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

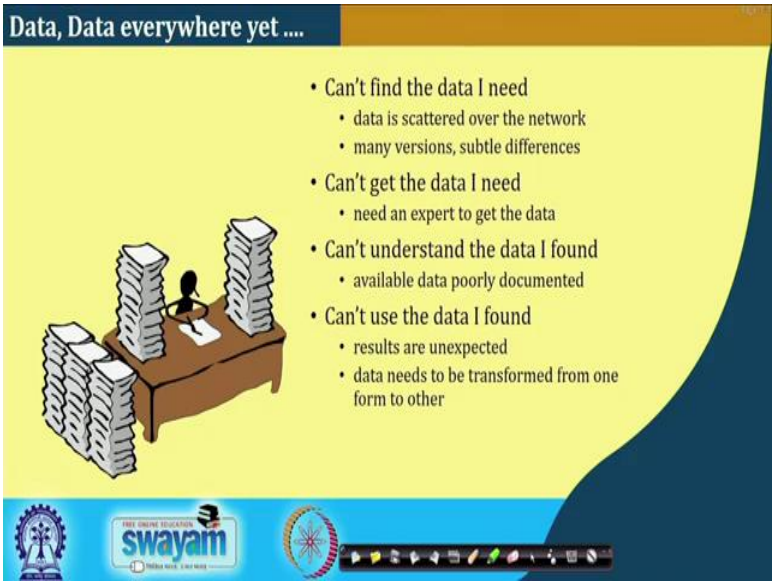
Lecture - 26
Spatial Analysis - I

Hello, today we will discuss on one major aspects of Spatial Informatics. So, long if you see that we talked about different things like one is that how data is stored modelled and also though we worked on spatial networks. So, different category things next couple of lectures or few lectures we will be discussing on this.

So, to say Spatial Analysis actually it will have lot of other things means lot of not other things lot of different dimension like looking an spatial data mining, then working on the spatial autocorrelation spatial computing. So, or also if feasible if we look at a so the spatial data analytics or the spatial data analytics aspects of the things right.

So, today specifically we look into one aspect of spatial data warehouse or more of a spatial data warehouse and data mining. Before that we I thought that it will be good for many of you who are not gone through for formal data warehousing or data mining courses. To have a background or the basics of data mining data warehousing and data mining few lectures on those things right. So, basically we plan to cover these aspects in next few lectures and today most specifically we look at the data mining a data warehousing and data mining.

(Refer Slide Time: 01:52)



So, if you look at today's data sets, so we have huge volumes of data right. So, as most of the most of our things are digitally enabled, so there are huge volume of data which are in place. The problem is now how to retrieve meaningful information from it which I need right, there may be a lot of sensor data. But for my particular purpose it may be research it maybe some metallurgical analysis etcetera, I need a something a data which mean for that.

So, cannot find the data I need data is scattered over network that several form of the data many version there are different sort of subtle differences right. So, there is a serious problem. I cannot get that data I need. So, first of all where to find the data, then how to retrieve or get the data need an expert to get the data right. So, I need a expert to tell that this data is particular for this purpose and so on I can understand the data I found alright, available data are poorly documented.

So, if may get a huge volume of data, but the available data is poorly documented cannot use the data. I found right results are unexpected like there are they are need to be clean they need to be find out that whether the data there is a any error in the data or there is something we need to have a appropriate scaling down of the data and or appropriate scaling of the data.

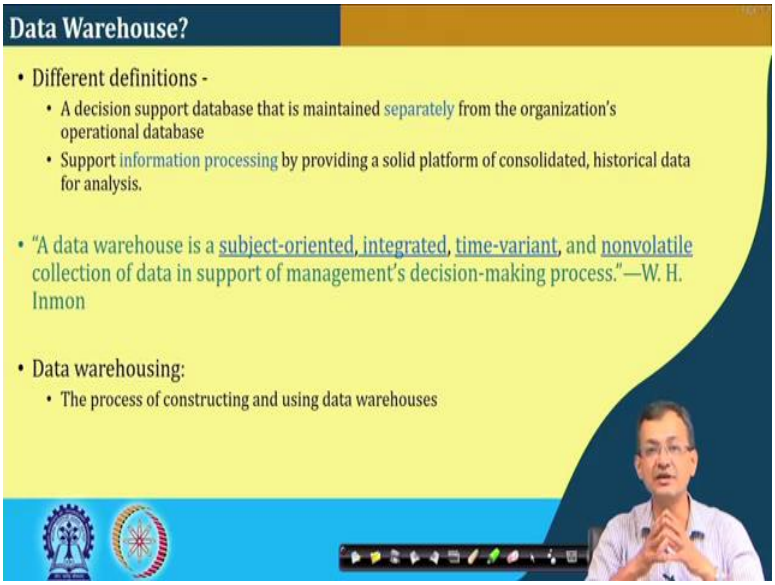
The data need to be transformed from one form to other to do that. So, this sort of problems we are facing, especially we when we want to do a large scale data analysis so to say.

(Refer Slide Time: 03:38)

So, in order to handle this so we may require some sort of a appropriate mechanism to store manage this data or what we say warehousing of the data. So, this is a generic statement or generic curve you might have seen in different aspects. So, what from the data we want to go to a informations right it is true for spatial data, it is true for non spatial data also right it is true for any data right.

So, a process of transforming data to information and makes it available to the users in a timely manner to make a will make a difference right. So, appropriately if we can make it to information etcetera. So, in other sense before I process the data, so data need to be so to say uplifted to a form, where it can be easily process able. So, that is one aspect of the sort of warehousing aspects right.

(Refer Slide Time: 04:38)



So, there are so when if we say what is warehousing there are different type of Definitions. A decision support database that is maintained separately from organizations operational database right. So, operational database means it is a transactional database is going on right, there are like say a banking operation. So, all these counters all the activities going on these are the operational data.

But if I want to know that profitability of the bank or how the bank growth is going on and etcetera that I need to work in the look at the some form which is not it definitely the input come from the operational data. but I need to analysis data in a some other form alright. Like I want to do a traffic analysis of a particular road network say Kolkata or Kharagpur or Pune or somewhere right.

In order to do that I one is the recording of the data of every instances right, either it can be the GPS from the vehicle or it is recorded at the junction by this intelligent sensory system like it may be cameras or other type of vehicle count etcetera and then but those are what we say operational data. But in order to do that what is the pattern of traffic movement on a particular time of a day on a particular day of a week, I need to analyze those data in a sample.

So, I required in order to write run this analytical tools I require a data which is not that the operational data, it need to be what to say transformed to a some other founded. So, that is why we are telling that has to be maintained which is more amicable for decision

support. Support information processing by providing a solid platform of consolidated historical data for analysis right.

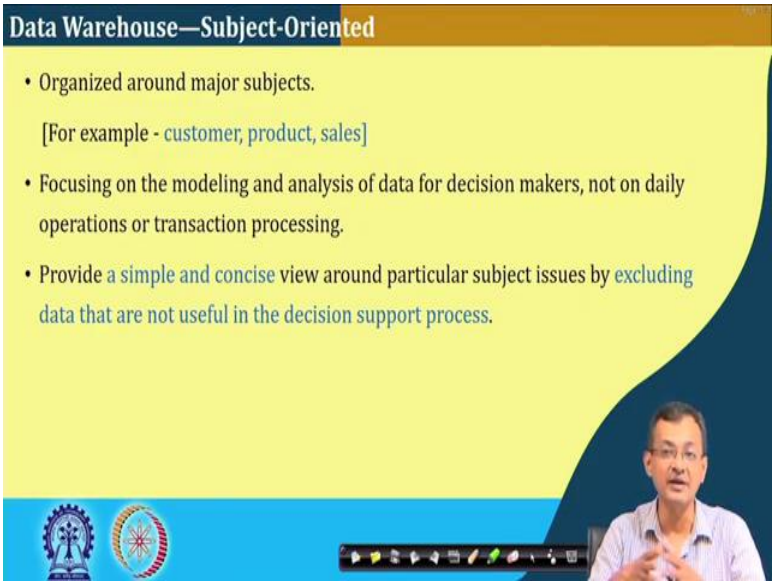
So, a data warehouse is a subject oriented, integrated, time variant and non volatile collection of data in support of management decision making process. So, that is one of the very state cut. So, it is something you see we can see that something highlighted right. Object oriented subject oriented this for a particular purpose integrated like I could able to integrate data from different source, time variant it is there is a temporal scale also and non volatile, it is not like that it will volatile and goes up.

Usually our one we go for operational data transaction that is the that instance the dataset is there right others dataset I overwrite on the things. It should be archived in a separate fashion, but the access the traditional data processing thing they are not responsible of maintaining those data per se right. So, there should be a data management process to handle that. But warehouse is non volatile, so the ones it is in warehouse it is a non volatile data right and it is.

So, if you see that if you look at the spatial data sources. So, this these requires fits in right it should be particular change like if I work on the transportation, I look at the transportation part of it right. So, I if I the same or not exactly the same data, but a good amount of overlap data or the same data sets, if it is used for disasters management only to look at those point of view right.

Planning for a longer time that to city planners etcetera that is we looked into different aspects. So, what we do we need to warehousing like we do warehousing of other products, the process of constructing and using data warehouse as we are doing that ware warehousing of the data alright.

(Refer Slide Time: 08:11)



So, what we say subject oriented organized around major subject, for example customer product sales if you have a things right. Like if I go to a departmental store or when or if we consider things it is a customer base particular product, sales of a product and type of things. Focusing on modelling and analysis of data for decision maker not on daily operational transaction other things right, it is more of a decision support type of things right.

Provide a simple and concise view around particular subject issues particular subject by excluding data that are not useful in the decision support process for that particular things. So, I can I may have all the data, but I can exclude those data or I can take an aggregated form of the data alright. So, that exactly what we say when you look the subject oriented, it should be integrated right.

(Refer Slide Time: 09:08)

So, it is constructed by integrating multiple heterogeneous data sources right. So, it is it can be different type of data sources coming into picture, like consider like I want to look for a traffic movement on a particular segment of a city network right. So, one of the data sources may be the time of the thing that what I can have that previous data based scenarios. I need to look from other type of things right, like it is I made to look for the weather condition if it is a rainy something if it is a normal day something.

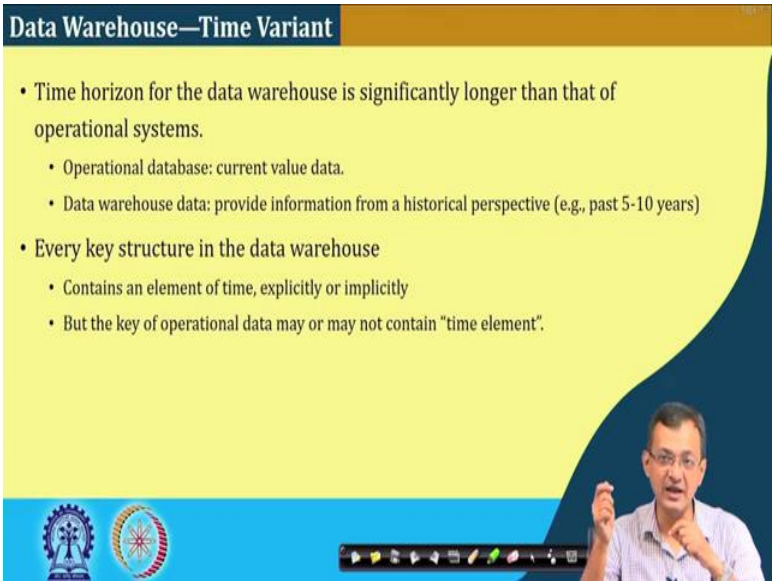
I need to look at that which time of the period like if it is a middle of the night some way it is the middle of the day some other things right. I also may look at two other aspects like if it is a festival seasons some aspects, if it is a week day weekend. So that means, there are different heterogeneous data sources which need to be integrated. So, there may be traditional relational database for flat file or online transactional databases which are which can be there right.

So, what we see that it is a integration of different data sources, data cleaning and data integration techniques are applied. So, I need to clean the data because as volume of data there, so there is a need to clean the data and data integration techniques how to integrate the data right. It may be I need a unified model if you remember that beginning of this things we, if we have a applications schema or a integrated model that it needs to fits in it.

Ensure consistency naming convention, encoding structure, attribute measure etcetera. So, you need to ensure among different data sources, so what in other sense quote

unquote interoperability needs to be supported. If you remember the spatial web services and others there where we talked about this those issues right where we things. The data is moved to the warehouse it is converted to a particular form which is amicable for decision support system for that particular organizations right or for that particular field. So, that is need to be that particular to be in the warehousing of the data.

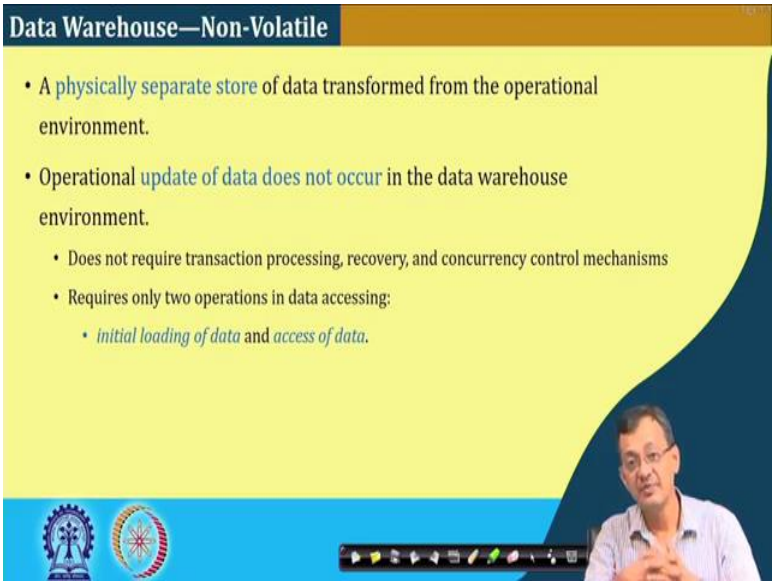
(Refer Slide Time: 11:19)



Next is data warehouse is time variant right, so time horizon of the data warehouse is significantly longer than that of the operational data. So, it is say operational data is a daily transaction or even may not maybe within that particular duration of the thing. But warehousing is much more that is yearly or over 5 10 years even monthly and different granularity I can go back scale down and scale up on the different scale right.

So, operational database is the current data value data, data warehouse data provide information from a historical perspective, that is 10 years 5 years type of long term data sets. Every key structure in data warehouse contain an element of time explicitly or implicitly. So, that is whether it is explicitly define a time access or implicitly it is the time access is there. But the key of operational data may or may not have the in case of operational data, we may or may not have that key concept.

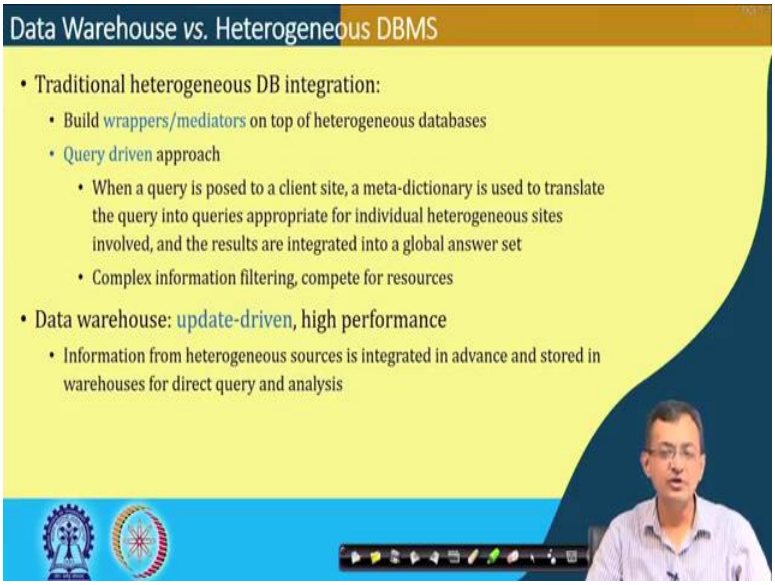
(Refer Slide Time: 12:28)



Data warehouse are non volatile that is another aspects. So, a physical separate store or the data transform from the operational environment. So, it is non volatile, operational update of the data does not occur in the data warehouse environment. Does not require transactional processing recovery and concurrency control mechanisms. Requires only two operations that is initial loading of data and the access of the data right.

So, it is what we require mostly that loading appropriate loading of the data in the particular form. So, that it is things are there definitely data cleaning etcetera will come into play and how to access the data alright. Maybe on the temporal scale, say I want to access before 5 years, 3 years, 10 years data if it is there the how that accessibility mechanism needs to be defined in the warehouse.

(Refer Slide Time: 13:22)



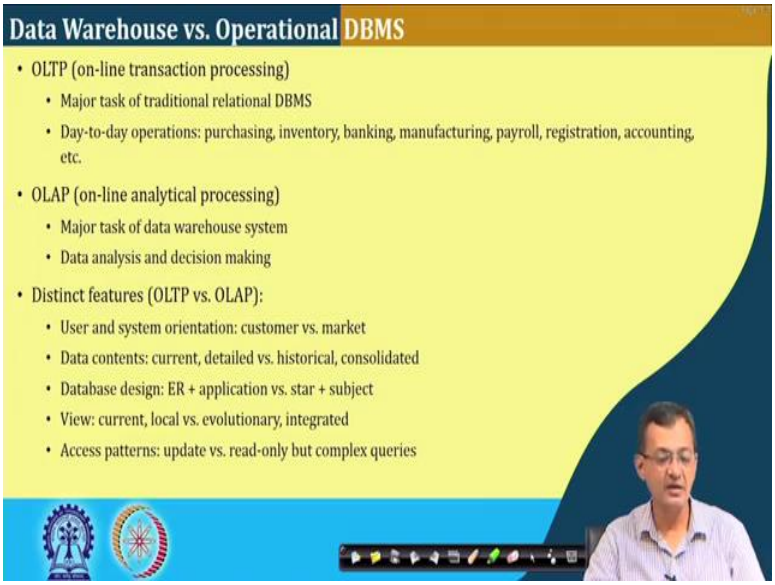
So, if we so this as we have seen these are the basic core properties. So, to say it is non volatile time variant integrated and subject oriented. So, these properties needs to be there right and obviously the question come that what is the gross different between warehouse and heterogeneous database management system right.

So, traditional heterogeneous DB database integration build wrappers and mediators around the top the other things. So, heterogeneous database is there I make a mediating layer which takes care of these putting them into format and type of things accessibility of the data and etcetera. And it is mostly query driven approach when the Query is posed to a client site the meta dictionary is used to translate the query into the queries appropriate for the individual heterogeneous sites.

So, the and the results are integrated into global answer sets. So, complex information filtering complete for resources, so these are the things right. Data warehouse so it is update driven high performance, so it is update driven and high performance. Information from heterogeneous sources is integrated in advance, that means I basically do some sort of a pre processing and stored in a warehouse for direct query analysis.

So, it is as if I what we say pre materialized the thing right, keep it in a form which can be quote unquote half cooked. So, that I can easily solved this decision support systems right, which is the traditional our heterogeneous database system is not meant for that.

(Refer Slide Time: 15:03)



So, another thing is that what is warehouse versus operational DBMS or queries type of things. So, it is like online transactional process processing it is a major task of the transactional traditional relational DBMS right. Online transactional processing day to day operation purchasing, inventory, banking manufacturing payroll, registration, accounting these are all operational OLTP type of operation is there.

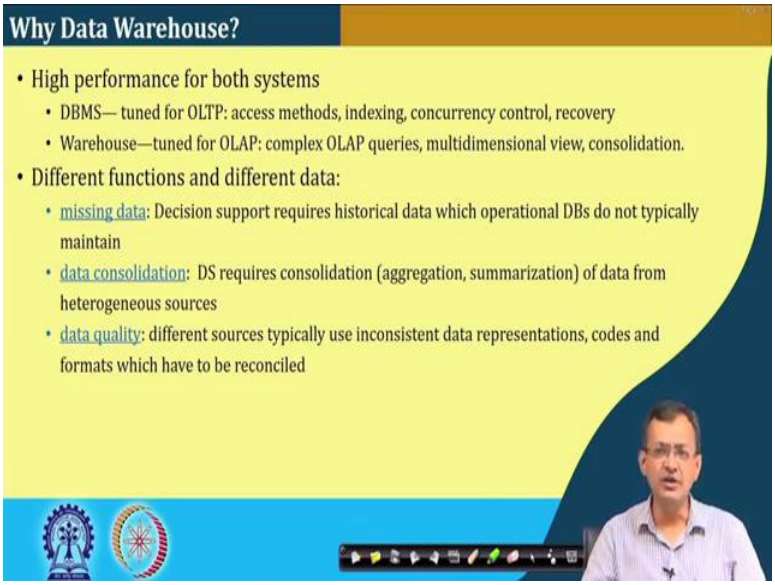
Whereas, what we look at this warehouse and other things is more of OLAP right. Online analytical processing major tasks of the data warehouse system data analysis and decision making is the major tasks. So, the type of what we say analytical, where that is more of a transactional processing right.

So, feature are OLTP versus OLAP, user and system oriented customer versus market, data content current details versus historical and consolidated database design ER plus application versus star plus subject, view current local versus evolutionary and integrated and access pattern update versus read only but complex queries right. So, these are typical things right.

So, one is user and system oriented the customer versus market, data content is in case of our OLTP is more of a current and detail and it is historical and consolidated. Aggregation of the data is there, database design here we have ER diagram application versus star plus subject view current local versus evolutionary integrated and access pattern it is update versus read only, but more complex queries.

So, usually data warehouse updates are not like transactional updates right. So, data warehouse is something which need to be preserved for a much longer time etcetera. So, the updated set there has to be very much regulated and usually it is note down on the offline right. It is from the I from the transactional data set of that data sources, I do a pre processing and put it on the warehouse where mostly the read only queries will be there.

(Refer Slide Time: 17:18)



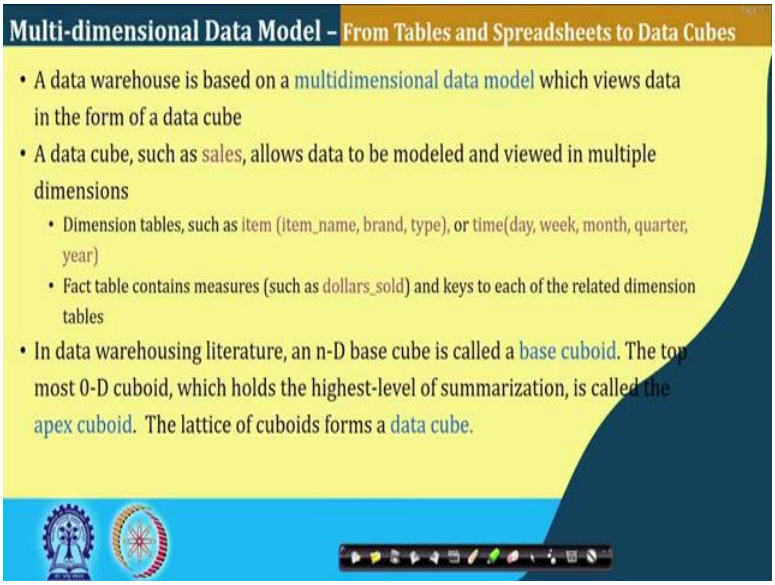
So, why we require a warehouse that is implied say high performance for both systems DBMS tune for OLTP access methods, indexing etcetera. Warehouse tuned for OLAP like online analytical processing and complex OLAP queries multi dimensional views and consolidation these are the major fact of warehousing. Different functions and different data missing data decision support requires historical data which operational DBMS do not typically maintained right.

So, we require historical data and if there is some missing data in the historical pattern then like I required ts from the t 0 t minus 1 t minus 2 etcetera, if that is t minus 2 t minus three instances are missing so that is to be there. Data consolidated requires consolidation aggregation and type of things and this will be DW data warehouse requires consolidation data from the heterogeneous sources.

Data quality different sources having typical use inconsistent that representation quality have heterogeneous data sets, so that need to be here absolutely consolidation into. So, that the appropriate data quality is maintained because, all later on pricing will be based

on this how quality data is there right. So, different function different data things are there.

(Refer Slide Time: 18:45)



So, for this what we see that from our this I require a multidimensional data models right, from the tables and spread sheet to data cubes. So, that is the one thing what we require. So, data warehouse is based on multidimensional data model which views data in form of a data cube. A data cube such as sales allows data to be modelled and viewed in multiple dimensions right. So, we will see that how a data cube looks like so it is a cube of data right.

So, dimension table such as item name brand type or time day week month quarter year. So, these are different dimension tables, like a time has different dimension say you require a daily sale you may require weekly sale you may require monthly sale you may require quarterly sale or yearly sale right, all are not aggregated in the things right.

When we do like I may take the daily sales to make a monthly sales, but as such weekly sales may not fir in to the monthly sale, like a some month it may have different type of different days like 30 days, 31 days sometimes 28 days, 29 days and type of things. Even that way week to year matching will be again difficult right exactly it does not fix.

In other sense like I if I have a concept hierarchy going up and down in the things will be there connected. But I can I what we see that we can have different type of dimensions of

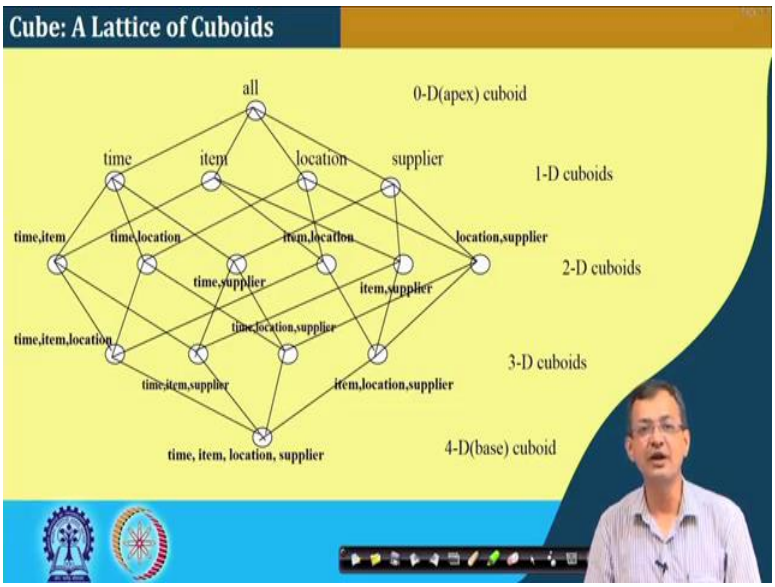
the things right. There are there is a concept called fact tables contain measure right. So, if what I want to achieve like I want maybe like dollar sold or total revenue and key to each of the related dimension tables right, so that type of things.

So, in a data warehouse literature so there n D base cube. So, n dimensional right 1 D 2D 0 D base cube the top of the thing is the 0 D. So, aggregated everything all. So, for all year for all products for all say item set I want to find out the sales right. So that means, one point it can be having 1 dimensional, we say that for a particular time scale or a product scale or type of things it can be 2 dimensional and soon and so forth.

So, which holds the highest value of summarization, so I summarize that. The total the manager wants to know that what is the total sales of the things, I want to know that what is the total amount of traffic flown through this particular region I do not care I am not looking for temporal and type of things. I may want to say that total number of traffic during this part of the type or that number of say now number of buses or number of four wheelers or two wheelers during this time on this region and type of things. So, there is those things are different granularity come into play right.

So, how do I do it if I go from the transactional data then I have to do a lot of multi layer processing, otherwise I can keep a multi data multi dimensional data cube ready from where I can process this queries. So, what we say that lattice of a data cube forms a lattice of the cuboid forms a data cube.

(Refer Slide Time: 22:13)

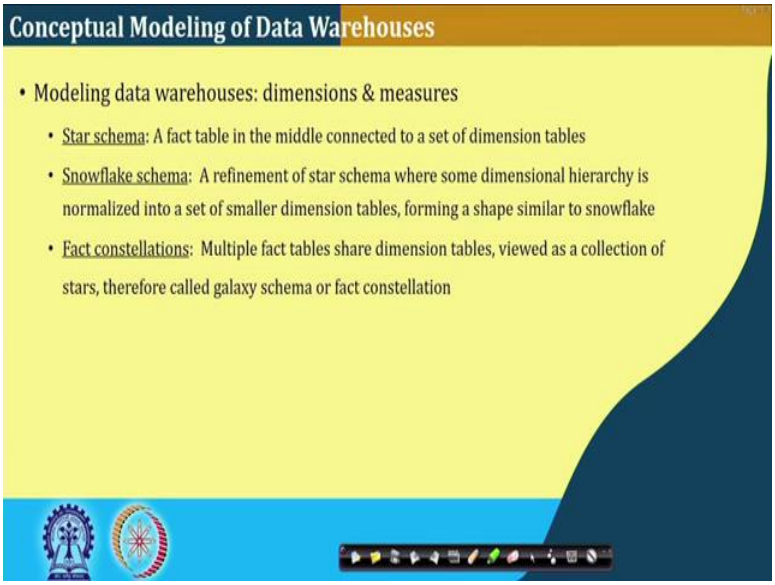


So, you see this is the thing if we look at, so there we have dimension time item location supplier. So, for a particular time for a particular item for a particular location for a particular what supplier this is the thing. Like at the other things that for all location for all time for all supplier and for all item what is the things, my objective may be finding the sales how much revenue is there. So, there is a for all.

Now, I have to say for a particular time for all item for all location for all suppliers. So, what is the thing say month of March, if it is my monthly thing or on particular day say today's date or something right or on particular week right. So, or I can say time and item for particular item and particular time for on location for all supplier, like here the dimension is 3D cuboid right time, item, location and then I have a 4D. So, all are there for a particular time period for a particular item for a particular location and supplier, so these are 4D cuboid. So, I can have different relations.

So, now when I now if you we can understand there is aggregation of the things are there and I can query at different things. So, warehouse should have a mechanism to maintain this data sets right. Now if you little look back to your traditional relational databases or the type of things what we do with it that, this is not exactly fit it into the things nevertheless I require this transactional database which the relational database to make this warehouse trading right. I require the data form I am not generating data here the warehouse is basically uploading the data in appropriate point where the queries can be there.

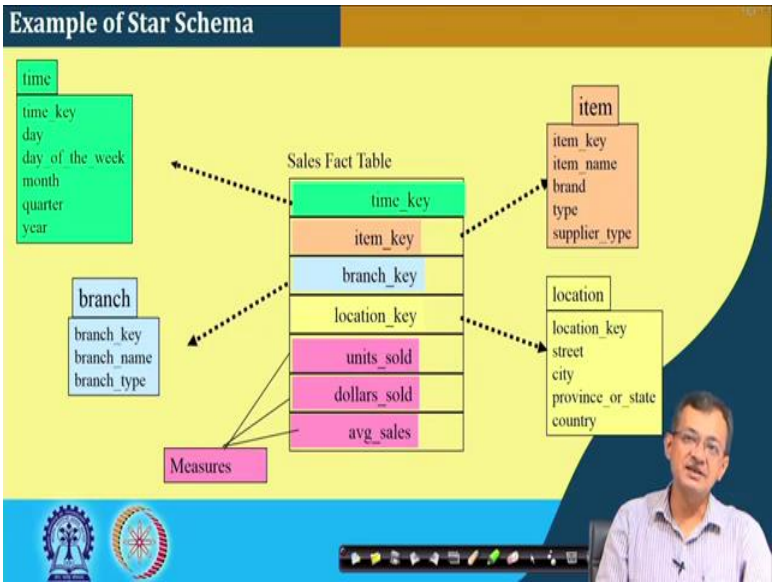
(Refer Slide Time: 24:09)



Conceptual modelling of the data warehouse, so how do I conceptually model. So, our traditional model may not work good to here. So, one is the star schema a fact table in the middle connected to a set of dimension tables right. So, we will see Snowflake schema, so this is another thing a refinement of the star schema where some dimensional hierarchy is normalized into a set of smaller dimensional tables and form a snowflakes. So, slow star is that one all connected star, snowflake is there again expanding right.

Another is the fact constellation table that multiple fact tables share the dimension tables viewed as a collection of the stars, therefore called galaxy schema or fact constellations scheme right. So, there are different things and connected into the things. So, that can be another way of looking at the things. So, we have star schema snowflake schema and the fact constellation scheme.

(Refer Slide Time: 25:18)

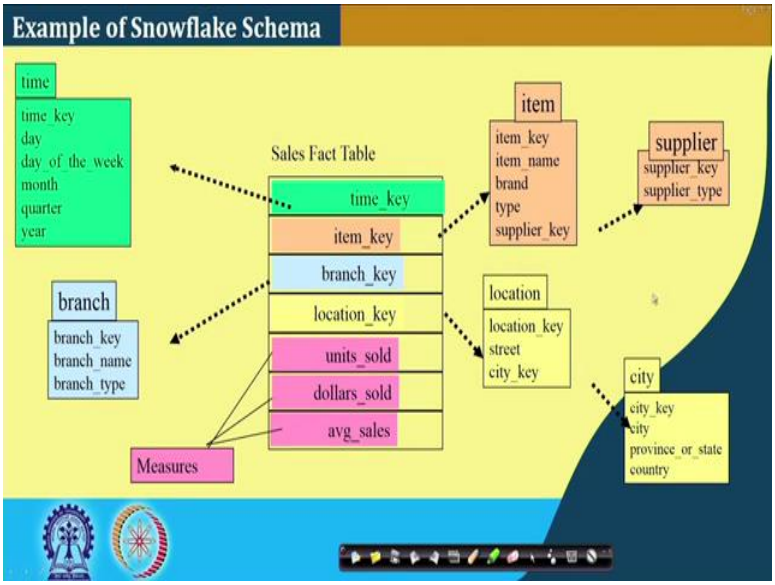


So, typically a star schema look like this, I can have different things like I have a time dimension, a branch key and the item key and a location key alright. So, for a particular time for a particular branch where particular think of a particular commercial chain even our chain of say different shopping or different malls and type of things which are a different place of the different places different cities etcetera and then different items are there and different location is there.

So, I have time in the time we have the time key which is the primary key day of the week month quarter year this is the time dimension, branch key name type is the branch dimension right item also and thing. So, you can say that particular time like for a weekly sale, for a particular type of branch, for a particular item set on particular location particular city or country or some province or state I can find out these things.

Now, if I could have stored then my query which is more of a decision support, we need to remember these are primarily the query which is main for the decision support system or supporting this higher management to do a planning and type of things right. So, this is typically a star schema right.

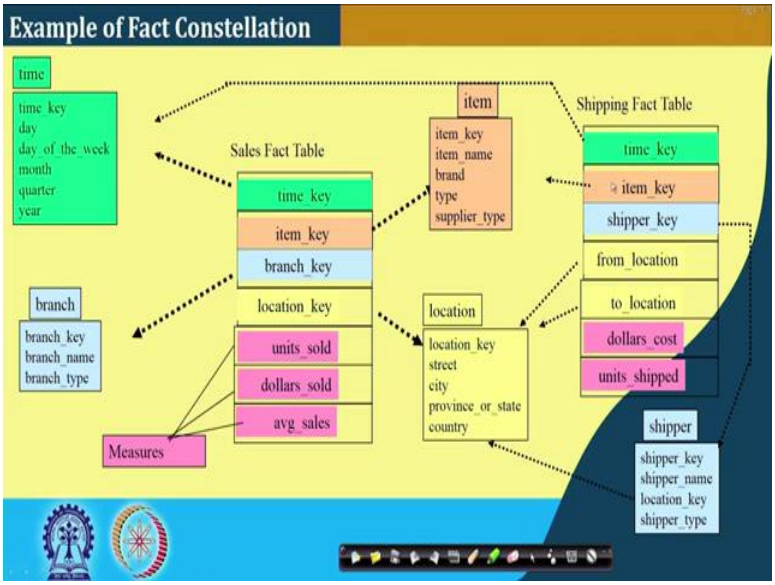
(Refer Slide Time: 26:52).



So, if we look at a snowflake schema you see it is the star schema, but some of the dimension are expanded. Like here from the I have a item which has a supplier key which links to a supplier or here is city key, city key beings to a particular things right city key city province state etcetera, so it is a snowflake schema.

So, one thing it is I just missed here also there are some units this is my goal, I want to find out units sold dollar and average sale this is my goal there are several goals based on your business policy. And here this is a measure to how to handle similarly here also we have same set of things. So, this is to when we took the spatial data set also we can have spatial item sets for a particular location and so and type of things right. I can have different type of goals into the or to be analyse on measure matrices.

(Refer Slide Time: 27:53)

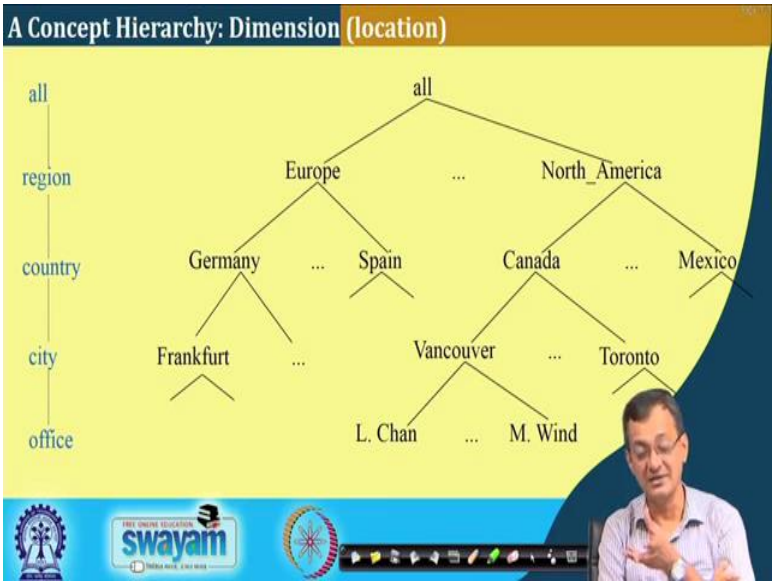


And this is what we say about fact constellation schema. So, it is again if you see this is the star like, but here I have a thing call in the location that what we not only item set. Item set is in time is the time key this is the shipping fact alright how the data are being shipped. So, it is the time key is connected to this time table alright. So, item is to this thing location is the location table form an two locations and it has a dollar cost unit etcetera for shipping right there can be a shipper right.

So, you see there are different so this is one sort of a fact table, this is another fact table right which are connected with different type of these dimension tables or the dimension tables which are connected to these two type of fact table, that can be multiple fact tables into the things. So, these sort of things is more of a what we look at a fact constellation.

Again come back, so one is that popular star schema it is just star like of things then we have a snowflake schema. So, it is star then again expanding and then we have a facts constellation schema, where we have multiple fact tables which are referring to different type of dimension tables and have the things like that. So, there are there can be different way of handling. So, is based on the need we can have; we can have this particular conceptual model of the data warehouse. So, that need to be plan a priori because, based on these the data warehouse will be build up right.

(Refer Slide Time: 29:29)

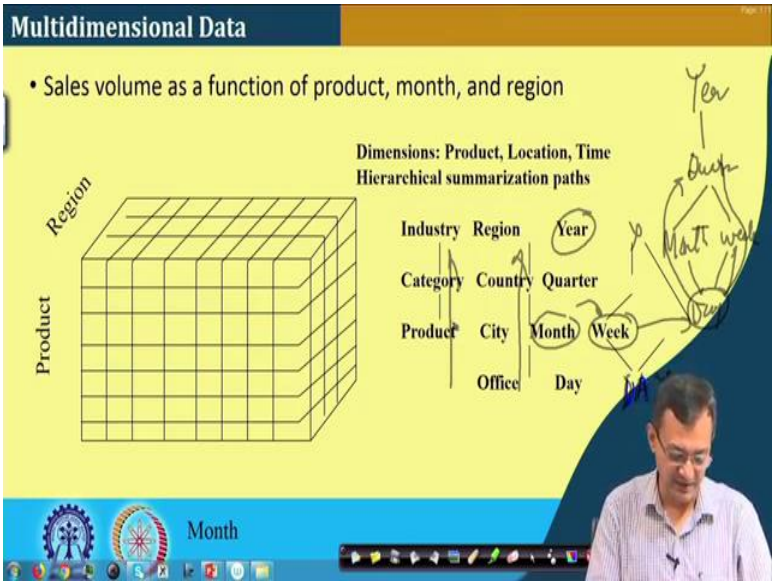


So, there is another thing called concept hierarchy like here if you say the concept hierarchy office to city, city to country, country to region to all as a departmental store or that stores are there it is for a particular city say Kharagpur or Kolkata. So, many stores or particular brand and then if I want to aggregate. So, if I want to find out the cells at the store level or the particular office level or outlet level, then at the particular city level country level region level all means at the top of the things alright.

So, like I can have a example a particular country over that it is a particular region, then country then particular city or state and then particular stores and go on. So, this is the way I can conceptually I want to do this hierarchy, how it is build up based on my analysis need or based on my requirement of my decision support system right. It could have been other way also I may not look at the city directly go to the country and type of things right.

So, that can be or I below city I can have a city regions right, so particular region southern part four or five blocks of the cities and type of thing. So that all depends at how you want to look at those type of things right. So, that the concept hierarchy based on that the warehouse is build up, because your query are over that targeted things.

(Refer Slide Time: 31:04)



So, you see this is a multi dimensional table. So, dimension product region and other this it is there it is little bit came down. So, it will be the temporal scale of the month out here. So, one is the temporal particular product in a particular region right. So cuboid, so every any cube in this cuboid represent that for that particular things what are the things. If I want to know all the regions then I basically aggregate this right on these two things will be there and if I have all the all have together then I can have for all type of things.

Now, you see I can take a piece of the thing either slice dies or a chunk of the things to find out the different type of suppose this is gives the cells right. So, product, month, region and we can have a different hierarchy like product can have a category and upon a particular industry, office, city, country, region time may have a different dimensions right. I can have it day will be here day and then this will be this week should be here month should be here and year should be here.

In other sense I have a year and quarter then month week day right. So, number of days will constitute this week right 7 days 30 days or 31 days or 28 day 29 days with the month right. So, 3 month is a quarter or I can define a quarter as 12 weeks as a quarter, again 4 quarter is in a year. So, there are two type of hierarchy; one path maybe this way one path maybe this way right. But these are all others way this is only one sort of hierarchy.

So, there can be different type of concept hierarchy over I want to calculate right routine calculation, unless there is some typical situation I want to calculate something else these

are the different hierarchies. So, what we are trying to see here the basic basics of warehousing some of you already knowing for them it is repetition, but basic of hierarchy things are there. Though we are looking on some other aspect like sales etcetera, but it fits in into different of the spatial data sets per se right. So, further spatial database.

So, we will continue our discussion on this spatial warehouse and spatial data mining in our next lecture also, so that we have a overview of that how things work. So, that when we talk about now this data mining or spatial data mining aspects, when the analysis part we are able to know that what this is those are build on this basic fundamental things only right ok. So, we will continue our lecture in our continue our discussion in the next lecture.

Thank you.