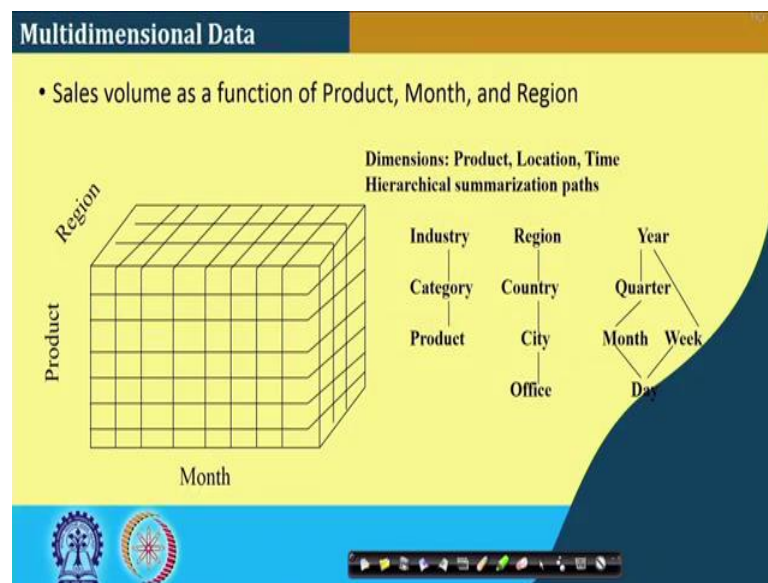


Spatial Informatics
Prof. Soumya K. Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 27
Spatial Analysis – II

Hello, so we will continue our discussion on Spatial Analysis, we are discussing on basics of data warehousing and data mining to basically for those who are not pretty accustomed with these things. So, we will continue that before we go to the things. One thing to keep in mind to so those this basic philosophy remains same when you go for any type of thing like spatial data mining aspects, so the basic thing will remain same. So, the same philosophy and same so to say same type of architecture will be using.

(Refer Slide Time: 00:48)



So, last lecture we were discussing about multi dimensional tables, like if you remember so we talked about a particular say sales volume one to calculate for product, month and region. And there are three type of hierarchy like one for the product is individual product wise category wise and then particular industry wise then there is a location is a office, city, country, region and there is a temporal scale.

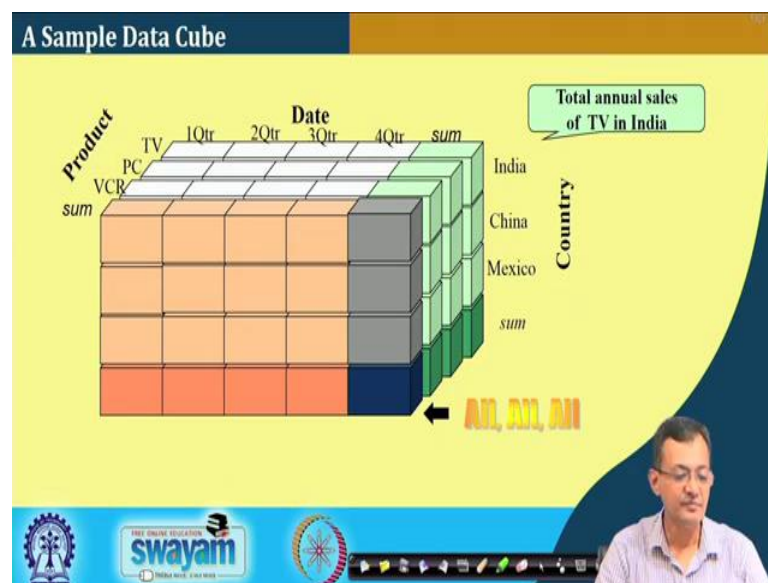
So, either it is a day week; week to year like 7 days a week and say we say 50 weeks a year, then I have a day month then quarter 3 months a quarter etcetera. If I see that the 1 quarter is also 4 weeks, then I could have gone through these line also. It all depends that

how we interpret the how my concept hierarchies there right. So, in other sense I require a daily sales, I require a monthly sales I require say quarterly sales and yearly sales.

Whereas, other channel maybe daily, weekly and then yearly sales right, so this can be things so based on your thing. So, this is the hierarchy how this data will be consolidated or hierarchy how the data will be expanded if you come down right, as I drill up or drill down to the things right. So, these are the operations we want to know.

So, this from the operational data sites when I formed this cube I do a pre processing along with cleaning etcetera and then fill up these things right. So, as and when I am getting the data this cuboids is being filled up or this particular cube is being data cube has been framed where we have this things right. The overall relational things that said the things is by these the models like what we have discussed, like your snowflake model star models and so on and so forth.

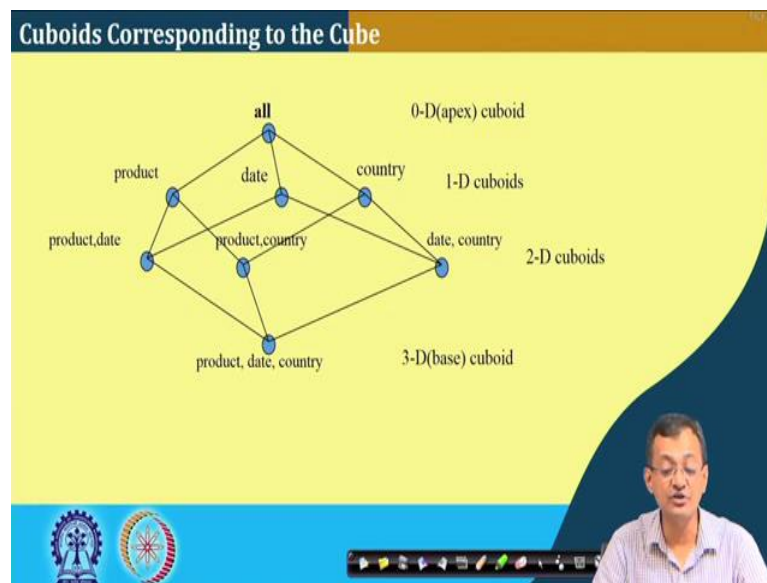
(Refer Slide Time: 02:34)



So, sample data cube like I have this is all like one is product one is date a in another is maybe the particular temporal scale or one is country date and product. And this cube is give the means all products all countries all over all time what is the thing and different I can have different slices that. So, this is the if we fourth quarter third quarter second quarter third quarter, if I want different type of products different country wise things I can get those things and these are the summation.

So, see this is a framework which gives you a some sort of a quote unquote pre cooked material, by which you can easily go to the decision support system. Below that of course, your traditional data sets are there below these the traditional data sets are there or traditional OLTP type of operations are there which are pre processed and filled into our materialized in the data cube right this is a conceptual thing. So, I can have the concept like this.

(Refer Slide Time: 03:48)



And corresponding if we cuboids corresponding things as you all is the 0 dimensional, it is a consolidated 1 dimensional is product date country, 2 dimensional or 2 dimensional cuboids is product date product country date and is 3 dimensional or means for a particular product particular day particular country things are there. So, this is the way I can have the decision instance of different type of cuboids 1 D 2D 3D 0D 1D 2D 3D right type of cuboids for this particular case we can have.

(Refer Slide Time: 04:24)

Typical OLAP Operations

- **Roll up (drill-up):** summarize data
 - by climbing up hierarchy or by dimension reduction
- **Drill down (roll down):** reverse of roll-up
 - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- **Slice and dice:**
 - project and select
- **Pivot (rotate):**
 - reorient the cube, visualization, 3D to series of 2D planes.
- **Other operations**
 - **drill across:** involving (across) more than one fact table
 - **drill through:** through the bottom level of the cube to its back-end relational tables (using SQL)

The slide features a yellow background with a blue header and footer. In the bottom right corner, there is a small video inset showing a man speaking. The footer also contains logos of institutions and a navigation bar.

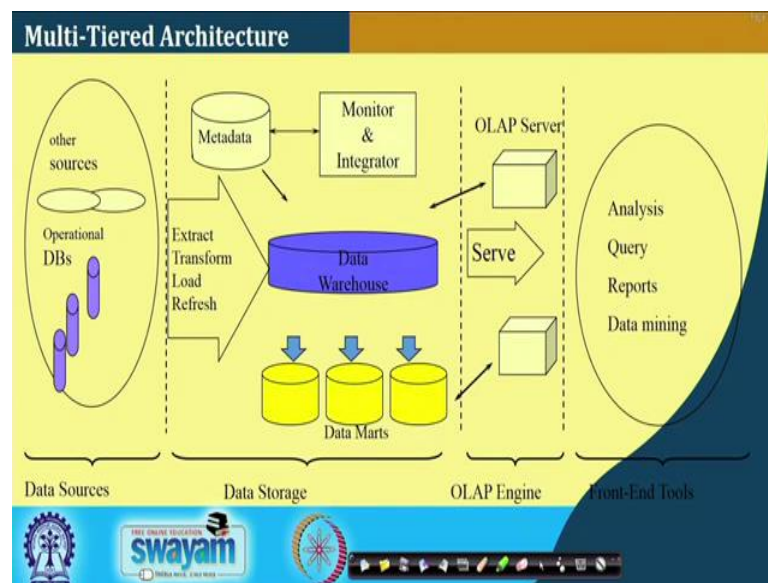
Now, what are the different type of operations or online analytical processing or OLAP operations we are having. So, one is roll up or drill up what you say drill down or roll down slice and dice pivot like over this I change that particular thing pivot another like drill across drill through and type of things. So, these are typical OLAP operations or rather if we have a data warehouse infrastructure like's say different type of databases like oracle and other things having data warehouses.

So, they have this type of OLAP engines or OLAP what we say interfaces or applications right. So, one is roll up that is summarize the data, so I have a daily sale, then weekly, then yearly or monthly quarterly etcetera. So, summarization of the data goes on right, by climbing up the hierarchy in the dimension reduction dimension is reduced right. So, it is all through here what we say all are individual, then I go for 2 dimensional it will aggregated and finally it is aggregated.

On other side drill down is the reverse of roll up right from the higher level of summary to the lower level of summary or detailed data or introducing new dimensions right. Slice and dice project and select right pivot rotate reorient the cube visualization 3D etcetera. So, reorient and see that it which dimension it what looks like that, there are other operations like drill across moving across more than one fact tables. So, these are if you see all these warehouses are built on these fact tables you have a fact tables based on that warehouse is built.

Now, I can have different other fact tables. So, I can if this is there if the product sale is there, then I want to see that which are the supplier which has supplied or some other type of things right aspects of the things right. So, or drill across drill through that through the bottom level of the cube to it is back end of the relational database, from the tube to the relations as I was telling that I have that cube below that some middle tier of pre processing and other things and at the down end we have the relational database. So, I can have a some sort of a drill through type of aspects.

(Refer Slide Time: 06:50)



Then multi so if we overall try to see it then it is a Multi Tiered Architecture. So, I have data sources or what we say traditional databases OLTP etcetera, that what from where we do extract transform load refresh right. So, that creates my data warehouse right. In the data warehouse in order to handle a data warehouse what we require a data meta data information's, because data warehouse will be huge volume of data.

So, I require a meta data about the data about the data right and for these I require monitor and integrator to look at that how the data warehouse can be there alright. And there are what you say a things called data marts which are small chunk of data warehouse. So, which are created based on that the thing, something analogical we can say that like we create views in the things.

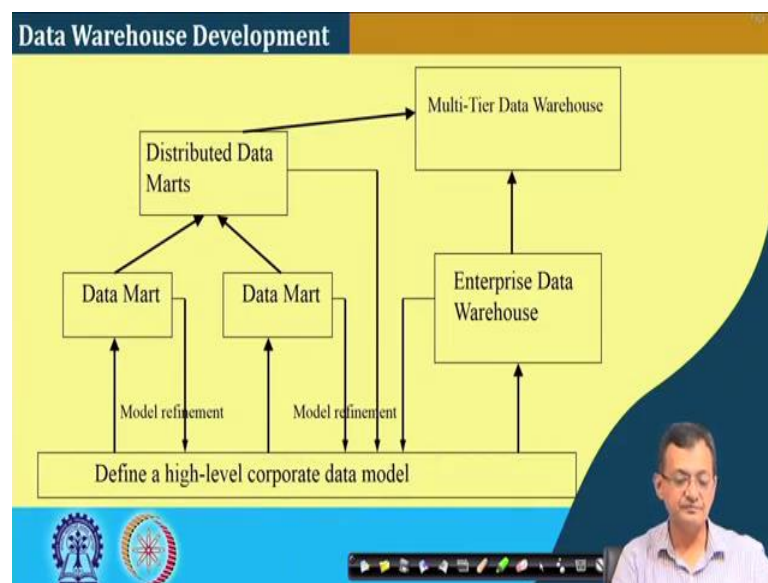
So, that the processes or in sometimes the data marts are in a particular suburbia type of things right domain. From there we have this OLAP server to solve different type of

operations like analysis query reporting and data mining, so there are which are there. So, this if you look at this is the way the data is collected or being acquired. This is a thing which the other end we add as a user at different level of management or different decision support system, we require analysis query reports and things.

This in this in between stuff which support this is this warehouse right. So, it is unlike our traditional query system or OLTP, when the query is fired it goes back to the database and gets here the data warehouses are pre metalized and ready to be served type of things right. Because it deals with huge volume of data temporal data and over a much larger scale and different dimensions right. So, this is the typical architecture.

So, one so what we see here this warehouse hills me to facilitate this data mining operations right. So, I could have done this data mining operation from the taking from the data warehouse structure also, but this primarily helped me to do that.

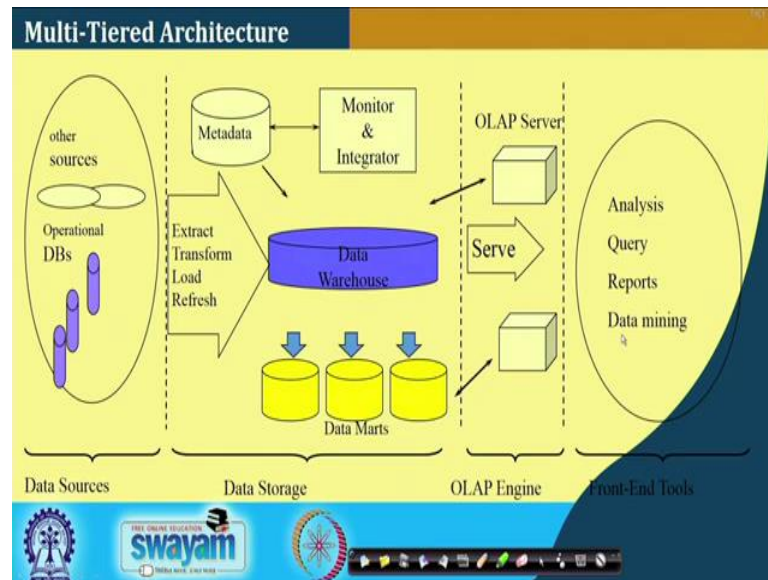
(Refer Slide Time: 09:13)



And if we look at the data warehouse development phases, so we have high level define a high level corporate data models and to be there right. So, there are different data marts which conformed to this model, there is a distributed data mart thing then multi tiered data warehouses. So, there can be enterprise data warehouse and type of things which also conform to this high level data models.

So, this is the way the warehouse overall development architecture fits in. With this background will come to the our other topic or which is which will be mostly one of the backbone for analysis of the things is the data mining.

(Refer Slide Time: 10:06)



So, what we see if you look at the thing this picture. So, one of the thing is that made mining the data, I want to mine the data to extract meaningful information's right. Where huge volume of data where usually a data warehouse facilitate I could have mined from other normal data set also right.

(Refer Slide Time: 10:20)

Data Mining

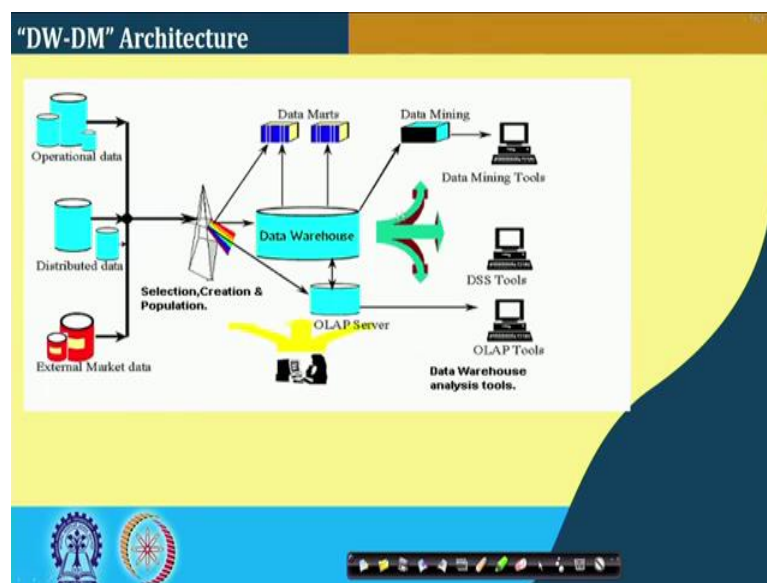
- Data mining refers to the discovery of new information in terms of patterns or rules from vast amounts of data
- **Data warehousing and Data mining**
 - The goal of data warehouse is to support decision making process
 - Data mining can be used in conjunction with a data warehouse to help with certain decisions
 - Data mining can be applied to operational databases but to make it more efficient and meaningful it is applied to data warehouses
- Data mining applications should be considered early during the design of a data warehouse

So, what is there so data mining refers to the discovery of new information in terms of patterns or rules from vast amounts of data right. So, discovery of new information in terms of pattern rules vast down data warehousing and data mining the goal of data warehouse is to support decision making support system. So, it prepare the data in such a way it can be solved

Data mining can be used in conjunction with a data warehouse or it can be used independently to help with certain decision, so mine and get the things. Data mining can be applied to operational databases, but it makes more efficient and meaningful apply to the data warehouse. It can be made to any operational databases or data stores and type of things, data mining applications would be considered early during the design of a warehouse.

So, in other sense I need to look at that what sort of queries or what sort of mining applications are there right, based on that I need to design the data warehouse right. So, that it is when you we designing the data warehouse. So, what are the frequent type of mining operations we require need to be there, for that matter even OLAP etcetera we need to; we need to looked in to the things though it comes hand in hand with the data warehouse suppress.

(Refer Slide Time: 11:38)



So, this is what we see it is a picture taken from some internet resources. So, we have operational database selection creation population then we have data warehouse, there

OLAP servers online analytical processing. Then we have data mining and data mining tools which acts on it and there are other tools like decision support tools OLAP tools and type of things.

So, there are different types of the tool different type of what to say processing, one of them is the data mining which takes care of the data warehouse again we want to emphasize. Data mining could have worked on the traditional databases also right, transactional data sets also right or some data repository also it mines you all your data, but it makes more sense or if you see when you talk about with the data warehousing type of things.

(Refer Slide Time: 12:30)



Data Mining and Knowledge Discovery

- Knowledge discovery in databases (KDD) --- more general than data mining
- KDD process consists of six phases
 1. Data selection
 2. Data cleaning
 3. Enrichment
 4. Data transformation
 5. Data mining
 6. Display and reporting
- Example
 - Consumer goods retailer
 - Association rule: whenever a customer buys product X he also buys product Y
 - Sequential pattern: whenever a customer buys a camera then within six months he buys photographic supplies
 - Classification trees: credit-card customers, cash customers, etc.

So, when we talk about data mining it is basically extracting information for meaningful purpose on, obviously another things come into play is the handy now is the knowledge discovery. So, knowledge discovery in database or sometimes we refer to that KDD type of things, more general than data mining. KDD process consists of six phases data selection, data cleaning, enrichment, data transformation, data mining display and reporting right.

So, these are the aspects and when you talk about spatial data mining we see that this type of things is comes automatically into the process right. We need to have that which data I required to be selected the cleaning operations would be there enrichment of the things data transformation data mining display and reporting as may there.

So, like one commercial application like goods retailer. So, one of the mining rule is the or one of the widely used is the association rule. Whenever a customer buys product x he also buys a product y right whenever I buy butter bread I buy butter, so it is a association rules right. So, this is while mining the data I can find out that this association is there.

So, what we will try that I will bring these two things together. Whenever there is a thing called some disaster or some natural calamity like flood it is likely that it will be follow up with some typical diseases which are maybe waterborne diseases and water contamination related diseases. So, this sort of disease or sometimes epidemic may fall out after this flood type of things right.

So, this is analyzing data particular thing. So, the region to region this may vary alright. So, those things has to be there is association if this happens this is likely to be happens. So now, whenever we talk about association you will say that how much confident it you are or how much support you are having alright. If you by taking two or a couple of tuples if you are telling it is may not be there. So, support and confidence comes hand in hand to measure of this data association rule miners right.

So, that is one is association rule another is a sequential pattern, whenever a customer buys a camera then 6 months down the line he buys a photographic material maybe the things right. Whenever we buy a color a printer color black and white laser XZ down the line after based on my printing every three to 6 months 8 months I required the cartridge right, so that is a sequential pattern.

If this happens this is going to happen after some time right, I can say if there is a road blockage or road accident etcetera. After some time there it will be congestion in that part of the things right, so there it can be different type of aspects. Classification trees credit card customer cash customer etcetera. So, that I try to classify into different categories right.

So, from that data mining I want to classify this is the things which are different category, I can say that if every departmental source to has different type of things like one garment section food section and different type of things. Then I can say this type of category falls into these and type of things or even particular section also some other subsection or I want to say that a cells of particular into the things right, so there can be

different aspects. So, association rule sequential pattern typically these classification trees are some examples of this mining categories.

(Refer Slide Time: 16:18)

Goals of Data Mining

- **Prediction** --- data mining can show how certain attributes within the data will behave in the future
- **Identification** --- data patterns can be used to identify the existence of an item, event, or an activity
- **Classification** --- data mining can partition the data so that different classes or categories can be identified based on combinations of parameters
- **Optimization** --- one eventual goal of data mining may be to optimize the use of limited resources such as time, space, money, or materials

So, one what are the goals of the data mining prediction right, data mining can show how certain attributes within a data will behave in future. Identification data patterns can be used to identify the existence of a item event or things identification. Classification data mining can partition the data, so that the different classes or categories can be defined based on the combination of the parameter right.

Optimization one eventual goal of the data mining may be to optimize the use of limited resources such as time space money etcetera. So, that these looking at the data mining I can do for other optimization things. So, one is Prediction one is Identification Classification and Optimization these are some of the major goals of a data miner.

(Refer Slide Time: 17:11)

Knowledge Discovery during Data Mining

- Raw data \Rightarrow Information \Rightarrow knowledge
- Deductive knowledge
 - Deduce new information based on applying pre-specified logical rules of deduction on the given data
- Inductive knowledge
 - Discover new rules and patterns from the available data
- Data mining addresses inductive knowledge
 - Discovered knowledge can be
 - Unstructured like rules or propositional logic
 - Structured like decision trees, semantic network, neural networks, etc

And knowledge discovery during data mining raw data to information to the knowledge. So, from the raw data we pre process and then using things I can have a knowledge like I find a pattern I find a association. So, that is a discovery knowledge, so it is not in the raw data neither is a in the information.

So, that can be deductive the deductive knowledge, deduce new information based on applying pre specified logical rules of deduction on that particular given data right. I can have a deductive knowledge or inductive knowledge discover new rules and pattern from the available data, I can have a new rule or pattern from the available data. So, I can inductive knowledge, so data mining at addresses by primarily the inductive knowledge.

So, finding out inherent pattern into the data set which is apparently not visible right, some sort of a what we say interesting patterns into the things. So, discovered knowledge can be unstructured like rules and proposition logic or structure like decision trees, semantic network or neural network and so on so forth which are more structure. So, this can be both inductive rule.

(Refer Slide Time: 18:27)

Types of Knowledge Discovered

Knowledge discovered during data mining can be described as

- **Association rules** --- correlate the presence of a set of items with another range of values for another set of variables
- **Classification hierarchies** --- create hierarchies of classes
- **Sequential patterns** --- sequence of actions or events
- **Pattern with time series** --- similarities detected within positions of the time series
- **Categorization and segmentation** --- partition a given population of events or items into sets of "similar" elements.

The slide features a yellow background with a blue header and footer. The footer contains two circular logos on the left and a video inset of a man speaking on the right. A Windows taskbar is visible at the bottom of the video inset.

So, types of knowledge discovered we some of them we already discussed, one is the association rules. Correlation of the presence of a set of item with another range of values for another set of items right this is the correlation. So, if these items are there then these items sets are going to happen right it can be one to one or means a set of items with a set of items right.

Classification hierarchy create hierarchies in the classes right, sequential pattern sequence of actions or events. Patterns with time series similarities detected within the position of the time series right. Then categorization and segmentation partition a given population of events or items into the set of similar events right.

So, that can be partition and type of things right, these are the things which are there association rules classification hierarchies sequential pattern, pattern with time series categorization and segmentation right. So, these are the things which we can which we see that type of knowledge discovered by the in using the data mining paradigm.

(Refer Slide Time: 19:44)

Association Rules

- An **association rule** is of the form $X \Rightarrow Y$
 where $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_m\}$ are sets of distinct items
 The rule states that if a customer buys X, he is also likely to buy Y
- The set $LHS \cup RHS$ is called an **itemset**
- **Interest measures**
 1. **Support** (*prevalence*) for the rule $LHS \Rightarrow RHS$ is the percentage of transactions that hold all the items in the itemset.
 2. **Confidence** (*strength*) for the rule $LHS \Rightarrow RHS$ is the percentage (fraction) of all transactions that include items in LHS and out of these the ones that include items of RHS.
 - Confidence is computed as $\text{support}(LHS \cup RHS) / \text{support}(LHS)$

Now, association rule those who are knowing fine, otherwise we say that X a set of item set Y set of item set are subject of distinct item sets right. The sets of distinct item set the rules state if a customer buys X he is also likely to buy Y like bread butter or if bread and butter two items then likely to butter into the by jam alright.

So, these are the things are there, if it is binding a customer is binding say bed or pillow etcetera likely to buy bed cover pillow cover and type of things right. So, these are which are item sets right. The set left hand to right hand side is called the item sets. So, these are individual items set and the whole item set is the this implies these means the whole the item set.

So, what need I measure how do I measure that who are related more how this association rule. So, two standard measures are there one is that support or prevalence of the rule in the left hand side right, that is right hand implies right hand side is the percentage of transaction that holds all the items in the item set. So, if I have those type of things that support for something, see I say bread butter implies purchase of jam and say something cookies or something alright.

So, I have this things left hand side implies right hand side. So, this is the percentage of the transaction that holds all the items. So, this has the item set like bread, butter, cookies, jam. So, these items set out of that old transactional things what is the particular support, all the items in the item sets of the all the items. So, percentage so that is the support. That means, what I am to say that this rule what I am trying to find out is

presence in so many percentage of the things right. I can say that my support should be more than 60 percent, support should be more than 70 percent that is frequently occur in thing.

And another is the confidence or the strength, that for the rule left hand side implies right hand side is the percentage fraction of all transaction that include the item set in the left hand side and out of these ones that include the item set in the item set right. So, all the transaction where the in the item set in the left hand side and those which include the item set on the right hand side.

Now, confidence is if we see that compute at support left hand side union right hand side divided by support of left hand side. So, it can be way of different. So, item set have two aspects one is that support another is the confidence in the things.

(Refer Slide Time: 22:57)

Example		
<i>Tid</i>	<i>time</i>	<i>Items</i>
101	6:35	milk, bread cookies, juice
102	7:38	milk, juice
103	8:05	milk, eggs
104	8:40	bread, cookies, coffee

Consider two rules $milk \Rightarrow juice$ and $bread \Rightarrow juice$

Support {milk, juice} is 50%

Support {bread, juice} is 25%

Confidence of $milk \Rightarrow juice$ is 66.7%

Confidence of $Bread \Rightarrow juice$ is 50%

Like for example, here the there are transaction there are four transaction t 01 t 1 t is 101 102 103 and 104 are the four transaction and these are the time of transaction 6:35 6:38 something 6:35 pm etcetera. And the these are the items which are being transact like meal, bread, cookies, milk and juice milk and eggs bread these etcetera right. So, these are or we can say that these are different in transaction item, these are the things which has been bought in particular shop or type of things.

Now, consider the two rules milk implies juice, but in other sense if a person is buying a milk he will purchase a purchase juice also alright or bread to juice. If it purchase bread it is likely to purchase juice right. So, these are the two things a implies b. So, if I say a implies b, so this is item set b item set bread item set juice item set. So, this what do I what we want to say how they are associated.

So, in other sense how much support and confidence we are having that these association holds right. If I am having this then there is a knowledge, I can say that if this is a purchase is going on or this is activities goes on this. Like you say that by default we say if there is a road blockage there will be congestion in that road.

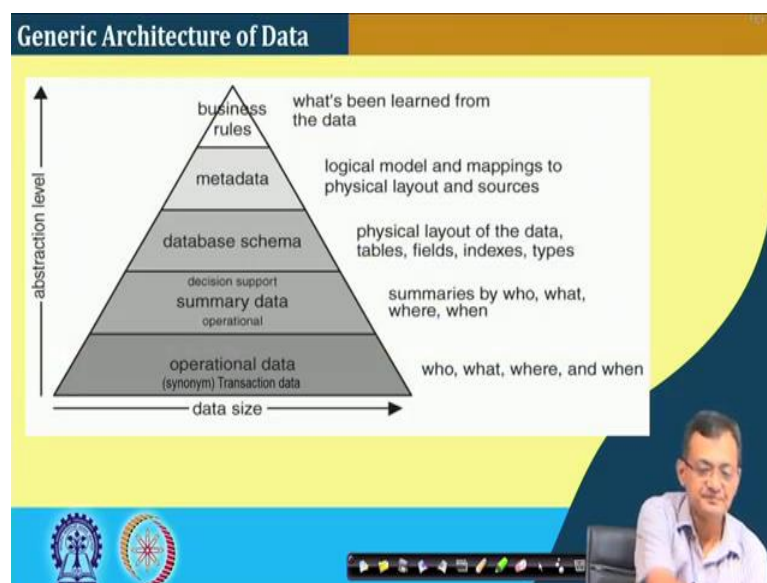
But that may be a road blockage due to some accident etcetera and that may be a congestion in some part of the network alright. Due to there is a perpetual things which are inherently not visible from the transactional data set by this association rule may say or I can say the during this time of the things the day there is more congestion or less congestion apparently which is not visible into the things right. So, these are the things which are may be there.

So, here if you see support for milk, juice is 50 percent right. So, out of four transaction milk juice is twice two places. So, 2 by 4 50 percent, support for bread juice bread and juice is there is 25 percent out of four one transaction is there. In reality there will be huge volume of transaction. So, there will be any way calculation if so this is just a customized tailor made calculations. So, this now confidence of milk to juice is 66 percent right.

So, if the milk is there juice is also there in this case is 66 point and sorry. So, the number of cases where milk is there and milk to juice is there, see milk juice is 2 milk is on the left hand side milk is there is 3 2 by 3 is 66 percent right and bread to juice is 50 percent right. So, bread is there in 2 places and bread juice is 1 places, 1 by 2 is 50 percent. So, this is the support and confidence.

Now, if you have a things that I require this much support from my transactional data sets and this much confidence that this will be bread to juice implies will be there. That means, I am selling that these activities implies that set of events or this type of inventory is so much support and confidence is there. So, that we require in case of this type of support and confidence. So, ad association rule is ready to be support and confidence.

(Refer Slide Time: 26:48)



So, now let us try to come back to our things that why we need all those things. Now this is a very generic diagram many of you might have seen in different places taken from again it. So, my operational or transactional data is that who what where when type of things, whether is spatial data non spatial data we store this data like this.

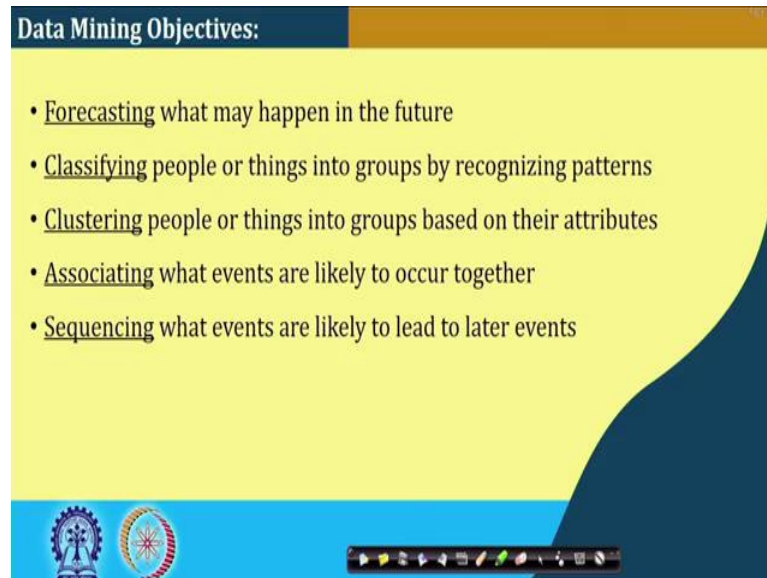
Then we have a summarization of the data right summary data, summarizes by who what something right summarizes by some aspects. Then we have a data schema physical layout of the data tables fields indexing and then we have a logical data and mappings of the physical layout and sources right.

So, meta data then we have a business rules. So, this is my objective from going from, this is our basic objective when we deal with data from this data operational data to business rule. Our this data mining and warehousing also warehousing these are this data mining helps us in finding out meaningful information is there.

So, one is the transitionally I know that these are the things analysis, another thing is that some interestingness measure comes out like this type of information may be interesting to look at it. Like if this happens this is going to happen and type of things, there are different type of if you look at the special country there are different type of studies like day to day traffic related things, there are studies on the things called say migratory bars or different aspects of the things right. Movement pattern of a city or particular climatic

prediction and type of things right. So those are the things which are there in these aspects of data mining aspects.

(Refer Slide Time: 28:38)



Data Mining Objectives:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events

So, if we look at our broad objectives of data mining, why we are doing there is one is forecasting what happened in future like forecasting. If these are the things whether with addition with our data sets we can say that what happened classifying people or things into groups or recognizes recognizing pattern so Classifying.

Clustering people or things into group based on their attributes, Associating what elements are likely to occur together, Sequencing what elements are likely to lead to other events see these are very interesting facts or interesting what we say metric or measure which are otherwise not like our traditional queries. Like forecasting into the things finding that what will be having in future or predicting something right.

So, classifying like categorizing into the things, clustering people things into groups based on their attributes. So, classification things into groups based on recognizing pattern, associating what happens what events are likely to occur on a to occur together. So, that is associating things right this if happen etcetera. There are different type of things that if this happened that may immediately happen or after on a temporal scale something after some time it happened and so on and so forth or this is happen there somewhere is something will happen.

Sequencing what elements are likely to lead to later events. So, these are the sequencing of the things. So, these are different mining data mining objectives which are we are trying to achieve in this type of studies or goal alright. So, this is the thing and if you if we try to look at this aspects, like if we now look at the thing at the same with the spatial data mining. So, this type of techniques will be using right.

So, association will mining or pattern type of things which are used to find out that different interesting measures or interesting facts in the spatial data. As we have seen spatial data is pretty large right large in size and with vary type of component or influencing factor, this may help us looking at the data of a or space time data spatio temporal data to have different type of interesting measures alright.

So, in today's lecture what we try to look at is the basic of data mining aspects we service the warehouses in place. So, data mining again can work on any data set. So, this is a type of rules and operations we look at to find out some of the interesting facts on knowledge from the data sets right. So, more deeply we will look into in coming on to lectures, that how this spatial data mining we service things are there all things will be true.

But the little the dimension wise it will be larger or different aspects need to be seen. So, let us conclude our discussion today and continue in the subsequent lectures on the same thing.

Thank you.