Unit 6.2

DATA WAREHOUSING

Data Warehousing Design Consideration and Dimensional Modeling: *Defining Dimensional Model, Granularity of Facts, Additivity of Facts, Functional dependency of the Data, Helper Tables, Implementation manyto-many relationships between fact and dimensional modelling*

- The promise of Business Intelligence (BI) is theability to make key strategic decisions quickly, intuitively and easily, from a uniform view of the core data of the organization in a timely manner and from a single nexus of control.
- One of the common methods of pulling together the corporate data for ease, speed and flexibility of analysis, is through OLAP cubes: multidimensional methods of storing a pre-aggregated version of the corporation's data.

Defining OLAP

- On Line Analytical Processing (OLAP) is now viewed asa key technology for providing knowledge workers with access to business data in a meaningful, intuitive fashion.
- OLAP was first coined by Dr E F Codd, the inventor of the relational model, to describe a genre of software that is used for analyzing business data in a top down hierarchical fashion.
- relational databases were never intended to provide the very powerful functions for data synthesis, analysis and consolidation that are being defined asmultidimensional analysis

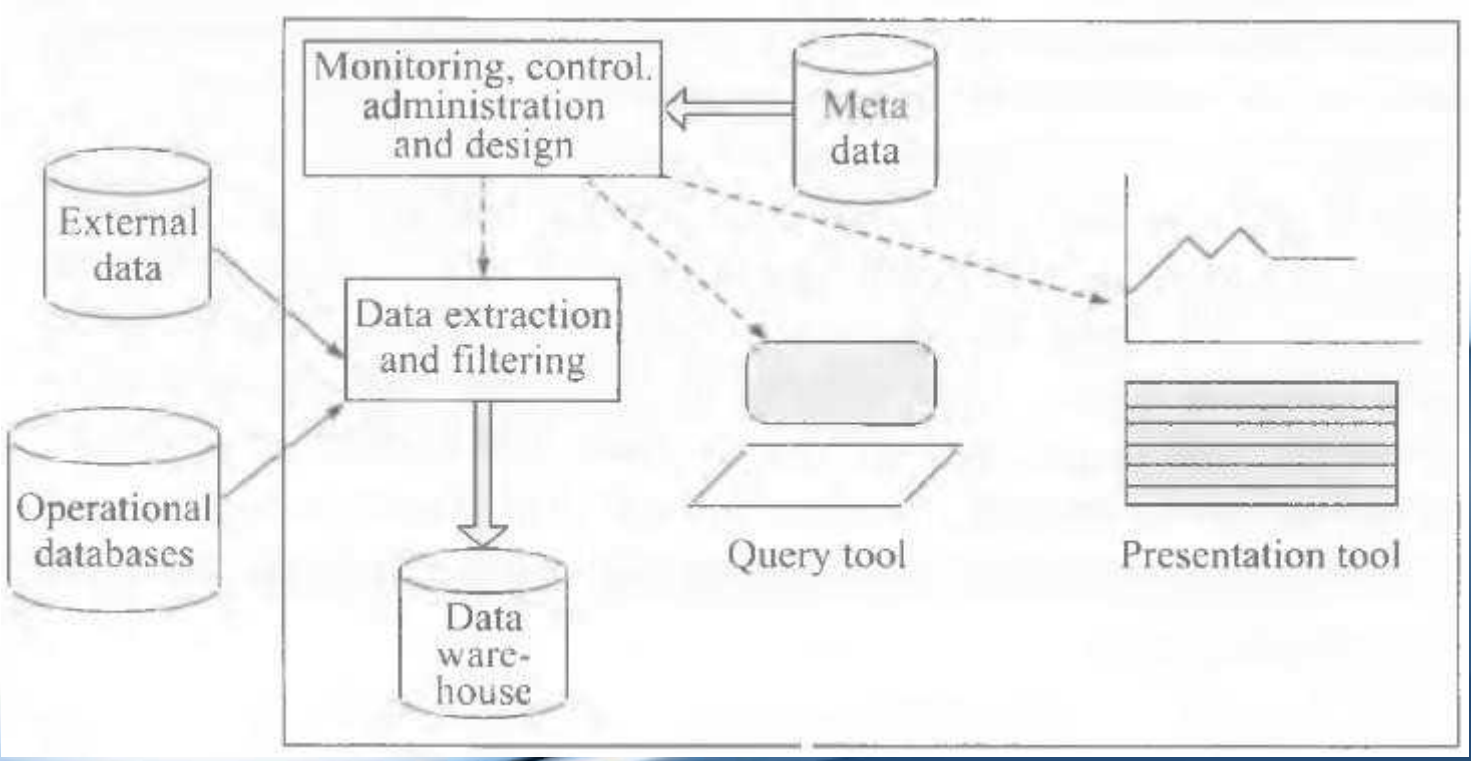
The Value Of Multidimensional Data

- Consider the example of the VP of Sales, of a retailing company. He wants to analyze how products are selling across his retail outlets.
- The dimensions are accounts (often also called variables or measures),products, time, channel, region and salesperson.
- One of the key features of OLAP technology is that the user is able to navigate through the data in any way that makes sense, without knowing in advance what the navigation route might be.

OLAP Terminologies

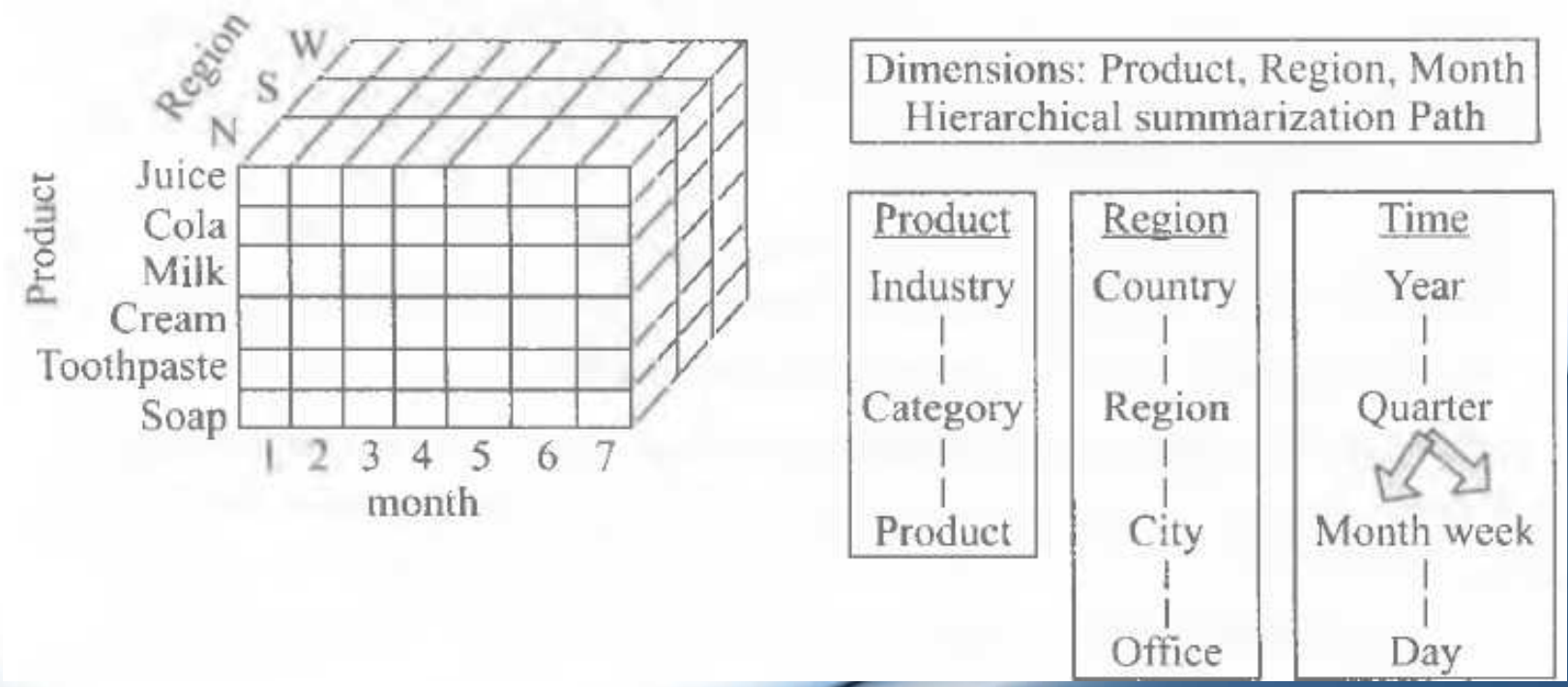
- There are three main architectures for delivering fast, dynamic, sophisticated analysis of multidimensional data.
- Relational OLAP (ROLAP), as its name suggests, uses a standard relational database to store the physical data.

Basic OLAP System Architecture



- Physical multidimensional databases (MOLAP) use a storage mechanism which is optimized for the pre-calculation, storage, and retrieval of multidimensional data.
  - MOLAPs are best suited for medium sized, static, and mostly static applications, which demand nothing less than sub-second data retrieval.
  - Used for analysis of historical sales and financial information
- Real-time Analytical Processing (RAP) deals with all the multidimensional input values in memory and creates the derived multidimensional values in real time, on demand.
  - RAP is best suited for dynamicapplications
  - The ability to perform calculations in real time (eliminating batch processing delays) makes RAP the best choice for dynamic applications such as financial reporting, budgeting, and planning
- and management in marketing, operations and sales.
- ROLAPs are best suited for large, transaction intensive applications such as high volume retail sales analysis.
  - Their advantages are the ability to handle extremely large data sets and having the same technology as existing RDBMS based systems

OLAP Multidimensional Cube



- A hypercube (implying a cube with many dimensions) refers to a collection of multidimensional data.

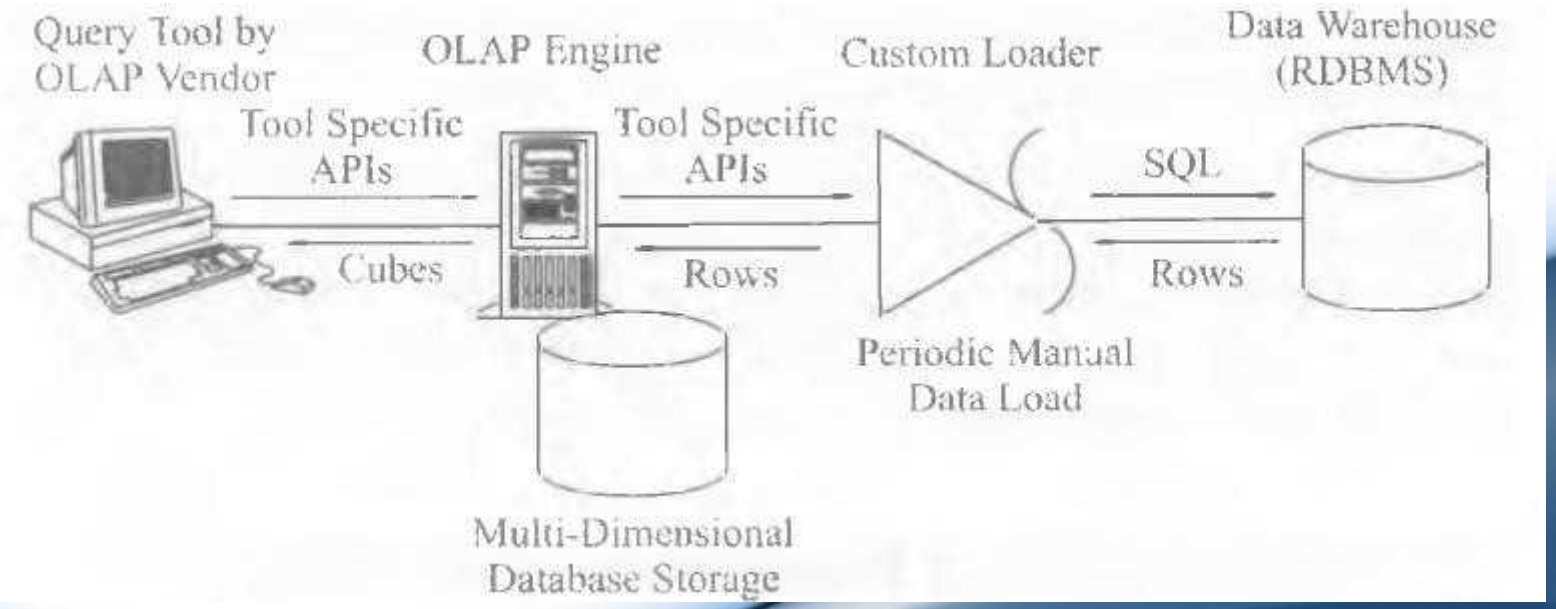
Understanding Multidimensional Data

- It is first necessary to understand the nature of multidimensional data.
- Multidimensional data is almost never 100 per cent dense.

Multidimensional Architectures

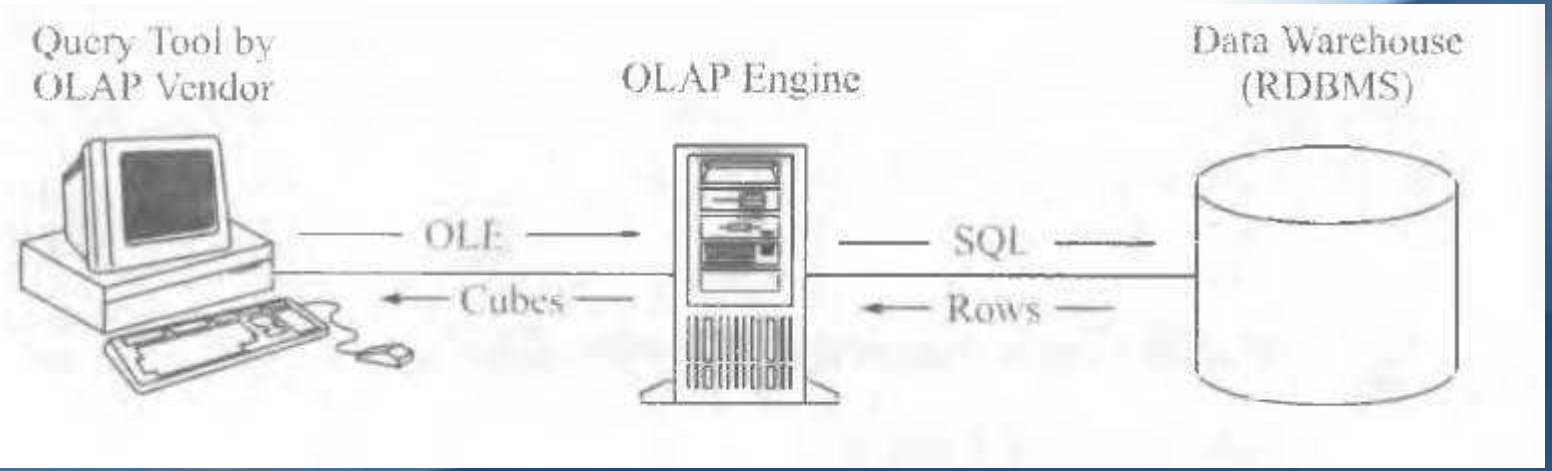
- One of the design objectives of the multidimensional server is to provide fast, linear access to the data regardless of the way the data is being requested.
- The simplest request is a two dimensional slice of data from the n-dimensional hypercube.
- The objective is to retrieve the data equally fast, regardless of the requested dimensions. In practice, such simple slices are rare; more typically, the requested data is a compound slice where two or more dimensions are nested as rows or columns.

Multidimensional Database Architecture



Multidimensional Views of Relational Data

- They provide a multidimensional view of this data. For this, all of the relational OLAP vendors store the data in a special way known as a star or snowflake schema.
- The data is retrieved from the relational database into the client tool by SQLqueries.



Physical Multidimensional Databases

- The next two major OLAP architectures, MOLAPand RAP, provide their own physical multidimensional databases.

Data Explosion

- It is not immediately obvious that a fully calculated hypercube is usually dozens of times, and in some cases many thousands of times, larger than theraw input data.

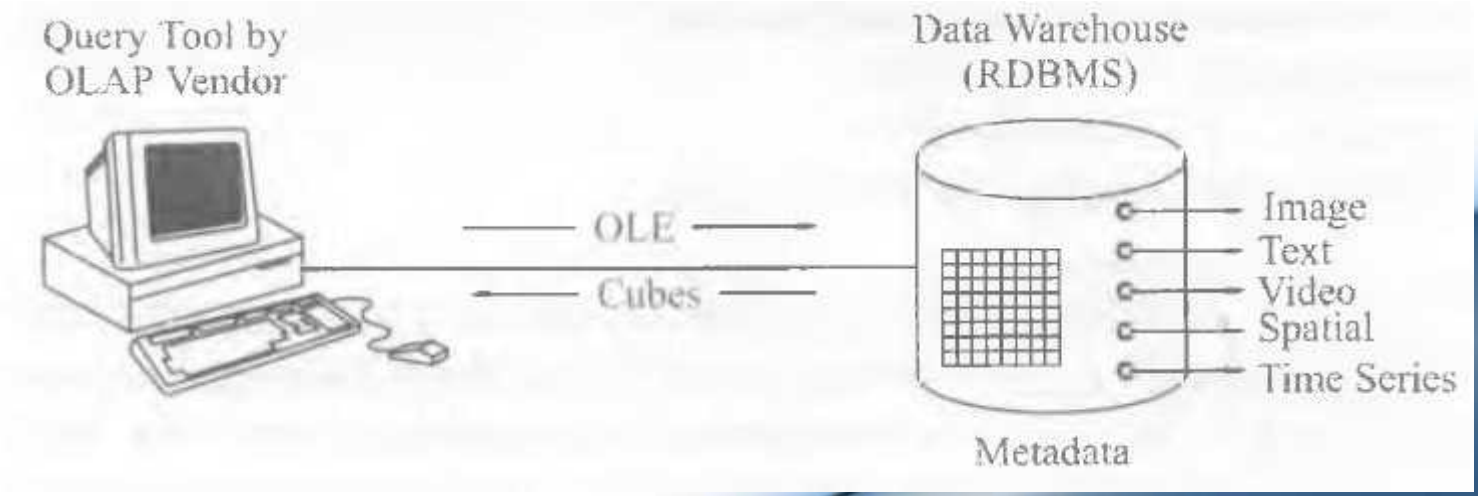
Physical Multidimensional Databases (MOLAP)

- The MOLAP approach is to provide a database specifically designed for multidimensional dataand then pre-calculate all derived values. The theory behind this is that pre-calculating all derivedvalues will result in very fast retrieval times, and that data explosion does not matter since disk space is cheap, relatively speaking

Real-Time Analytical Processing (RAP)

- RAP takes the approach that derived values should be calculated on demand, not pre-calculated. This avoids both the long calculation time and the data explosion that occur with the pre-calculation approach used by most OLAP vendors.

Integrated Relational OLAP



Problem in traditional ROLAP system

- data travels over the networktwice:
  1. once to the analysisserver
  2. and again to the client application.
- The next logical step in the evolution of OLAP technology is the integration of the ROLAP engine with the scalable, parallel RDBMS itself: integrated relational OLAP
- With an integrated database server and cost-based optimizer for both relational and multidimensional analysis, the “universal” server can giveunprecedented performance and scalability for the data warehouse.

Some of the critical features of integrated ROLAP functionality are

- **Parallelization**
  - One of the key problems with data warehouses is the sheer volume of data.
  - Parallelization techniques are critical to performance
  - breaking down complex actions into smaller parts, each of which can be executed in parallel.
  - The net result is faster execution.

- Data Partitioning
  - Partitioning enables the database to automatically distribute portions of a table or tables into more than one logical and/or physical file, which enhances the ability for the database to parallelize operations and eases maintenance on the large data sets.
- **DSS indexes**
  - An index is a means by which an RDBMS can very quickly access information from a table (or tables) without scanning the entiretable.
- **Sampling**
  - sampling allows users to estimate results based on a partial representation of detailed data.

Data Sparsity And Data Explosion

- The reality of data explosion in multidimensional databases is a surprising and widely misunderstood phenomenon.
- in a one-dimensional matrix, you can suppress all zero values and, therefore, have a 100 percent dense matrix.
- In a two dimensional matrix, you cannot suppress zeros if there is a non-zero value in any element in the two dimensions

100 Per cent Dense Structure

	Year
Product A	10
Product B	20
Product C	8
Product D	15

25 percent Dense

	Q1	Q2	Q3	Q4
Product A	10	0	0	0
Product B	0	20	0	0
Product C	0	0	8	0
Product D	0	0	0	15

100 Per cent Dense Structure—No Data Explosion

	Year
Product A	10
Product B	20
Product C	8
Product D	15
Total	53

Data Explosion is 3.75 Times

	Year	Q1	Q2	Q3	Q4
Product A	10	10	0	0	0
Product B	20	0	20	0	0
Product C	8	0	0	8	0
Product D	15	0	0	0	15
Product A + Product B	30	10	20	0	0
Product C + Product D	23	0	0	8	15
Product A + Product B + Product C + Product D	53	10	20	8	15

Thank You