stamp and the temperature ok. Assume that you are storing the temperature initially in degree Celsius and that said 10; 10 AM that was it was 120 degree Celsius at 10: 20 AM; it was 190 degree Celsius like this. And somewhere in between let us assume that the system got changed to Fahrenheit then these temperature values will immediately change.

So, changing the rule of degree Celsius to degree Fahrenheit measures the values ok. So, this is an example of the violation of the data rule ok. Then the last part we are going to talk today is about the missing values. Missing values are where data values whose information never got stored in the data.

So, an example of this is let us say you are storing transaction banking bank fund transfer data ok. And let us say in this process you have to know the source account then target account then money; how much of money is being transacted currency which currency you are transacting and all those kind of things.

So, let us say that you are transacted between USD and INR and on that particular day let us say you did not store the or did not collect the data for that particular day the exchange rate; then that missing value would probably result in the wrongful transact wrongful conversion of the USD dot money; the United States dollars to Indian rupees in that regard ok.

Let us say you did not store that conversion value then that missing value what value that you are used to convert this USD to INR will be missing because you might also result in using a commission out of this if you are doing this transaction. So, we will not be able to decipher the amount and bank commission as part of this ok; so, that kind of examples.

So, this is a vast field as I said earlier, but the aim here is that we will we will take a quick look into it and so, that you are exposed or aware about these are the main errors data errors that you see. And data stewardship is to a large extent identify such type of anomalies and mistakes in the data and correct them. And so, that the data is suitable or as a fit for the analytics process.

With this we conclude our today's lecture on data business intelligence and as well as its importance or why it is a very important component of doing data analytics.

Thank you very much.