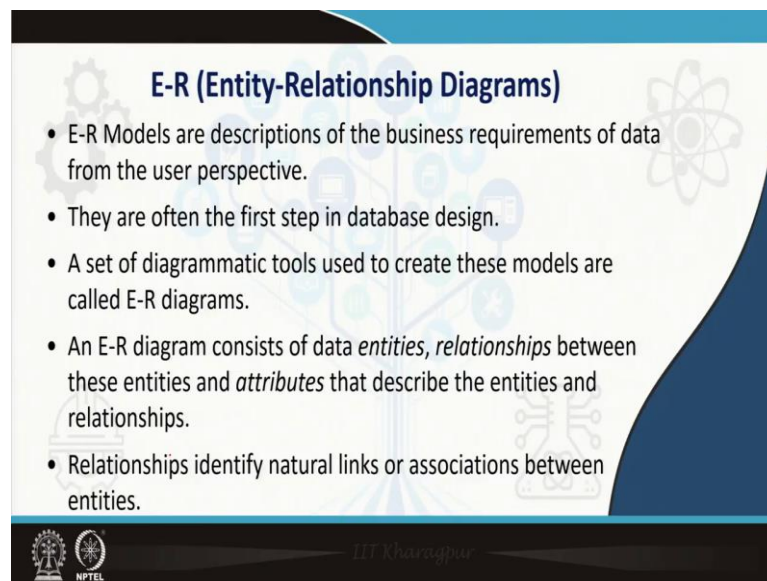


Management Information System
Prof. Saini Das
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Module - 02
Foundations of Business Analytics
Lecture – 07
Data Warehouses & Business Intelligence

Welcome back! In the previous lecture, we had discussed ‘the evolution of file or data organization’ and ‘file organization in information systems’. So, we had discussed the traditional methods of file storage, data storage, then we had moved ahead and we had discussed about database management systems. We had also spoken about relational database management systems in organizations. So, today we will be talking more about ‘data warehousing’ and ‘business intelligence’.

(Refer Slide Time: 00:49)



E-R (Entity-Relationship Diagrams)

- E-R Models are descriptions of the business requirements of data from the user perspective.
- They are often the first step in database design.
- A set of diagrammatic tools used to create these models are called E-R diagrams.
- An E-R diagram consists of data *entities*, *relationships* between these entities and *attributes* that describe the entities and relationships.
- Relationships identify natural links or associations between entities.

IIT Kharagpur

NPTEL

So, let us proceed. In order to you know design a database in an organization. Organizations do require the business requirements of the data. So, because databases are generally used in large organizations or small organizations, which have some specific business requirements. Though a database is a very technical concept the designing of database has certain business requirements behind it only then will it be able to suit the context of the organization.

Therefore, today we are going to discuss about entity relationship diagrams or entity relationship models. E- R models as they are called are descriptions of the business requirements of data from the user perspective. Often they are the first step in database design.

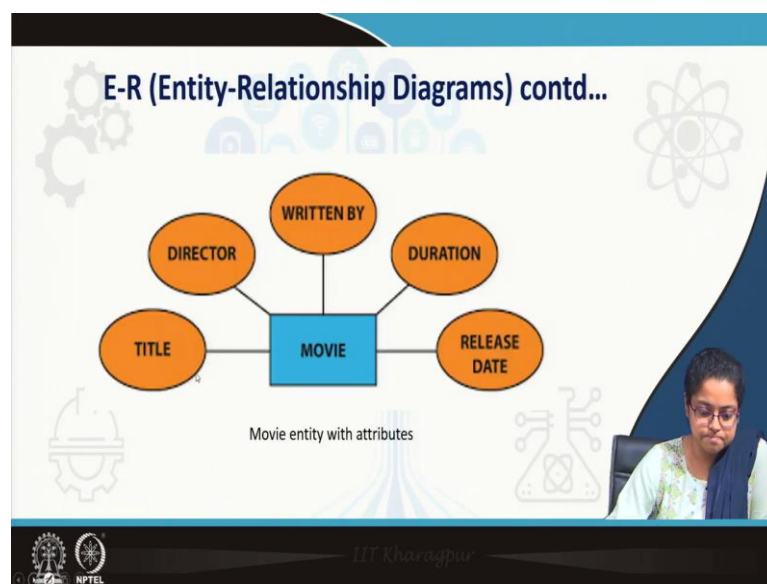
Since database design is a very technical concept, in order to derive more business value out of the database and in order to make the database suited to the context of the organization. It is very important to prepare these entity relationship models from the perspective of the end user. And, they are often the first step in database design.

So, a set of diagrammatic tools that are used to create these models are called entity relationship diagrams. An E-R diagram consists of data entities, relationships between these entities, and attributes that describe the entities and the relationships.

So, in the previous lecture we had spoken about entities. And, we had discussed that entities are could be anything, they could be a person a thing, a place, about which we want to store data. An attributes are those features that describe these entities. So, we had; when we had spoken about the context of a ‘student’ entity, we had discussed attributes such as the course, the grades, the demographic background of the student, etc.

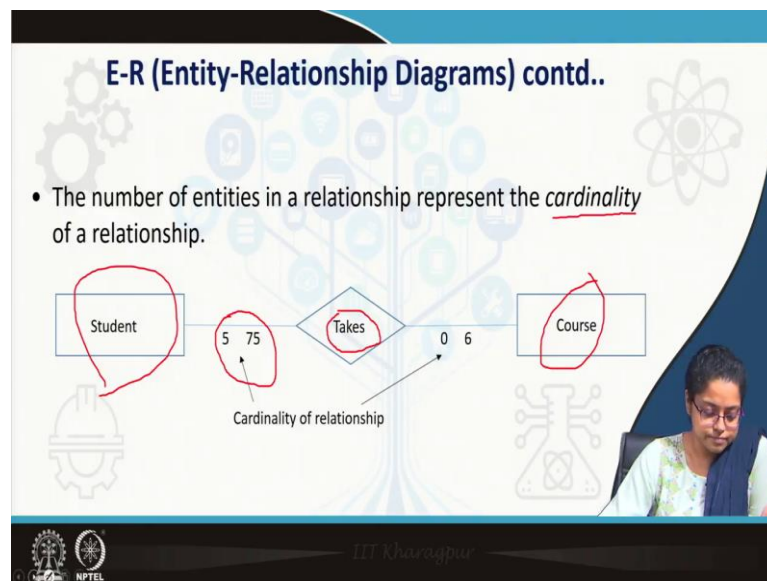
Now, relationships identify natural links or associations between these entities. So, we will see this with the help of an example.

(Refer Slide Time: 03:06)



Here, in the previous lecture we had spoken more about student entity and we had said that this entity could be anything. So, today we will talk about movie as an entity. So, a movie entity with its attributes. So, here we see different attributes such as the title of the movie, the director, the writer, duration release date; we could have N number of other attributes such as you know the box office collection, the music director and so on. All of these are characteristics or attributes of the entity movie.

(Refer Slide Time: 03:45)



Moving ahead see in this particular you know diagram here, we talk about two entities; one is a student. A student entity and the other is the course entity. So, here we have two entities and we are trying to find out the relationship between these two entities. So, here in this particular example a student opts for a course or a student takes a course. Therefore, takes here is the relationship between student and course entity.

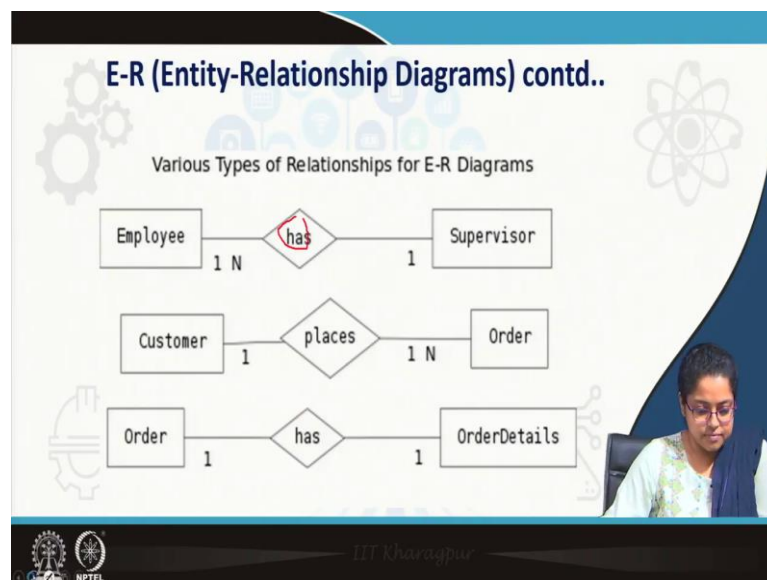
Now, here we see 2 numbers. So, what are these 2 numbers? What do they represent? Right. The number of entities in a relationship represent the cardinality of a relationship. So, this is a very important concept called cardinality, which talks about the number of entities that are party to a relationship. In this particular example we see student takes a course right.

So, here we have 5 and 75. These represent the cardinality of the relationship, which means the minimum number of students and the maximum number of students that can take a particular course.

So, a minimum of 5 to a maximum of 75 students can take a particular course. Why a minimum of 5? That depends upon the university, because many universities say that you know if there is a less than a minimum enrolment of 5, then the course will be dropped. So, here we have assumed that the minimum number of students required to enrol for a course is 5. So, 5 to 75.

Similarly, each student can take 0 to a maximum of 6 number of courses again that depends entirely on the universities rules and regulations. So, university may permit a student to take 0 no courses in a particular semester or a max up to a maximum of 6 courses in a particular semester. So, this is what represents the cardinality of the relationship.

(Refer Slide Time: 05:48)



Moving ahead, so, this particular slide talks about different examples of relationships and cardinalities for E-R diagrams. So, here we see in the first example employee has an supervisor. So, has is the relationship between the employee and the supervisor. So, here a minimum of 1 to a maximum of many number of employees have 1 supervisor.

So, 1 N here means minimum of 1 to maximum of N number. So, many number of employees can have 1 supervisor. Similarly, 1 customer places 1 to many number of orders. So, each customer can place 1 or many number of orders in any particular company.

And, thirdly a third example talks about order. So, 1 order will have only 1 order details. So, it is a 1 to 1 relationship here. So, these are certain examples. Now, when we try to do design a database in an organization prior to that, considering the different entities that would be party to the database, an entity relationship model is prepared, considering all the different entities their attributes and their relationships.

So, this is used, this is designed from the end users perspective and has the business requirements in mind, eventually this would be utilized in designing the database.

(Refer Slide Time: 07:23)

Normalization
Streamlining complex groupings of data to minimize redundant data elements.

Students Engg. Mechanics

Roll No.	Name	Dept.	HoD	Dept. Contact no.
101	Sachin	Electrical	Prof. X	1234567
102	Rahul	Mechanical	Prof. Y	4567899
103	Saurav	Electronics	Prof. Z	6789048
104	Virat	Mechanical	Prof. Y	4567899
105	Dhoni	Electrical	Prof. X	1234567
106	Anil	Mechanical	Prof. Y	4567899

The slide features a table with six rows of student data. Red underlines are drawn under the 'Dept.' and 'Dept. Contact no.' columns for rows 101, 102, 104, 105, and 106, highlighting the repetition of department names and contact numbers. The slide also includes a small inset video of a woman speaking in the bottom right corner and logos for IIT Kharagpur and NPTEL at the bottom.

Let us move ahead. So, we will talk about a very interesting concept called normalization, which is a very important feature in reducing redundancy in databases. Because, earlier in the previous session we had spoken about the presence of redundancy in databases, we had said that you know in databases there could be redundancy data could be there could be an overlap of data, the same data could be present in multiple locations. So, that would create a problem.

Therefore, normalization is the concept of streamlining complex groupings of data to minimize redundant elements. Here, we take a very simple example of a table in any engineering college; in the first year we see that students take up some core courses. So, these courses are taken by students from different departments. Say in this particular example, this is the students table for engineering mechanics, which is a core first year subject in any engineering curriculum.

So, here in this particular table, we have data about say multiple course students, here we see roll number 1 is Sachin from Electrical Engineering Professor X is the H of HOD of the Department Electrical Engineering and the department contact number is this. Similarly, we have 100 and 2 for Rahul from Mechanical Department with the respective HODs name and the department contact number.

Similarly, for Saurav and again we see 104; Virat department is again Mechanical Engineering; same Professor X is the HOD and the department's contact number is again the same.

So, here we see a pattern right. We see that, there is a lot of repetition or what we call redundancy. Here, we see electrical engineering Professor X and the department contact number is getting repeated again here.

Similarly, mechanical engineering details are getting repeated thrice. So, if there is a huge table there would be a lot of repetition, a lot of redundancy and lot of overlap that would result in a lot of space wastage. We do not want that creates unnecessary confusion unnecessary wastage of space and you know storage space is money right.

So, how do we try to tackle this particular problem? Here comes the concept of normalization, which is streamlining complex data groupings to minimize redundant data elements. So, let us see how this can be improved through normalization.

(Refer Slide Time: 10:10)

Normalization (contd..)

Roll No.	Name	Dept. Id
101	Sachin	001
102	Rahul	002
103	Saurav	003
104	Virat	002
105	Dhoni	001
106	Anil	002

Dept. Id	Dept. Id	HoD	Dept. Contact no.
001	Electrical	Prof. X	1234567
002	Mechanical	Prof. Y	4567899
003	Electronics	Prof. Z	6789048

DT Kharagpur

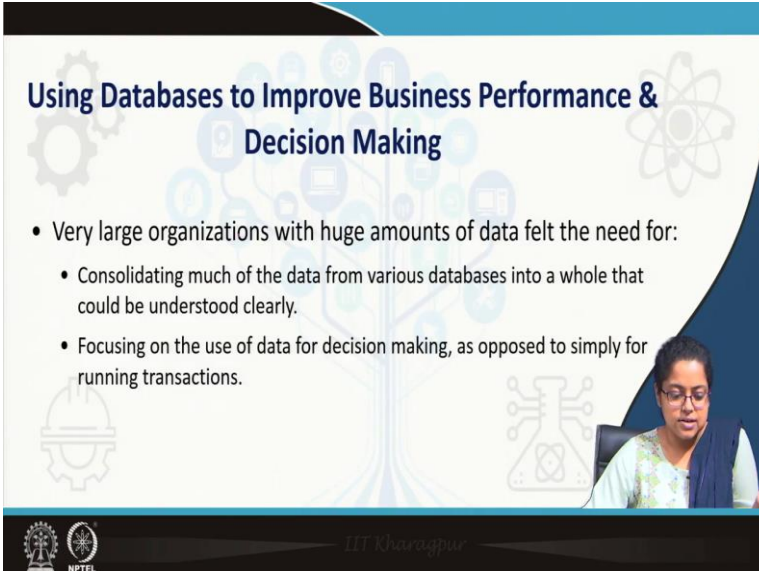
NPTEL

In order to normalize that one particular table that was the previous table that we saw is now split into 2 smaller tables. One table which has the details of the students for the engineering mechanics course and the third column here is this the Department Id. Department Id is also present in the department table, in the department table department Id is the primary key which we had discussed again in the previous lecture.

So, this each of these Ids represent a particular Department with the HODs name and the Department Contact Number. So, here we simply put the Department Id and we get rid of all the redundancies. So, here we maintain two tables; one for student engineering student for engineering mechanics and the other for the department. So, unnecessary redundancy on repetition of data is eliminated in this process.

So, this is a method of normalization in very simple terms I have explained it here, though you know technically there are a lot of complexities involved such as first normal form, second normal form, I did not go into all of those details. Rather for a course like MIS it is important to understand, the concept of normalization and the purpose of normalization in an organization. Rather than all the trivial technical details right.

(Refer Slide Time: 11:38)



Using Databases to Improve Business Performance & Decision Making

- Very large organizations with huge amounts of data felt the need for:
 - Consolidating much of the data from various databases into a whole that could be understood clearly.
 - Focusing on the use of data for decision making, as opposed to simply for running transactions.

The slide features a background with various icons related to technology and business. A video inset in the bottom right corner shows a woman speaking. The NPTEL logo is visible in the bottom left corner.

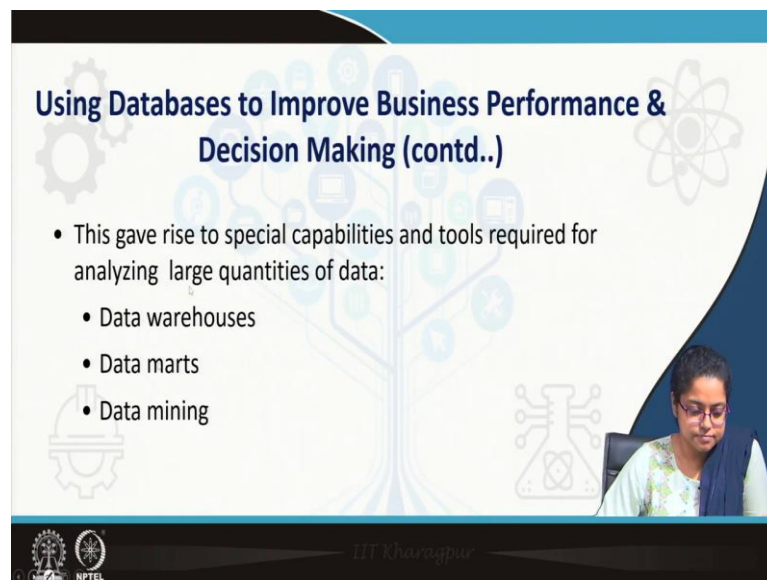
So, moving ahead, we are now going to focus more on using databases to improve business performance and decision making. We see that, today organizations have huge amounts of data.

And, over a period of time as organizations are becoming larger, they have multiple databases we having a lot of data and it is difficult to maintain all of these databases at the same time, it is also difficult to perform analytics on data from multiple databases or to query multiple databases to fit some relevant data. Therefore, very large organizations with huge amounts of data felt the need for primarily two things.

First consolidating much of the data from various databases into a whole that could be understood clearly. Secondly, focus on the use of data for decision making as opposed to simply for running transactions. Prior to this data was stored in databases transactional data was stored in databases and they were used primarily only for recording data, storing data, and that is it nothing beyond that.

But, over a period of time as more and more data got stored and collected, there came to be a need for analyzing the data, for decision making rather than simply running transactions. So, these two needs actually drove a bigger need and we will talk about that.

(Refer Slide Time: 13:14)



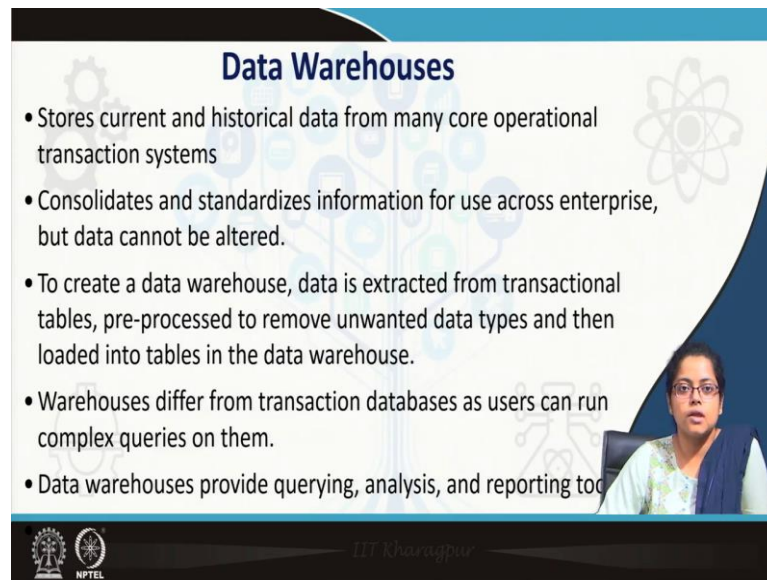
The slide features a light blue background with a central graphic of a tree whose branches are composed of various data-related icons like gears, a laptop, and a network diagram. The title is in a bold, dark blue font. Below the title, a bulleted list is presented. In the bottom right corner, there is a small inset video frame showing a woman with glasses and a blue vest. The bottom of the slide has a dark blue footer with the NPTEL logo on the left and the text 'IIT Kharagpur' in the center.

Using Databases to Improve Business Performance & Decision Making (contd..)

- This gave rise to special capabilities and tools required for analyzing large quantities of data:
 - Data warehouses
 - Data marts
 - Data mining

These two needs give rise to special capabilities and tools required for analyzing large quantities of data. So, what are these tools and capabilities data warehouses data marts and data mining. So, the rest of this particular module is dedicated to data warehouses, data marts and data mining. We will try to understand the concepts and we will try to understand certain other details.

(Refer Slide Time: 13:48)



Data Warehouses

- Stores current and historical data from many core operational transaction systems
- Consolidates and standardizes information for use across enterprise, but data cannot be altered.
- To create a data warehouse, data is extracted from transactional tables, pre-processed to remove unwanted data types and then loaded into tables in the data warehouse.
- Warehouses differ from transaction databases as users can run complex queries on them.
- Data warehouses provide querying, analysis, and reporting tools.

DT Khanna

NPTEL

So, proceeding ahead, what are data warehouses? Data warehouses store current and historical data from many core operational transactional system. So, we have just mentioned that there are large organizations have multiple transactional data bases. So, a data warehouse stores current as well as a lot of historical data from the core operational transactional systems, it consolidates and standardizes information for use across enterprise.

So, from all the multiple data bases across the enterprise data is consolidated and stored in the data warehouse. But, here there is a word of caution data in a data warehouse can never be altered or modified.

To create a data warehouse from multiple databases, data is extracted from transactional tables – ‘pre-processed’ to remove unwanted data types and then loaded into tables in the data warehouse. So, three steps extraction of data from relevant transactional databases pre processing of data and then loading the data into tables, this entire process is called ETL or extraction transformation loading. So, data is then loaded into the data warehouse. Warehouses differ from transactional databases as users can run complex queries on them.

So, data warehouses also provide certain features such as querying, analysis and reporting tools. So, you can query them you can analyze them and you can report create reports or generate reports.

(Refer Slide Time: 15:37)

Data Mart

- Subset of data warehouse with summarized or highly focused portion of firm's data for use by specific population of users.
- Typically focuses on single subject or line of business.
- Used to identify problems and find solutions pertaining to a particular domain.
- Example: Sales data mart.

The slide features a blue header with the title 'Data Mart'. The background is white with faint icons of gears, a tree, and a molecular structure. A red circle highlights the word 'summarized' in the first bullet point. In the bottom right corner, there is a small video inset showing a woman with glasses and a blue vest. At the bottom left, there are logos for IIT Kharagpur and NPTEL.

So, these are the purposes of data warehouse data mart is another separate concept which is again very important data mart is a subset of a data warehouse. With summarized or highly focused portion of firm's data for use by specific population of users. What this means is data warehouse could have data from multiple databases and multiple domains. But data mart is a very focused specific subset of data warehouse which has highly summarized and focused portion of firms data for a specific use by a specific population of users.

It typically used focuses on single subject or a single line of business. It is used to identify problems and find solutions pertaining to a particular domain. So, I think you would agree with me that, if data is there in multiple you know from multiple domains is there in a data warehouse, it might be very difficult to scrutinize the data and to find out identify problems, to resolve those problems and find solutions.

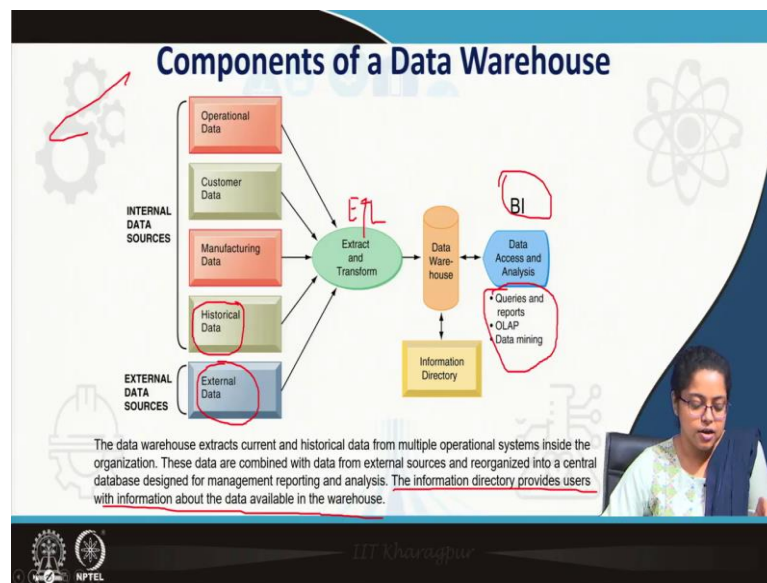
Therefore, data mart is a specific subset which contains data pertaining to only one particular domain. Therefore, it is very it is much easier to delve deeper and to analyze the data in the data mart to find problems pertaining to a particular domain.

So, and a very you know there are a lot of examples. So, here we are talking about a sales data mart. So, though our organization has multiple you know data, repositories in a or data belonging to multiple domains in a data warehouse, a sales data mart pertains

only to sales. It does not have data pertaining to distribution; it will not have data pertaining to warehousing, or manufacturing, but only to sales.

So, if you want to find any problems with sales, you can look deeper into the sales data mart and try to analyze it to find problems there. So, this is the concept of a data mart and this is how it differs from a data warehouse? Moving ahead a components of a data warehouse.

(Refer Slide Time: 17:47)



So, earlier we mentioned that the data warehouse has data from multiple databases. So, a core operational databases within the organization such as operational database, customer database here, and manufacturing database that we see, these are the current data. And, these are all internal data sources pertaining to inside of an organization. Along with these data sources core transactional data sources, historical data is also maintained.

So, in historical data is also maintained in a particular database and that also forms a part of internal data source. Along with this there is a lot of data that resides in external data sources. By external data sources we could mean maybe share market data, it could mean industry data, it could mean a competitor data. So, all of these this data is now together extracted from the multiple databases and transformed and then loaded into the data warehouse.

So, again we will talk about E-T and L extraction transformation and loading. So, it is loaded into the data warehouse. Now, data warehouse has an information directory, which provides users with information about the data available in the warehouse. So, information as the term directory suggests it gives you a information about data that is there in the data warehouse.

Now, data from all of these systems selected data, resides in the data warehouse and if you have to derive BI. Here, we are speaking about BI. So, if you have to derive BI, which is business intelligence, you have to perform certain operations or from on the data from the data warehouse.

So, data is accessed and analyzed using certain tools and features here to derive business intelligence from the data in the data warehouse. So, these tools and features are primarily querying and reporting tools OLAP which stands for online analytical processing, we will discuss this soon and thirdly data mining.

So, these three tools and techniques help an organization; derive business intelligence out of data in the data warehouse. I hope this concept is clear.

(Refer Slide Time: 20:13)

Business Intelligence

- Tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions
 - E.g. Harrah's Entertainment analyzes customers data to develop gambling profiles and identify most profitable customers
- Principle tools to derive BI from data in a warehouse include:
 - Software for database querying and reporting
 - Online analytical processing (OLAP)
 - Data mining

The slide features a blue header with the title 'Business Intelligence'. The background is light blue with faint icons of gears, a network, and a person. A small video inset in the bottom right corner shows a woman with glasses speaking. The bottom of the slide has a dark blue footer with logos for IIT Madras and NPTEL.

So, moving ahead; what is business intelligence? Tools for consolidating, analyzing and providing access to vast amounts of data to help users make better business decisions is 'what is business intelligence'. So, here we have taken an example, Harrah's

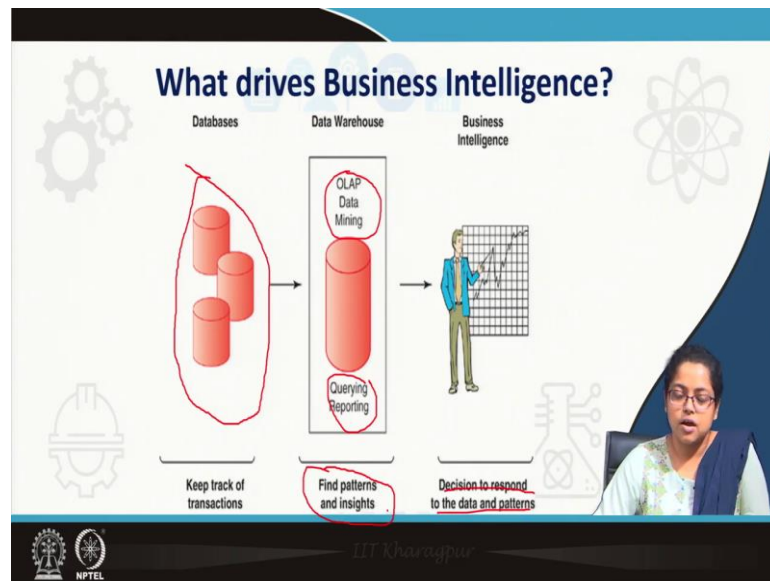
Entertainment, there is a chain of casinos which are there in Las Vegas in the United States.

So, casinos as you know people engage in gambling there. So, you would be you know surprised to know, that in the world of gambling also, data warehouses and bi play a very important role. So, Harrah's Entertainment, analyzes customer's data to develop gambling profiles and identify most profitable customers.

So, once they identify the most profitable customers through bi they would be able to drive their strategies accordingly. So, this is how Harrah's entertainment in the far world of gambling uses business intelligence (BI), you know and data warehousing to serve their customers better.

A principle tool to derive BI from data in a warehouse, we have described this earlier also; so, software for database querying and reporting – OLAP and data mining; ok.

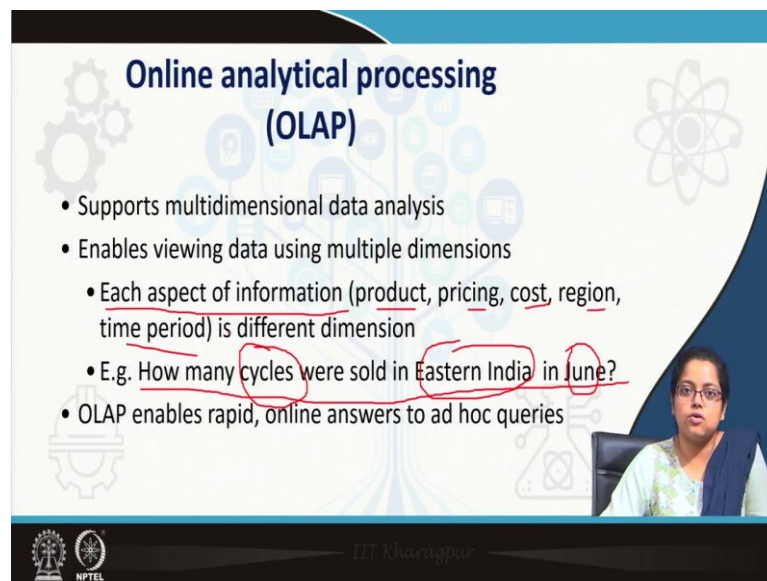
(Refer Slide Time: 21:36)



So, let us proceed. What drives business intelligence? Again this is highlighting whatever we had discussed few minutes back. So, here databases there are multiple databases in organizations that keep track of transactions. We had earlier discussed that we do have a current transactional databases, we have historical databases, we have internal as well as external databases or data sources. So, all of the data is extracted transferred and then loaded transformed and then loaded into the data warehouse.

Now, data warehouse has tools for querying, reporting, performing OLAP and data mining in order to find very important patterns and insights. So, once you find patterns and insights out of the data by all of these tools that we have discussed, you derive business intelligence that is decision to respond to the data and patterns. So, business intelligence helps you make better business decisions out of the data that resides in the data warehouse.

(Refer Slide Time: 22:53)



Online analytical processing (OLAP)

- Supports multidimensional data analysis
- Enables viewing data using multiple dimensions
 - Each aspect of information (product, pricing, cost, region, time period) is different dimension
 - E.g. How many cycles were sold in Eastern India in June?
- OLAP enables rapid, online answers to ad hoc queries

The slide features a blue header and footer. The footer contains the IIT Madras logo and the NPTEL logo. A video inset in the bottom right corner shows a woman with glasses and a blue vest speaking.

All right. So, moving ahead, what is online analytical processing or OLAP in short? So, we have discussed about OLAP before I want to highlight the concept of OLAP now in this particular slide. OLAP supports multi dimensional data analysis. So, earlier we said data resides in a data warehouse, and data warehouse has you know you can perform analysis on the data in the data warehouse using certain tools and techniques and one of them is OLAP. So, OLAP or online analytical processing supports multi dimensional data analysis.

So, what do we mean by dimensions? What are dimensions? OLAP enables viewing data using multiple dimensions. Now, a dimension is each aspect of information. So, for example, you may have a lot of information in an organization, but dimension represents each aspect of that information.

So, aspect could be the product, pricing, cost, region, time, period; all of these are examples of different dimensions. So, the data could be split based on any of these

dimensions or based on multiple dimensions. So, two or more dimensions data could be split based on that. So, here there is a concept called slicing and dicing the data.

So, OLAP basically helps an organization slice and dice the data based on the organizations requirements. Now, when you try to slice and dice the data, if you have a huge data, you know having 3 2 3 or more number of dimensions, you can split it up into smaller you know smaller data of smaller chunks based on the based on the number of dimensions. So, basically if there is a huge data set, you slice and dice it into smaller chunks based on the dimensions. So, each of these are examples of dimensions.

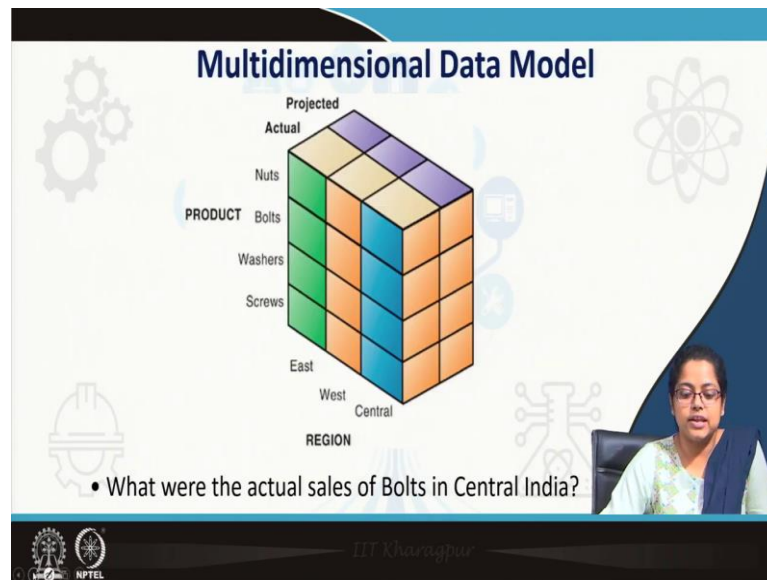
Now, talking further about dimensions here is an example. Now, how many cycles were sold in Eastern India in June?. So, if you have data pertaining to say you know a particular chunk of data, you can slice it based on multiple dimensions. Here, we have three dimensions.

So, I hope you can understand, what are the three dimensions? So, the three dimensions clearly are product that is cycle. So, what is this, what is the product here that is the first dimension that is cycle?

The second dimension is the region, that is Eastern India, that we are talking about there could be other regions, but here we are talking about Eastern India. And, the third dimension is time period right. So, we are now slicing and dicing the huge chunk of data based on 3 dimensions; the first one is product, the second one is region, and the third dimension is the time period. OLAP enables rapid, online answers to ad hoc queries.

So, if there are ad hoc queries OLAP will help you give rapid answers to those ad hoc queries.

(Refer Slide Time: 26:11)



Now, we will see this with an example, multi dimensional data model. So, here we see a data model wherein there is data again pertaining to three dimensions as we had discussed. So, here the three dimensions are product which has four categories, nuts, bolts, washers and screws, region again east zone, west zone, and central zone and this the third dimension is sales.

So, the projected and the actual sales. Now, if somebody were to ask you, what were the actual sales of bolts in Central India?. So, what you would do is basically slice and dice this huge chunk of data, based on the three dimensions that is sales, bolts and sales product and the region.

And, what you would end up getting is so, central India bolts and this particular chunk of data you would get, this particular chunk or cube of data. So, this is how OLAP helps in slicing and dicing the data based on multiple dimensions or is used for multi dimensional data analysis.

(Refer Slide Time: 27:24)

The slide is titled "References" and features a background with a stylized tree of icons representing various technologies and business concepts. The references listed are:

- K. Laudon and J. Laudon (2016). Management Information Systems Publisher: Pearson. Edition 14e.
- R. De. (2018). MIS Managing Information Systems in Business, Government and Society. Publisher: Wiley. Second Edition.

In the bottom right corner, a woman with glasses and a blue shawl is visible, likely the presenter. The bottom of the slide includes the NPTEL logo and the text "IIT Kharagpur".

So, I think we will stop here in this particular lecture. In the next lecture, we will take up you know, details about data mining, and we will see how data mining is used in organizations. In this particular session, we focused more upon data warehouses and how data warehouses can be used to derive business intelligence.

Thank you, see you around!