Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name: Indian Institute of Technology Madras
Week:03
Lecture:12

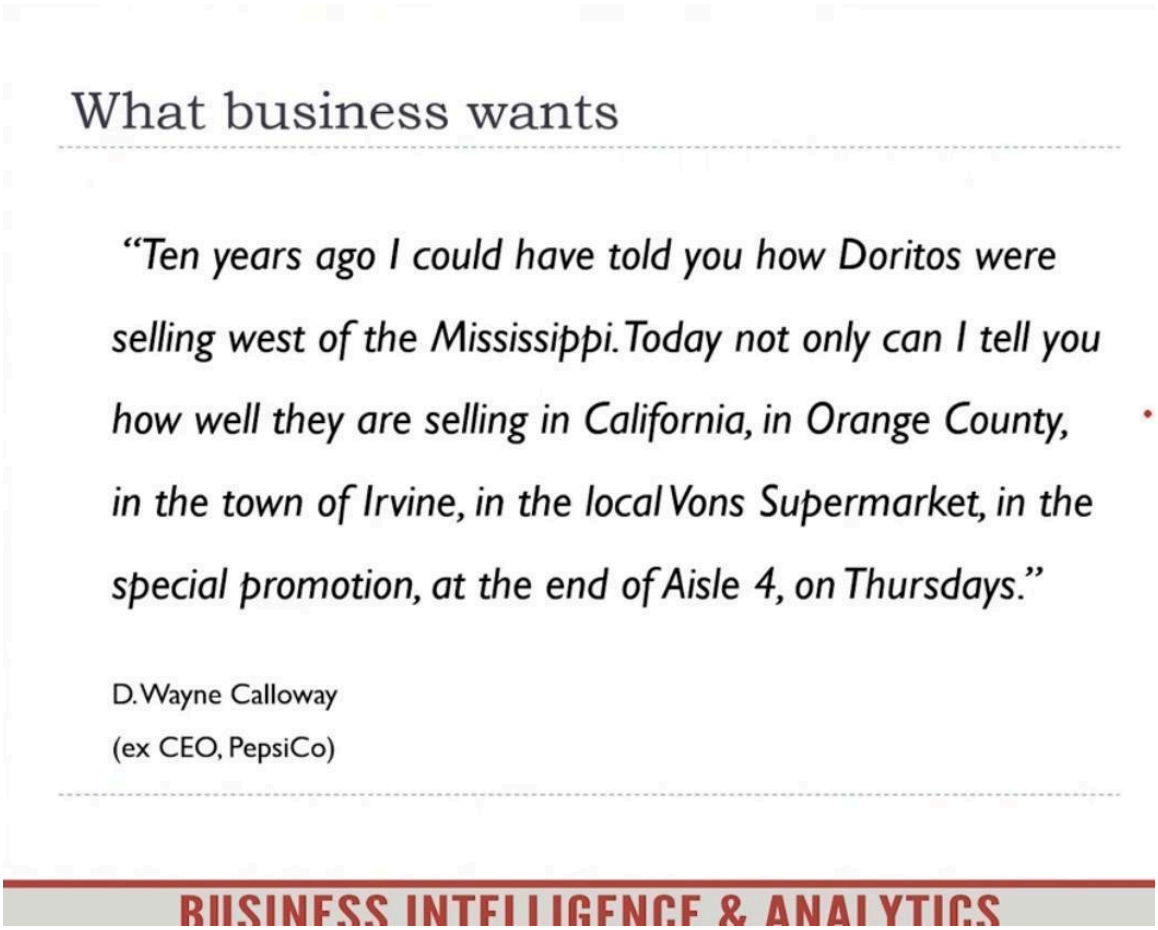
ONLINE ANALYTICAL PROCESSING | Business Intelligence & Analytics

Hello, and welcome back to this session on online analytical processing which is a very powerful technique for descriptive analytics, online analytical processing or in short it is known as OLAP. Technically this can also be called multi-dimensional data analysis or queries. It is a technique or a technology to facilitate multi-dimensional analysis of data. In the previous session, we have seen what is online transaction processing or OLTP. OLTP is a database, it is a raw data.

But in online analytical processing, we move that raw data to a data warehouse. We have seen that what is a data warehouse and then use the data warehouse to analyze data from multiple dimensions. So, there is a technology that ensures that data is captured and pre-processed in a way that when managers or decision makers ask complex questions, the results can be delivered fast. And the results can be delivered without any ambiguity.

So, OLAP data structure or OLAP form of structuring data, which is slightly different from how data is structured in transaction databases, that ensures faster processing of multi-dimensional queries. So, if I summarize why there is a technology called OLAP, it is something like this. If I put that idea into the mouth of Waine Calloway, who was the CEO of PepsiCo, we will hear from the CEO like this. Ten years ago, I could have told you how Doritos were selling west of Mississippi. Today, not only can I tell you how well they are selling in California, in Orange County, in the town of Irvine, in the local Vons Supermarket, in the special promotion at the end of Aisle 4, on Thursdays.

Look at how some of these managers or top executives want to look at the performance of their products. They want to look at the product performance from multiple dimensions like time, location, store, promotion. So, there are multiple views that they want to take of the data. That is one aspect and they also want to look at data from multiple levels. They want to look for the same dimension.



They want to traverse through multiple levels from top down and bottom up. They want to move from country level to store level. They also want to move from a A level to a region level and so on, to take how the product is performing. So, business is all about business performance. A leader or a manager should know how the business is performing on a dynamic basis, not once in a year or once in a month, but constantly.

One has to perform, one has to monitor the performance of business. So, you need certain dashboard. A dashboard, you know, in an aircraft or in an automobile displays how the automobile is running or it provides you running information or current information about certain important characteristics, certain important business, follows certain indicators, how the business is performing. And in short, they are known as KPIs or key performance indicators or KPIs. Business has a need to monitor its key performance indicators.

And only if they know, they can take action, that which you cannot measure, you cannot control. So business to take control actions, they need to know. This knowledge or information provides actionable insights to decision makers and then therefore, they can

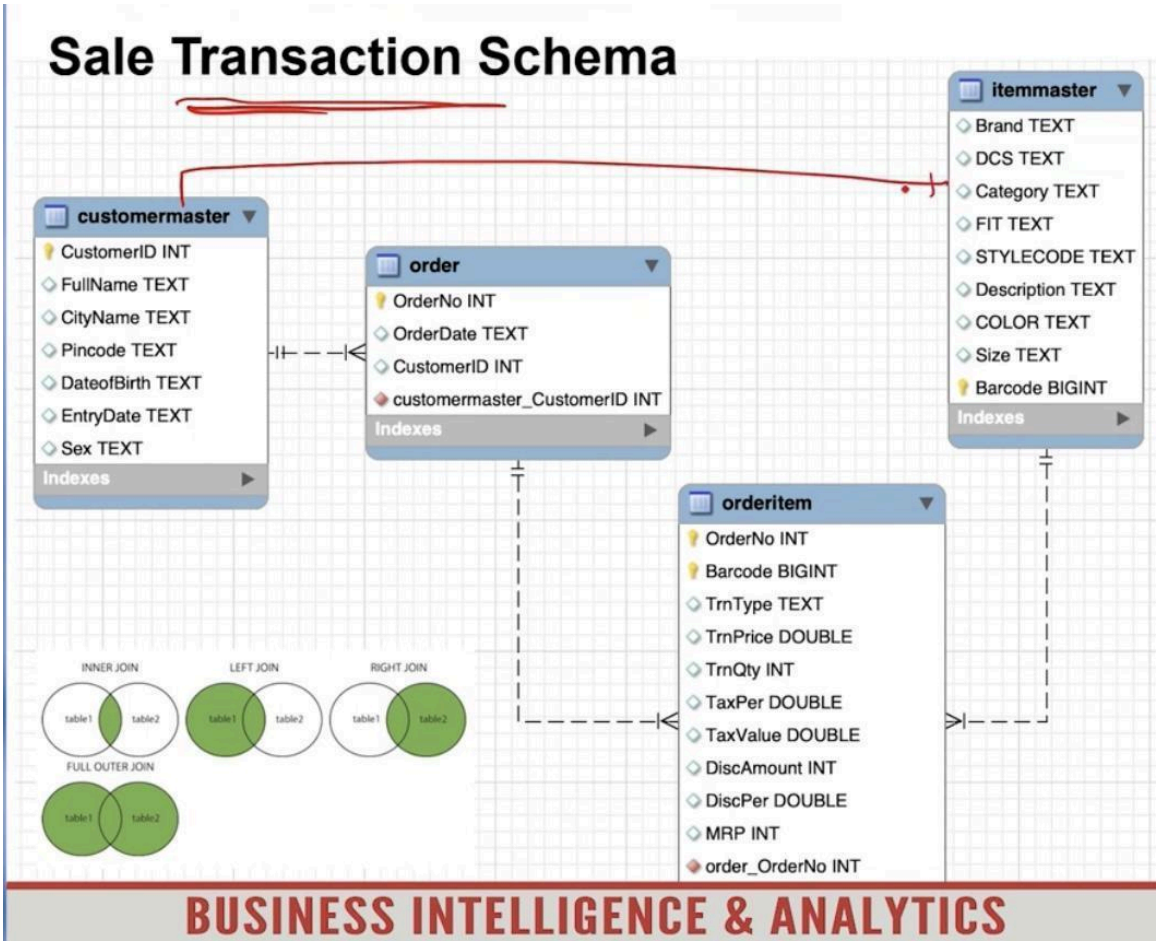
act. So therefore, information is the key for action. And so therefore, an information to be actionable, they have to be relevant or they have to be KPIs.

So therefore, you can imagine, you need a system, you may have a database, data warehouse and so on. But what matters to business from a front end is, inform me about how products and services and people are performing. So, design a system based on data from the backend such that I am able to monitor my KPIs, I am able to browse or sort of look at my KPIs at from different perspectives. So, that is the business need. So, what business wants is this.

So this is, you know, technically if you convert this business problem into an analytics problem, this is descriptive analytics problem. The manager is only asking for a description of data on a continuous basis, but the description is actually multi-dimensional. One has to look at the location of the customer in a hierarchical way. And also have the CEO wants to look at it from promotional perspective and also time perspective. There are multiple dimensions here with respect to which a data has to be retrieved or a query in more technical terms, you have to query a database multi-dimensionally, multi-dimensional query.

That is what the translation of this problem, business problem into analytics problem would mean. So, let us actually, since we have looked at database design at a, you know, at a fundamental level, the entity relationship diagrams. Let us look at how this kind of a query can be facilitated by a schema, a database schema. So, this diagram is a representation of the schema which was used in the particular demo exercise that you listened to or you observed. So in that particular exercise, you have seen that there was a customer master table, there was an order table, there was an item master table and then there is a order item table.

The relation between order and item master is resolved here using order item table. And so it is a schema, it is a transaction schema, basically designed for a transaction database. This particular schema is designed for a transaction database and that is the purpose. This is, in other words not designed for a analytical database. It is not designed for an analytical process.It is designed for transaction process and therefore, when you try write queries based on this table for multiple dimension or multi-dimensional queries, you will find that you have a difficulty, a real difficulty in writing table, writing queries.



Suppose you have a query where you have to retrieve data from the order item table by joining all the tables, you have a difficulty. So, you can see that customer master table is related to order table. So therefore, you can join this, you can give a join command because there is a common attribute between these two. Similarly, order, order item, item master, order item they are related and therefore, there are common attributes based on which you can join these tables. But when you look at a customer master and item master and suppose you have a query where item master is involved and customer master is also involved and you have to join them, you have a difficulty there because there is no common attribute between an item master and a customer master.

And item master carries the characteristics of the items and customer master carries the characteristics of a customer and you cannot join them because there is no common attribute. So, that becomes a difficult situation for writing a simple query. So, the difficulty here is basically that the tables are not designed or the schema is not designed for facilitating queries or multi-dimensional queries. That is not what someone kept in

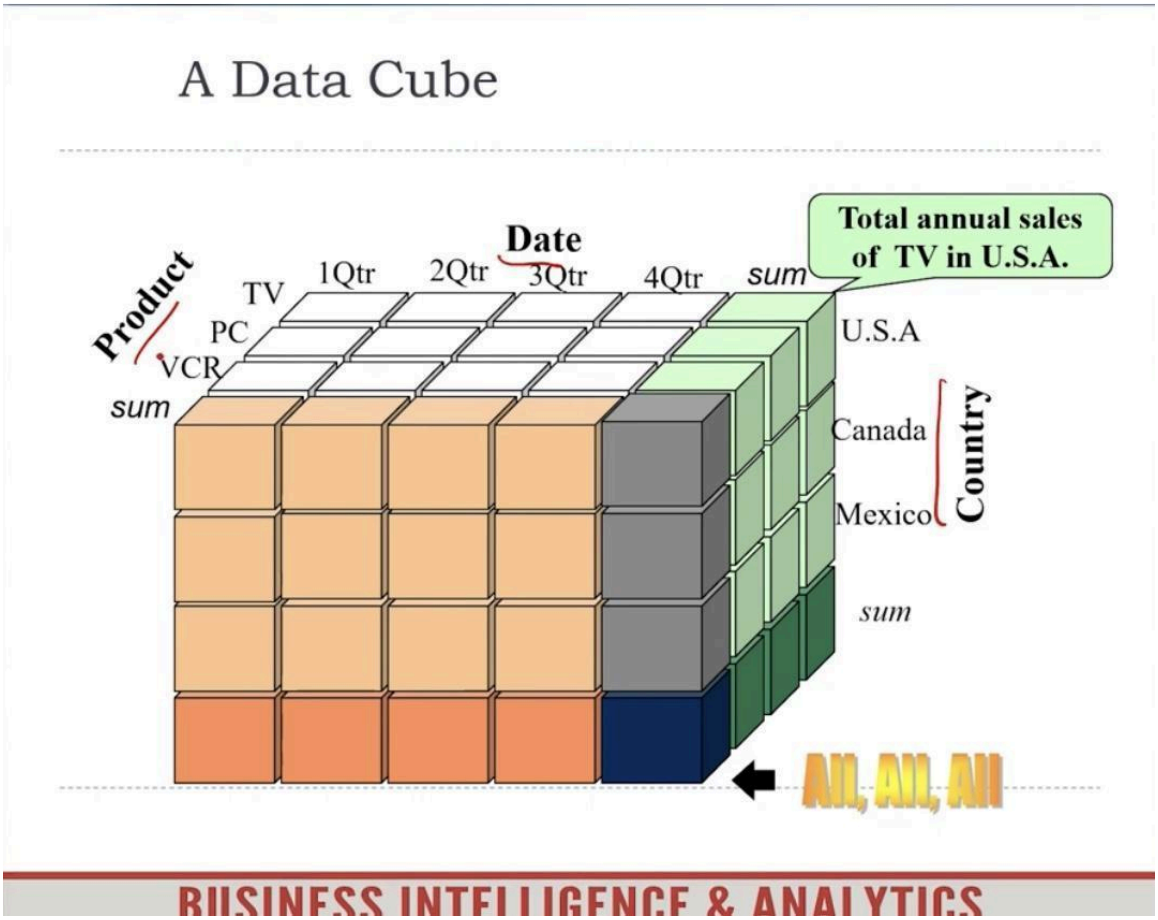
mind while setting up a table structure like this. The designer had the transactions in mind. So therefore, this structure is not good for multi-dimensional queries because it is a transaction schema.

So, what would be better? Look at how for, only for the purpose of understanding what a different structure could have been to facilitate multi-dimensional queries. Instead of the transactional schema, now I am presenting you a different schema wherein you have something called a sales table, which I can call a fact table, which is a central table and every other table like the customer item order, all are related to the sales table through a 1 to n relationship, meaning the sales table has the primary keys of all the other tables. You see the primary key of the sales table is a concatenation of the customer table key, the item master key and the order table key. In doing so, in order to conduct any query, you can actually query for any data, you can see the content of the sales table is actually certain measures. We call them as measures.

Measures like the discount, the tax value, the sales volume or the MRP, all that is of interest to a decision maker. So in order to aggregate, in order to count, in order to sum, all that actually can be done on the measures that is stored in the sales table or the fact table, it can be called technically, the fact table has the measures and the fact table connects with the other tables. The other tables here in this sort of a design would be known as dimension tables. Dimension tables are tables, which are the basis for analyzing the data. Data is stored in the fact table or the indicators or the KPIs are stored in the central table, but they have to be analyzed with respect to certain important dimensions.

Who decides this? The decision about what is the content of a fact table and what is the content of a dimension table. It is Wayne Calloway's decision or Wayne Calloway's requirement or the decision maker's requirement? I have these questions. You design a system such that I can analyze or I can explore relevant KPIs multi-dimensionally and hierarchically. And the answer to that question is a design, which is having a cube structure. This is known as a cube structure of tables.

The schema is having a cube structure and why it is called a cube structure? Because it can be visualized as a cube, although it is not physically a cube, it can be related to or visualized and explained using a cube structure. This particular slide illustrates what is a cube structure here. So, look at this cube, you know, it is, it looks like a rubik's cube. And the idea is to convey what is a cube structure in OLAP, online analytical processing, and how it facilitates multi-dimensional queries. So, in this particular cube, there are only three dimensions, the product dimension, the date or the time dimension and the country dimension or the customer or location dimension more generically.



And you can see that product has three values or there are three products that the company has, a VCR, PC and TV, I would call them the three values of a categorical variable called product. This is another way of technically describing it or product has three values. So, product dimension has three values, date dimension has three values, first quarter, second quarter, third quarter and fourth quarter, very simple. And country dimension has three values, USA, Canada and Mexico. So, this is a simple scenario to understand how a cube structure works.

Now, what does this mean? If I just randomly look at one cell of the cube structure, what does it actually represent? This particular cell represent PC sales in third quarter in USA, PC sales in third quarter in USA. This has all the three dimensions involved, all the three dimension, with respect to three dimensions, somebody asked, what was the sales of PC in USA and the third quarter? It is, it can be pre computed through a query and stored there, in a different layer of computing called OLAP layer. Or you know, there can be the query ready to run, it is not pre computed, but the query can be easily run multi dimensionally. So, that depends on how an OLAP is implemented. But you

can see that is, this structure enables us to visualize how an OLAP works.

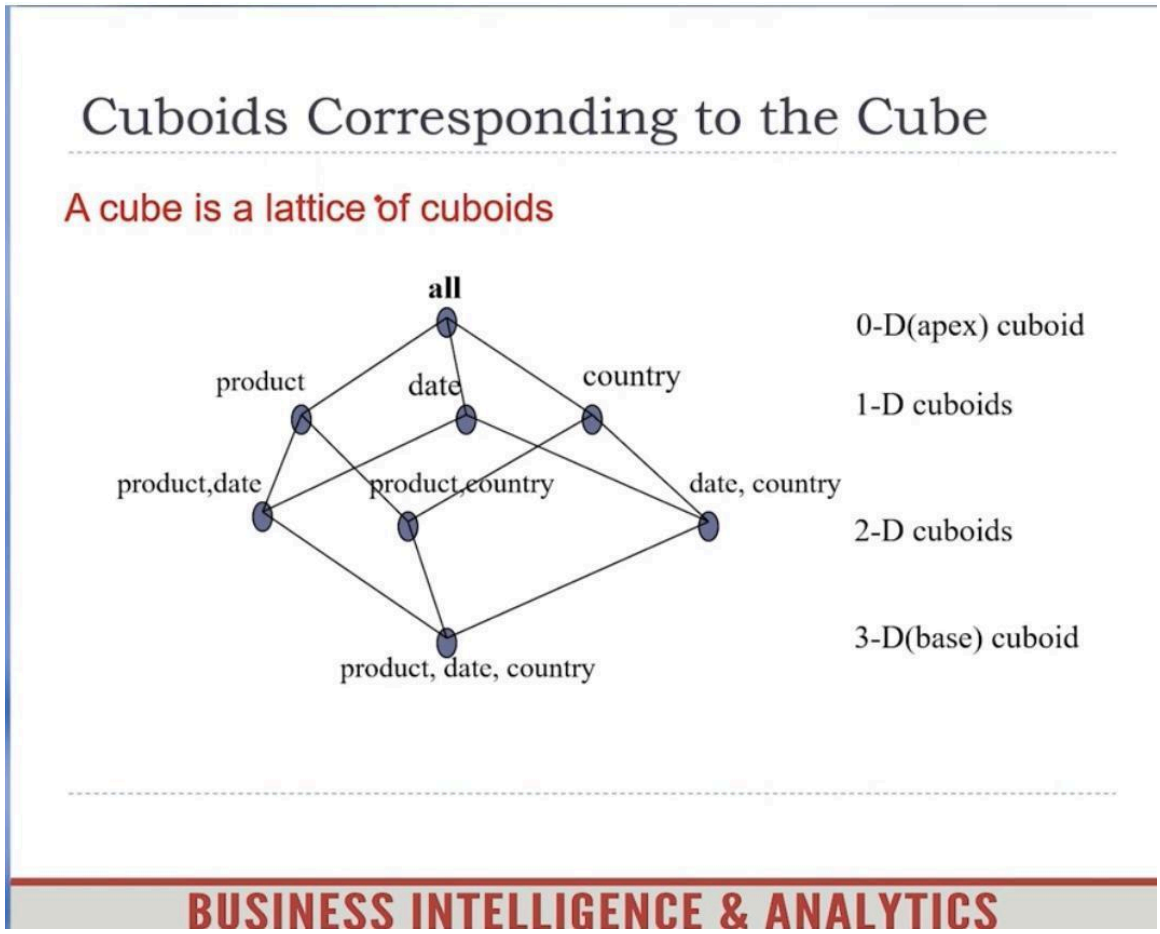
So, it facilitates multi dimensional query. But suppose somebody is interested instead of looking at a one singular piece of data, one wants to see how various products are performing. One is indifferent about the time and the country, but how the products are performing overall. That is an interesting scenario. And for that, you have to go back to your lesson on queries and you learned an interesting query instruction called group by, group by, the group by command.

So suppose, there is this particular key called product key, so group by product. If I say group by product, and I do not add any other dimension here, I do not add group by date or group by country, I just say group by product. So what I want is, what is the TV sales, total TV sales? What is the total PC sales? What is the total VCR sales? Where do I get, where do I go to get this sum? This particular representation of the cube actually helps you spot it. You see this particular cells, these are summing cells. These particular three cells have summed VCR, PC and TV sales for all quarters and for all countries.

This is a group by product command. And you can see, you will be summing up say, sales, whatever, followed. The sum sales of course will come first. Select sum sales and then you would say group by product, approximately I am giving a skeleton code,a skeleton sql code that will implement this and you will get these values here. Similarly, you can think of a group by date and group by country. Now looking at this structure, you should also imagine, suppose somebody wants product sales by date or by time, then one is indifferent to country, total sales, product wise and date wise.

You can see that you will add one more dimension here, group by product, date. So that will actually provide sum of sales for all countries, but product wise and date wise. Now you have to spot the cells where that information will be present. You can easily imagine that that gets inside here. You have to traverse inside, it is not visible outside.

But this cube enables you to imagine how multi-dimensional queries in OLAP system is facilitated by, is particular design of tables called the cubic structure of tables. And then using certain features of the, certain characteristics of the SQL like the group by command, data can be aggregated for responding to multi-dimensional queries and that is what a OLAP is. A data cube is the particular schema structure of OLAP. Now this particular slide represent what I explained to you using a physical cube structure.



So a cube is a lattice of cuboids. Look at that statement. That is a very interesting, I would say a more metaphorical statement. It uses cube, lattice and cuboids which are terms that belong to a physical cube. But here in OLAP, you actually can relate with that particular physical artefact and try to understand how an OLAP cube works. What is a cuboid? A cuboid is particular vertex.

This is a cuboid, this is a cuboid, this is a cuboid. So there are different types of cuboids. And these cuboids have information. And what kind of information? They have summarized or aggregated information. So this aggregation happens at different levels.

Look at a zero-dimensional cuboid. Zero-dimensional cuboid is not having any dimension. It is like answering the question, what is the total sales? Does not matter product, does not matter which country, does not matter which time. Some of the wholesale, how much money business has made. That is what the zero-dimensional cube is. Well, at the next level, somebody is more inquisitive.

Get me product wise sales. So there is the, a lattice known as group by product. There is a lattice, which is, there is a cuboid, there is a cuboid, group by product, group by date, group by country. Then someone becomes more inquisitive, group by product, group by date, group by product, group by country, group by date, group by country. So this is actually two-dimensional query. It becomes three-dimensional cuboid when you make it group by product, group by date and group by country.

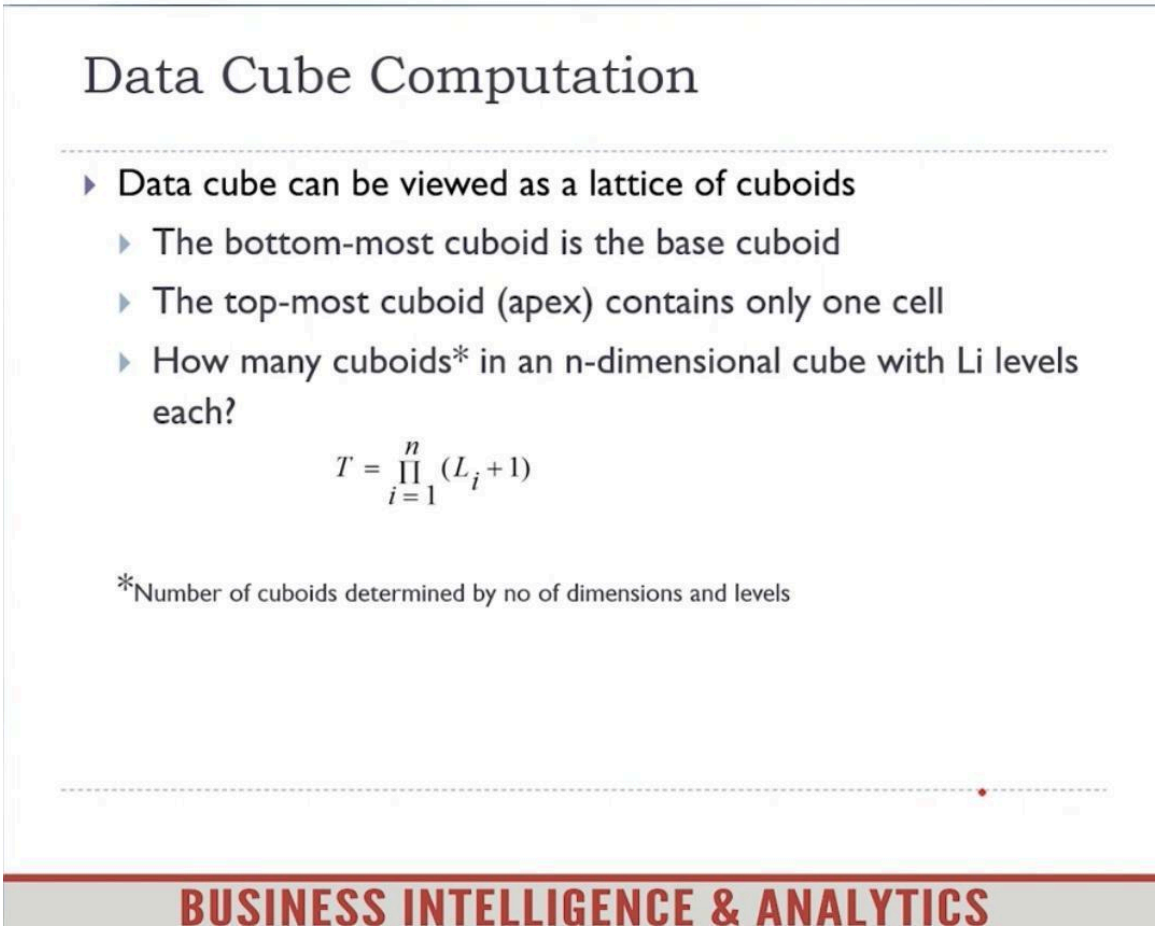
That is the most detailed information that you get in a format in response to SQL query. So you can see that the heart of the matter is that the cube structure is a particular data structure in OLAP system, which enables computation of data from different tables into certain, into a certain centralized table called the fact table, where you can have the choice of pre-computing that data and storing it, or you can run this queries at the time of, at the time of the actual query when somebody queries for it. You can pre-compute, you can compute at the runtime. That depends on how much storage you have, etc. So but these are the different requirements that business has and you develop a system, develop a structure for the data such that this multi-dimensional queries can be responded to efficiently and that is what an OLAP system is.

So this is a three-dimensional cube and that is what you expect a cube to have. But in OLAP, the dimensions are not restricted to three. There can be more than three dimension. And as I said, here the cube is just a metaphor to visualize how the system works.

It is not a physical cube. Here you can see the structure or the representation of a four-dimensional OLAP cube, four-dimensional OLAP cube. And there is always the apex cube, which does not have a, which does not have a particular dimension, any dimension. It is neutral. But the number of dimension, the maximum number of dimension is four here because there are four dimension tables here. So this slide is a summary of the information that I already gave you.

There are, a cube is a multi-dimensional structure for organizing data for the purpose of responding to complex queries, multi-dimensional queries. So at this point, you also need to appreciate that aggregation of data is done through SQL queries, but when you aggregate and store that data, then there is a demand for storage space and which is also a function of the number of dimension and the number of levels, number of dimensions and number of levels. Number of levels is represented by l and number of dimensions is

represented by n. So in order to answer, answer how many cuboids in an n-dimensional cube with Li, levels that can be computed using this formula. And the, why it is important? The number of cuboids will determine the extent of computation required and the extent of storage required, if you go for full materialization or full pre-computation of all the cuboids.



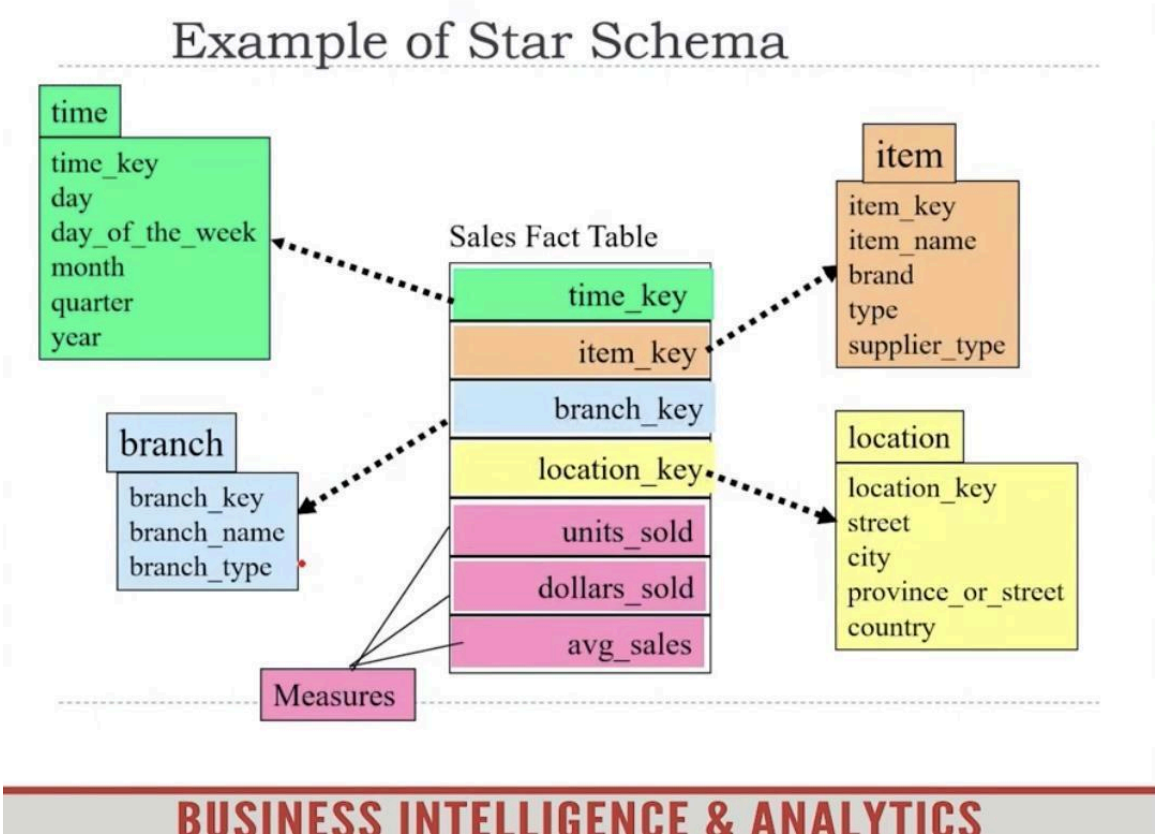
All the cuboids can be pre-computed and that will be called full materialization. Part of the cuboids can be pre-computed if there is frequent queries about certain multi-dimensional queries, you can actually have partial materialization or no materialization. So this is related to storage requirement. But full materialization will be very efficient query because data is already pre-computed and when you query it, it is there in the dashboard, it is in the sense that it is constantly updated. And this formula is useful in sort of determining the number of cuboids and therefore, the load on materialization.

So for example, if there are three dimensions D1, D2 and D3 and say, it is say product,

time and say location. So product, there are three levels of product. And suppose time there are, say you can imagine year, let me give an example here, year, month, and date, these are the three dimensions. So this is again three dimensions. And I already imagined that product has three dimensions, this is also three with an example with a hierarchy here.

This is a conceptual hierarchy in a particular dimension, year, month, date. Also indicating that this kind of hierarchies is something that you can design based on user requirement. And suppose the third dimension of location is say country, state and province. So and this is also, so let me also add a particular city.

So therefore, suppose this has four dimensions. So it is 3, 4, how, what would be the total number of cuboids? That is the question. The total number of cuboids, this is a product. So therefore, the first dimension is d1, which has 3, so 3 plus 1, the 1 is added to account for the apex cuboid, so which is not having any dimension. So 3 plus 1, and 3 plus 1 is the next one into 4 plus 1.



What is the total number of cuboids or 4 into 4 into 5? 80. Correct. So that is the answer. So you can work this out. And you can imagine how the computational cost as well as the storage cost increases as a function of number of dimensions and number of levels in each dimension. Now, in order to implement a data cube, there are different configurations or different designs possible.

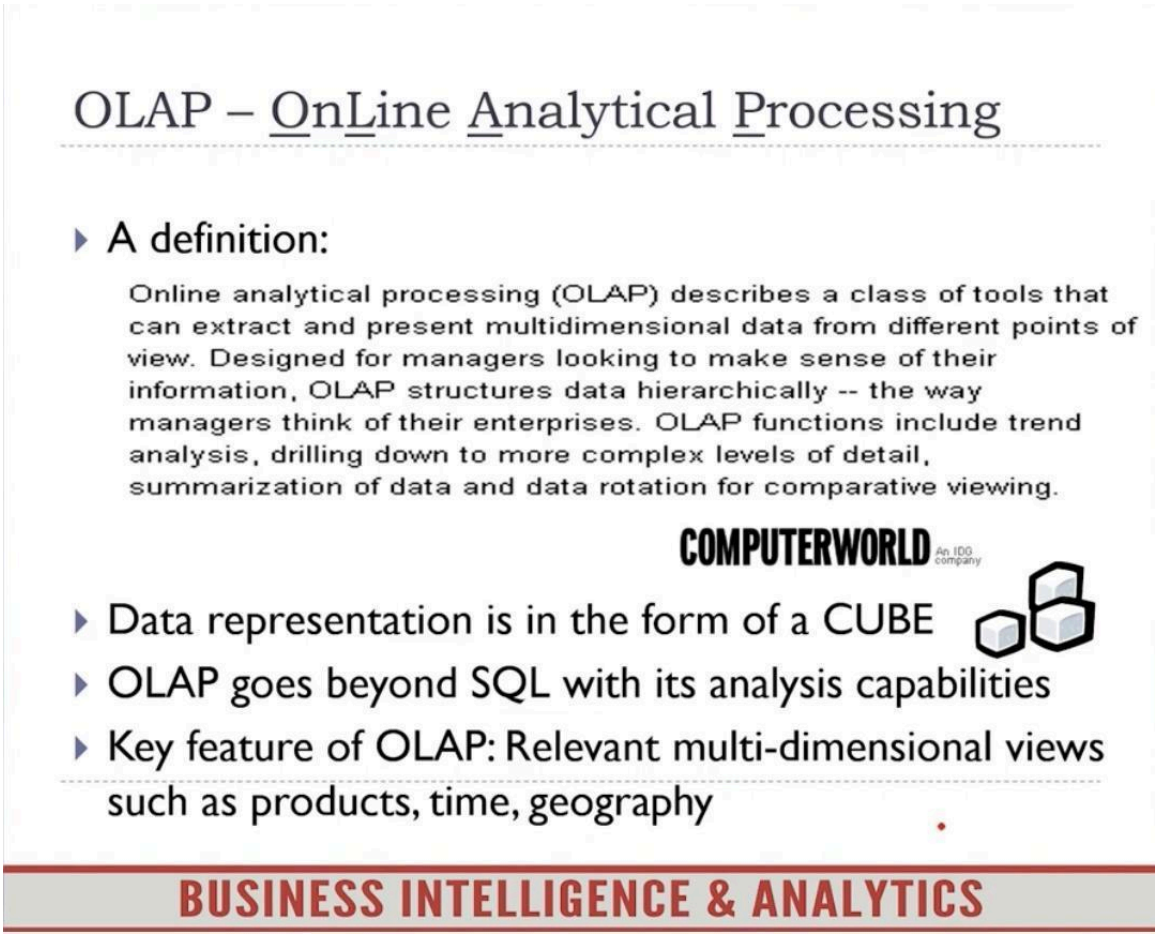
A schema or a data cube schema can be designed as a star schema, as a snowflake schema or as a fact constellation. What does a star schema mean? It is easy to imagine now because we have related it with the, with the previous table structure, where I showed you how a table consists, a table structure of a cube consists of a fact table and dimension table. Here you can see that there is a central fact table here. And then there is dimension 1, dimension 2, dimension 3 and dimension 4. The dimension tables all connected through the primary keys you can see to the central fact table.

That is a simple structure or a star schema. But what is a snowflake schema? When you come to the snowflake schema, you can see that the dimension, the fact table is the same, but the dimension tables are normalized. The dimension tables are normalized or they are split. Because you know, you know how normalization is done, you have already seen that. So some tables are normalized here. So a snowflake of course, when you normalize the other tables hang, like the snowflake.

So that may be the reason why it is called snowflake. So a snowflake schema, the dimension tables are normalized. And as soon as you hear the term normalization, you know the purpose. When you normalize, they are more efficient for storage or they reduce redundancy. So they are efficient in storage. But what is the trade off? But you trade off, trade it off with query, less efficient query.

Queries are not efficient, that efficient because you have to join multiple tables here and so on. So therefore, a snowflake schema is good to save space, whereas a star schema is more efficient for queries. That is a simple and direct implication. And so therefore, two designs are available.

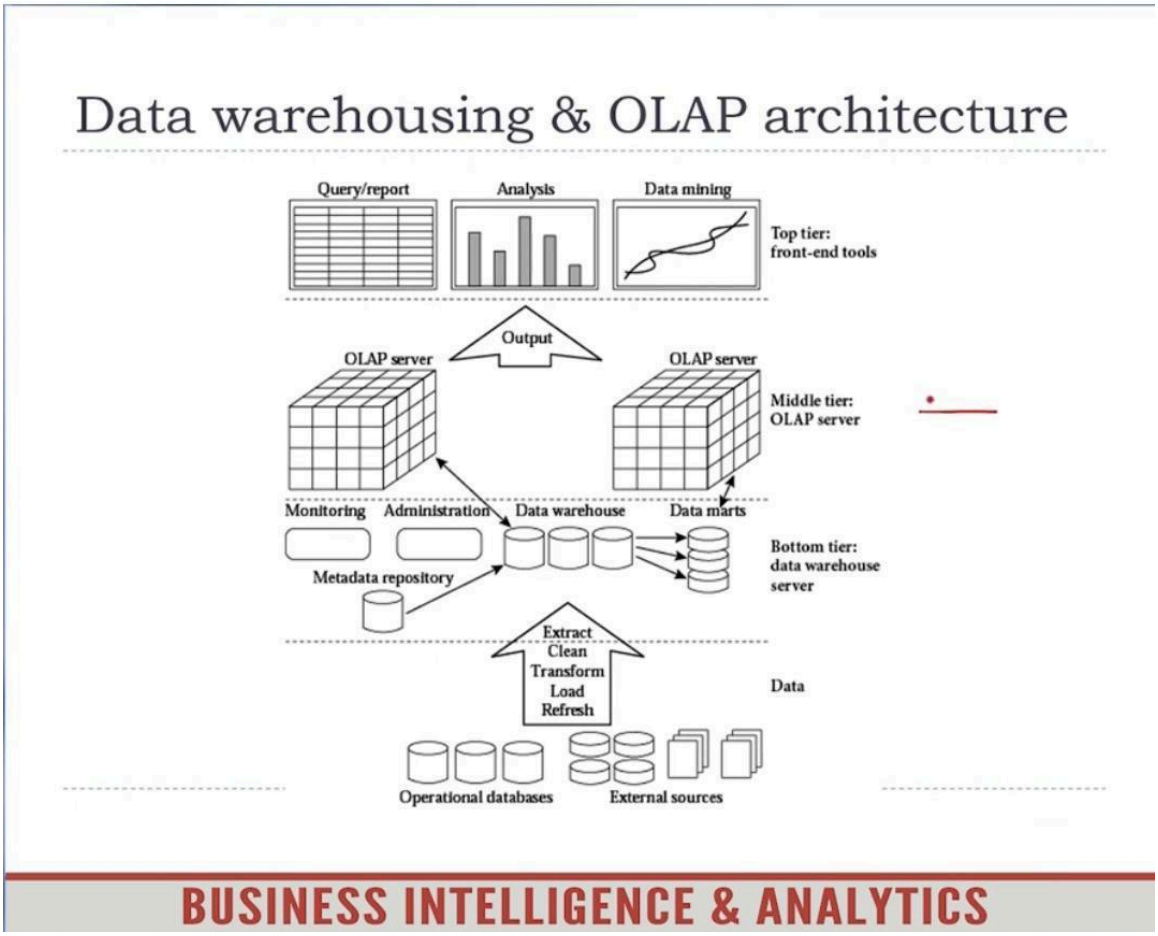
And there is also fact constellation that is shown in the previous slide. Fact constellation is about number of fact tables you have. In the simple design, you see there is only one fact table, but in fact constellation, where in real life scenarios, you may have multiple fact tables, multiple fact tables. So when a cube design has more than one fact table, it is called a fact constellation. And this slide actually tells you about the definition, the practical definition of what an OLAP is.



It is a data representation in the form of a cube. It goes beyond a simple SQL query and so on. So let us move on. And this slide provides you a representation of the architecture of which involves OLAP. You can see that a layer for an OLAP server is added here.

And this you can see as the, you know, we talked about query reporting, etc. in the previous BI architecture. Essentially it is the same, but stressing on the data visualization or the, you know, the visualization of data is stressed here in a business intelligence context. When you add an OLAP server, it provides for multi-dimensional views, multi-dimensional queries and multi-dimensional visualization, say in the form of a dashboard or things like that which enables business users to understand KPIs in a more effective way. So you can see that it starts with databases, it moves on to data warehouse. So till that point, it is the same as the BI architecture we see, we have already seen, but it adds an OLAP element to it in a data warehouse working with an OLAP server either to

pre-compute all the multi-dimensional queries and then deliver it to the front end or partially materialize as we have already seen.



The purpose of the OLAP server is to feed the front end constantly and the OLAP is designed or the OLAP structure is designed based on the business requirements or the query requirements that comes from the business users. We have seen this already. Before I close, let me explain to you some of the fundamental terms that are associated with slicing, that is associated with slicing. A slice, what is a slice? A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimension or in other terms, you see it is shown here, what is a slice of a multi-dimensional data space.

It is actually a slicing with respect to, you can say, it is like a group by location. If you do a group by, you actually fix one dimension and then explore the data, that is a slice of the data. So for Asia alone, the rest of the data is shown, that is a slice of the data. So slicing is to fix one dimension and then explore the data, whereas dicing is, it is not one

dimension alone, it can be multiple values of the different dimension that is explored. It is a subset of the data, but it cut across multiple dimensions and then it becomes dicing of data.

Drill down and, drill down operation is about top down exploration of data. As we have seen, the top down can be done multi-dimensionally or with respect to a few dimension, but essentially you traverse down from an upper level. Like if I use the time hierarchy, then say I drill down from year to month and a date, this is drill down, this is drill down operation. There is also roll up, which is just the opposite of drill down. It is bottom up, not difficult to understand.

And then, there is of course the pivoting of data, which can be very intuitive. You do not change any data, but you look at, you change the view, you change the order in which the dimensions are placed. Instead of product, location, time visualization, you use product, time, location and it entirely changes the view or the perspective. And a person who visualizes the data is able to imagine how the KPIs are performing with respect to different dimensions when you change the sequence. It actually enables your mind to visualize the data differently.

All right, so I will close here. This is nothing but Gartner Magic or Gartner's Magic, a market research firm Gartner publishes how different commercial providers of the BI solutions are and how they are performing etc. And this is of course a commercial, this is a market research firm and this is about commercial products and those who have interest in this kind of understanding of the market of BI, can look for Gartner's Magic for, you know, this is not updated, this is 2019. Thank you very much and look forward to the next session, which will be on analytics process or data mining process, how we move from descriptive analysis to explanatory and predictive analytics and what process you follow. And see you then. Thank you very much.