Data Warehousing - II

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

School Of Computer Engineering

Dr. Amiya Ranjan Panda
Assistant Professor [II]
School of Computer Engineering,
Kalinga Institute of Industrial Technology (KIIT),
Deemed to be University,Odisha

3 Credit

Lecture Note 07

Acknoledgement

A Special

Thanks to

J. Han and M. Kamber.

&

Tan, Steinbach, Kumar

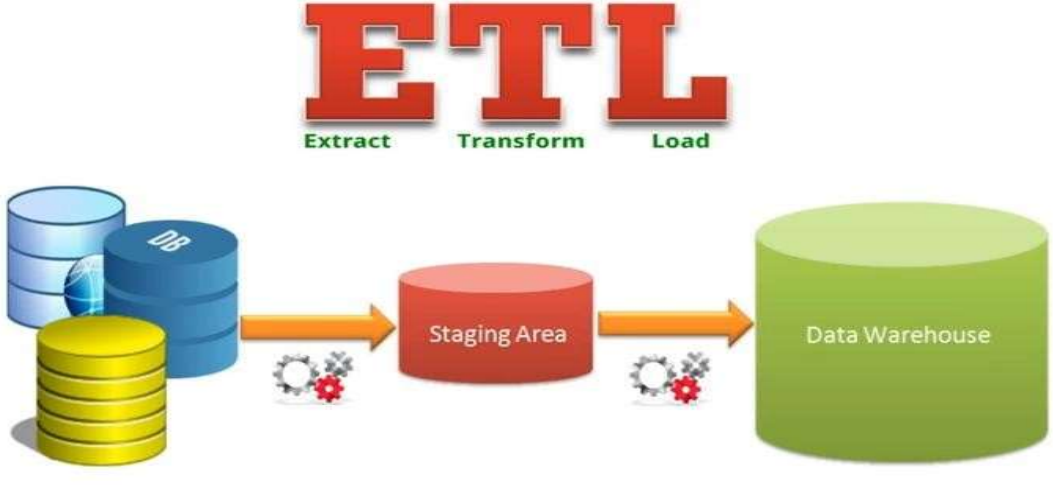
for their slides and books, which I have used for preparation of these slides.

Chapter Contents

- ETL (Extract, Transformation, Load)
- ETL Tools
- Metadata (data catalog)
- Schema Design
  - Star Schema
  - Snowflake Schema
  - Fact constellation schema
- Dimension Table
- Fact Table
- OLAP Cube
- Operations on Datacube
  - ✓ Drill Down
  - ✓ Roll Up
  - ✓ Dice
  - ✓ Slice

Extraction, Transformation, and Loading (ETL)

- Data extraction
  - get data from multiple, heterogeneous, and external sources
- Data cleaning
  - detect errors in the data and rectify them when possible
- Data transformation
  - convert data from legacy or host format to warehouse format
- Load
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse



ETL Tools

- ETL tools are the equivalent of schema mappings in virtual integration, but are more powerful
- Arbitrary pieces of code to take data from a source,
- convert it into data for the warehouse:
  - import filters – read and convert from data sources
  - data transformations – join, aggregate, filter, convert data
  - de-duplication – finds multiple records referring to the same entity, merges them
  - profiling – builds tables, histograms, etc. to summarize data
  - quality management – test against master values, known business rules, constraints, etc.

Metadata (data catalog)

- Metadata repository is an integral part of a data warehouse system.
  - Identify subjects of the data mart
  - Identify dimensions and facts
  - Indicate how data is derived from enterprise data warehouses, including derivation rules
  - Indicate how data is derived from operational data store,
  - including derivation rules
  - Identify available reports and predefined queries
  - Identify data analysis techniques (e.g. drill-down)
  - Identify responsible people

Metadata (cont..)

- Normally, it contains the following metadata:
  - **Business metadata** - It contains the data ownership information,
  - business definition, and changing policies.
  - **Operational metadata** - It includes currency of data and data lineage. Currency of data refers to the data being active, archived, or purged. Lineage of data means history of data migrated and transformation applied on it.
  - **Data for mapping from operational environment to data warehouse** - It metadata includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.
  - **The algorithms for summarization** - It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

Schema Design

- Schema refers to the structure or organization of a database.
- It contains a logical description of the entire database, which includes names and descriptions of tables, records, views, and indexes.
- While a relational model is used to describe a database, data warehouse schemas get more specialized because the structure is optimized for reporting and analysis
- Database organization
  - must look like business
  - must be recognizable by business user
  - approachable by business user
  - Must be simple
- Schema Types
  - Star Schema
  - Snowflake schema
  - Fact Constellation Schema

Dimension Tables

- Every dimension contains attributes, which are grouped in the form of a dimension.
- They are essentially a collection of information that can be referenced to answer meaningful business questions when used together with fact tables.
- Hold descriptive information about a particular business perspective
- Define business in terms already familiar to users
  - Wide rows with lots of descriptive text
  - Small tables (about a million rows)
  - Joined to fact table by a foreign key
  - heavily indexed
  - typical dimensions
    - ✓ time periods, geographic region (markets, cities), products,
    - ✓ customers, salesperson, etc.

Fact Table

- Central Table: Multiple dimension tables are linked to one fact table, which contains 'keys' and 'measures'. By 'keys', we're referring to the foreign keys of every associated dimension.
- Keys are used to perform joins with dimension tables to run queries. 'Measures' refer to numeric data like price and quantity, which represents business events or transactions, used to add detail to dimension data, so that effective reports can be generated.
- Key value is a composite key made up of the primary keys of the dimensions
- Joined to dimension tables through foreign keys that reference primary keys in the dimension tables
  - Typical example: individual sales records mostly raw numeric items narrow rows, a few columns at most large number of rows (millions to a billion)
  - Access via dimensions

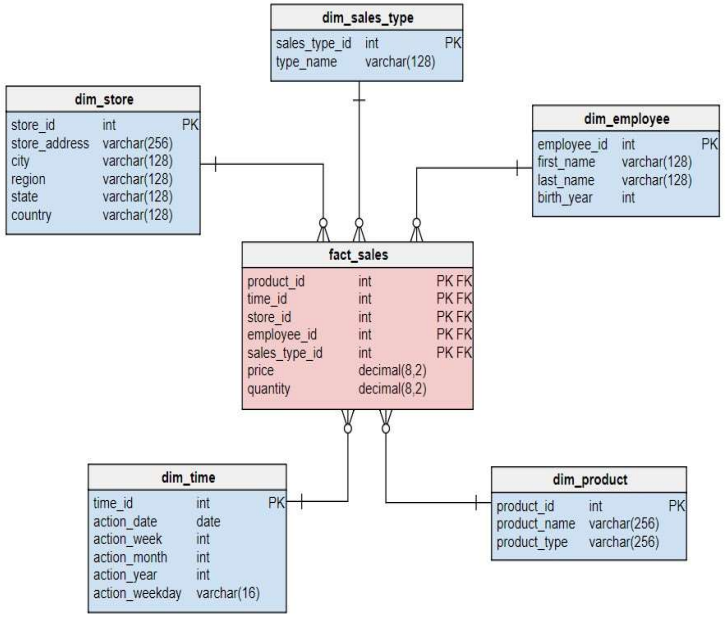
Work space 1

Work space 2

Work space 3

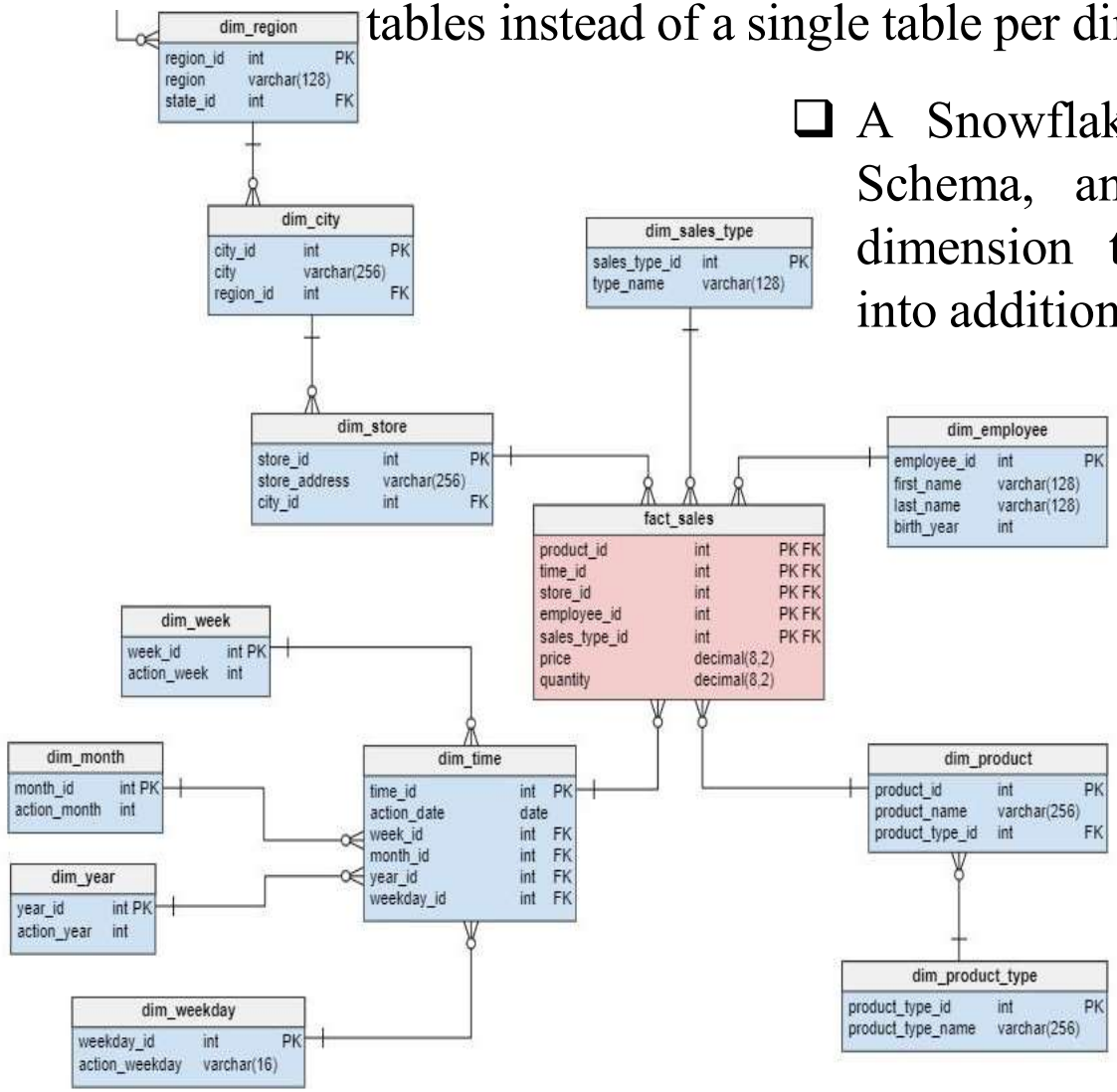
Star Schema

- In the STAR Schema, the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star.
  - The dimension table should contain the set of attributes.
  - The dimension table is joined to the fact table using a foreign key
  - The dimension table are not joined to each other
  - Fact table would contain key and measure
  - The Star schema is easy to understand and provides optimal disk usage.
  - The dimension tables are not normalized. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
  - The schema is widely supported by BI Tools



Snowflake Schema

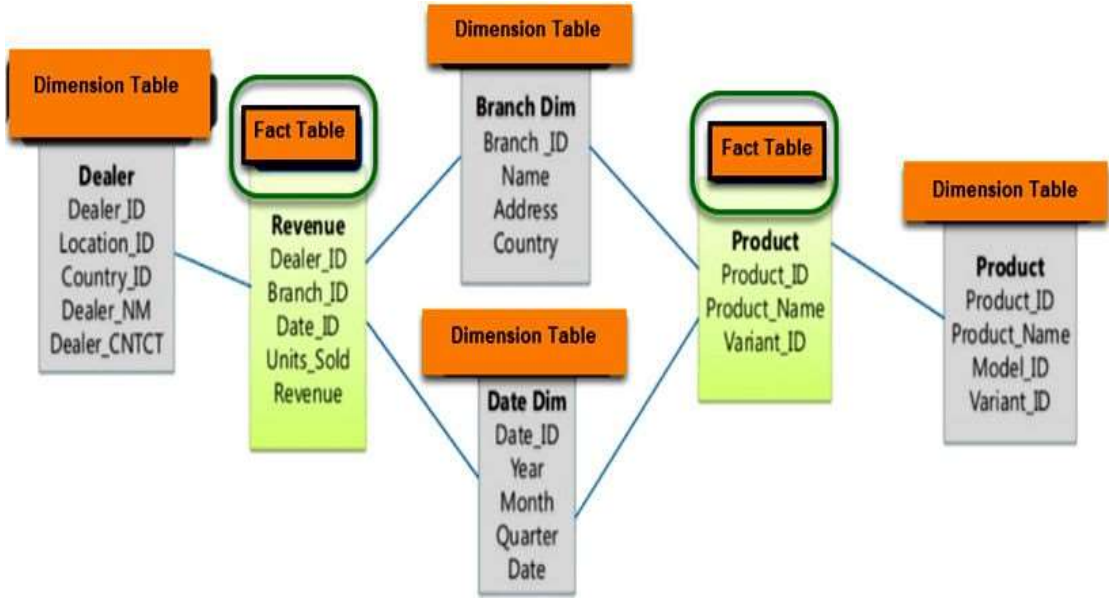
- In the snowflake schema, dimensions are stored in multiple dimension tables instead of a single table per dimension.

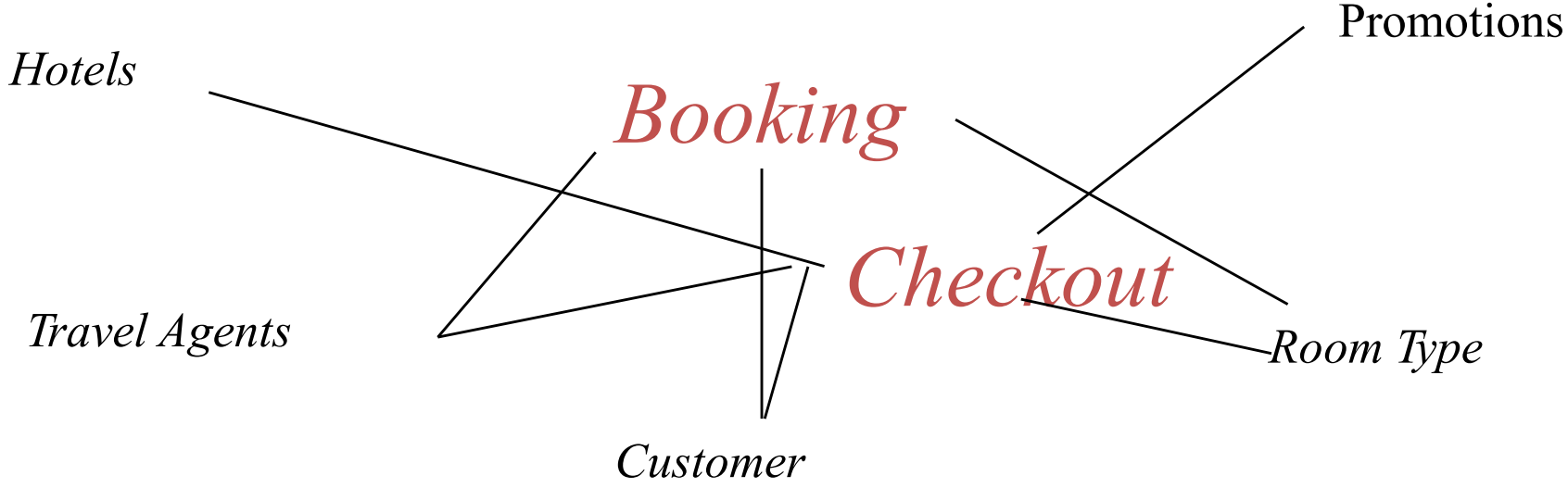


- A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. The dimension tables are normalized which splits data into additional tables.
  - The main benefit of the snowflake schema it uses smaller disk space.
  - Easier to implement a dimension is added to the Schema
  - Due to multiple tables query performance is reduced
  - The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

Fact Constellation Schema

- Multiple fact tables that share many dimension tables.
- The schema is viewed as a collection of stars hence the name Galaxy Schema.
- Booking and Checkout may share many dimension tables in the hotel industry





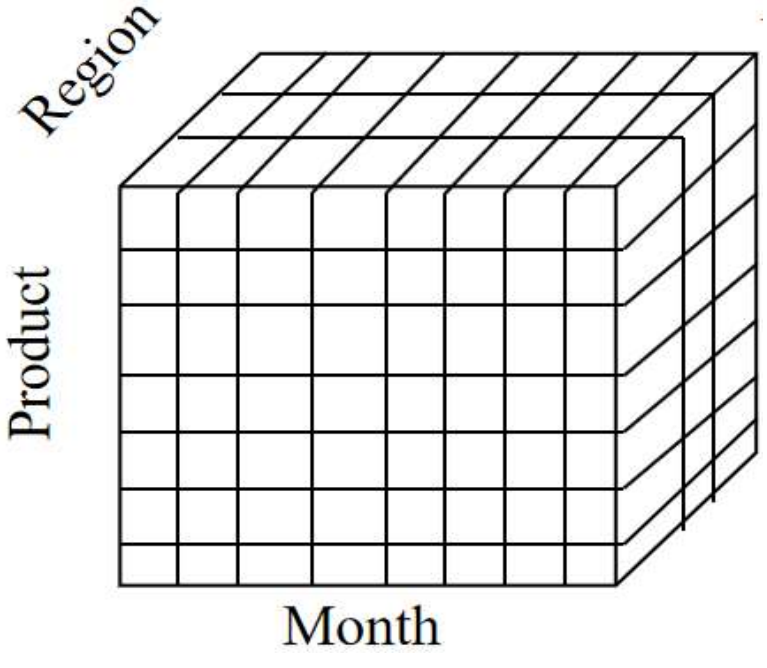
Star Vs Snowflake Schema: Key Differences

Star Schema	Snow Flake Schema
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized Data structure and query also run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.
Cube processing is faster.	Cube processing might be slow because of the complex join.
Offers higher performing queries using Star Join Query Optimization. Tables may be connected with multiple dimensions.	The Snow Flake Schema is represented by centralized fact table which unlikely connected with multiple dimensions.

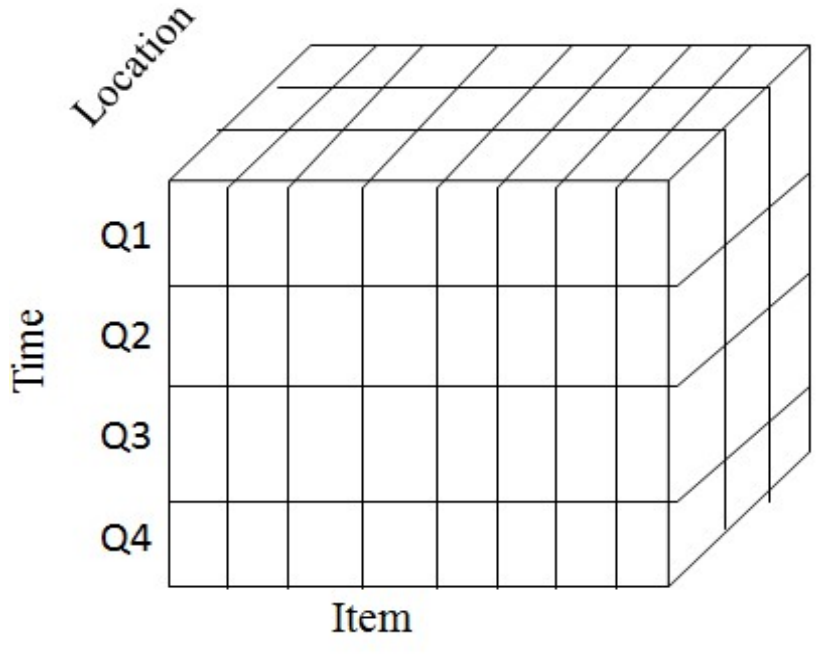
Multi-Dimensional Data

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube.
- OLAP databases are divided into one or more cubes. The cubes are designed in such a way that creating and viewing reports become easy.
  - Measures - numerical data being tracked
  - Dimensions - business parameters that define a transaction
  - Example: Analyst may want to view sales data (measure) by
  - geography, by time, and by product (dimensions)
  - Dimensional modeling is a technique for structuring data around business concepts
  - ER models describe “entities” and “relationships”
  - Dimensional models describe “measures” and “dimensions”

Multi-Dimensional Data



- Three dimensional data
- X-axis Item
- Y-axis Time
- Z-axis Location

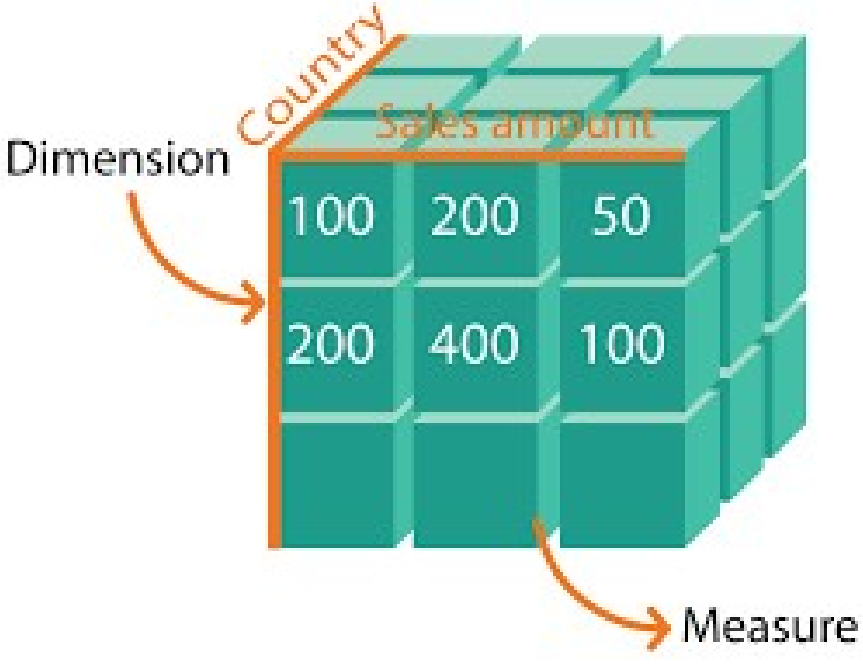


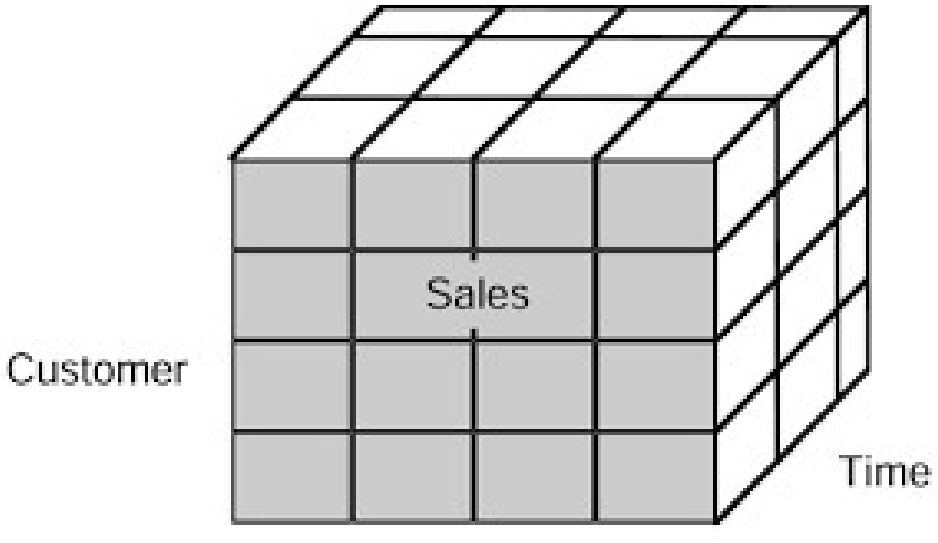
OLAP - Cube, Operations

- **OLAP:** It is a category of software that allows users to analyze information from multiple database systems at the same time.
- **Data cubes**
  - A data warehouse is based on a multidimensional data model which views data in the form of a data cube
  - Data cube helps to arrange a complex data in a simple format.
  - Data cube represents the data along some measures of an interest.
  - It can be of 2-dimensional, 3-dimensional and higher dimensional
  - Mainly used for the retrieval of the data
  - It consists of categories of data called dimensions and measures.
  - Measure and dimension represents fact such as cost, time and locations.

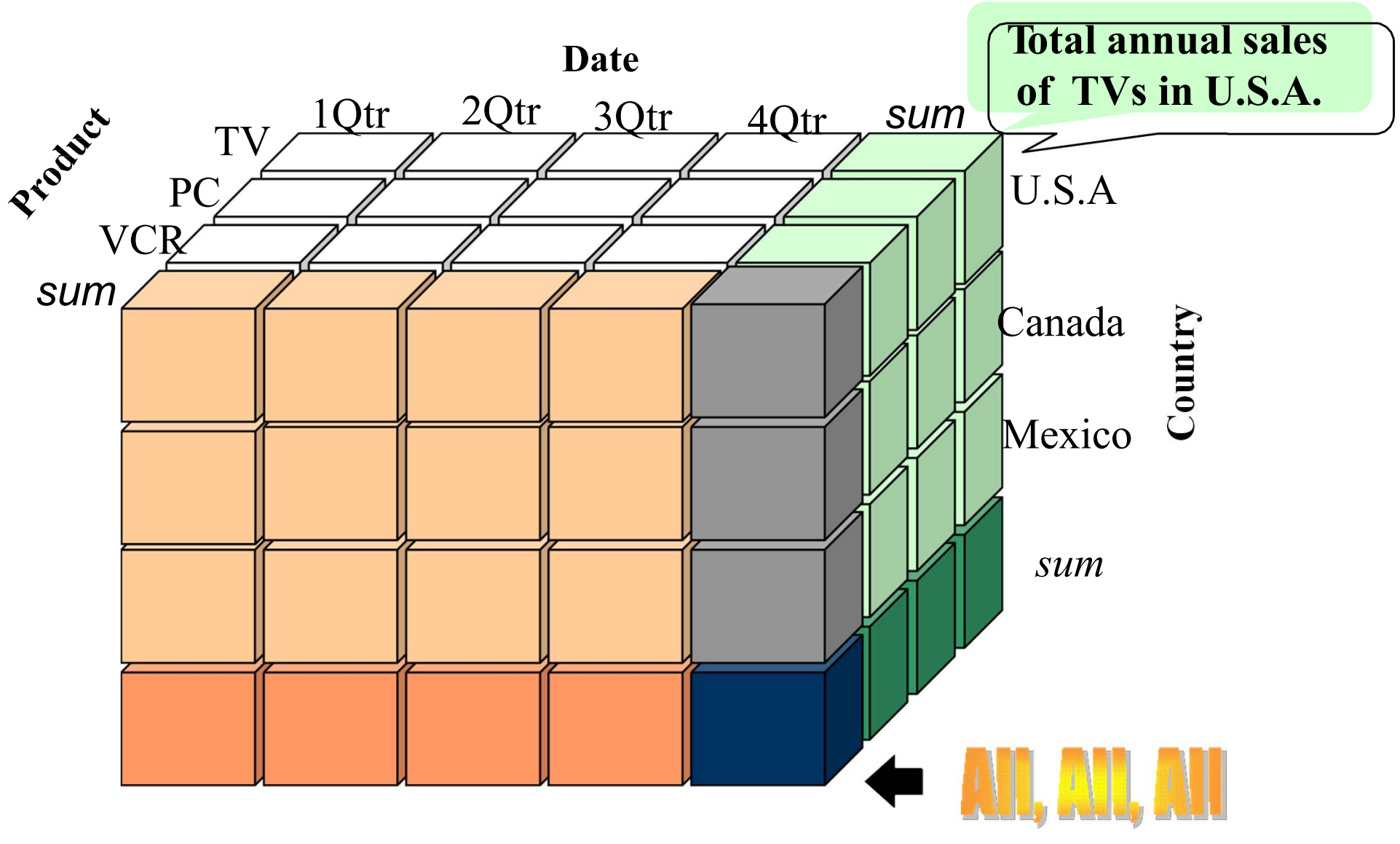
Measures and Dimensions

3-Dimensional Data

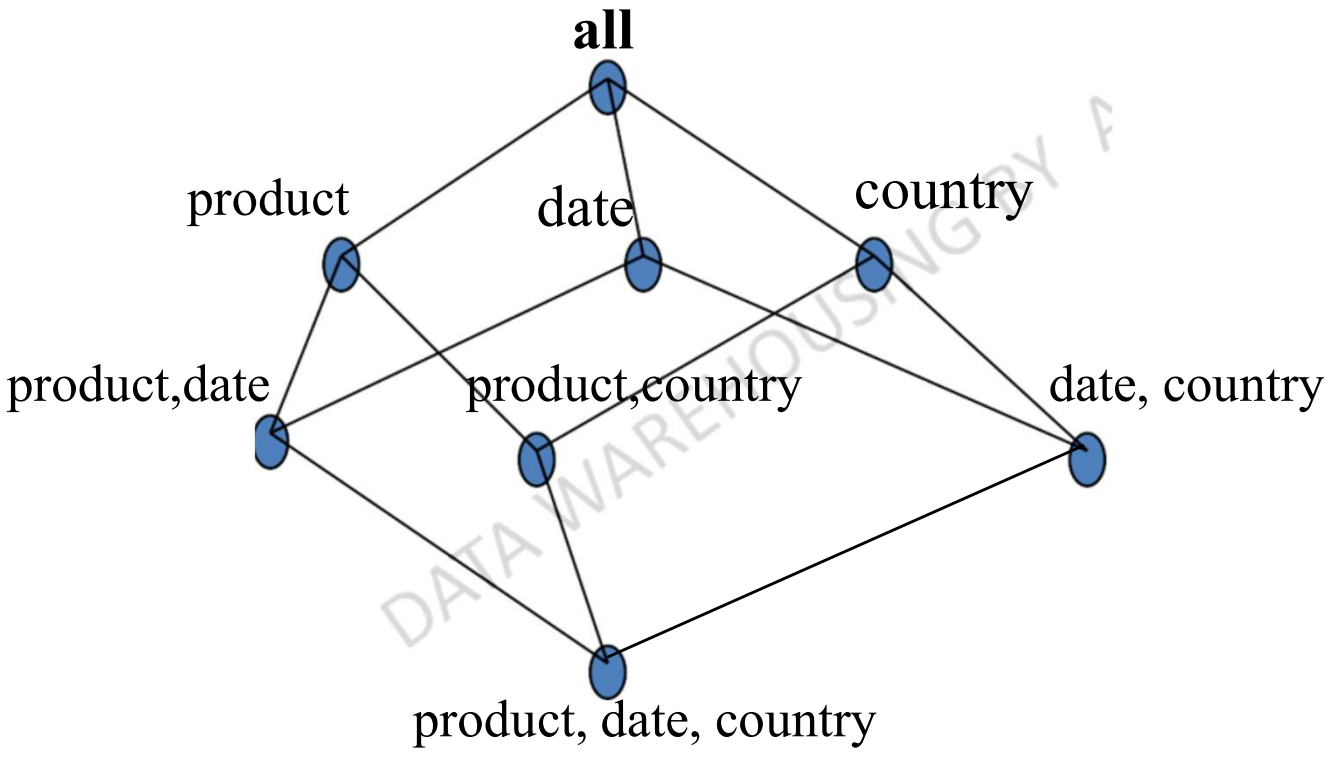




A Sample Data Cube



Cuboids Corresponding to the Cube



0-D (*apex*) cuboid

1-D cuboids

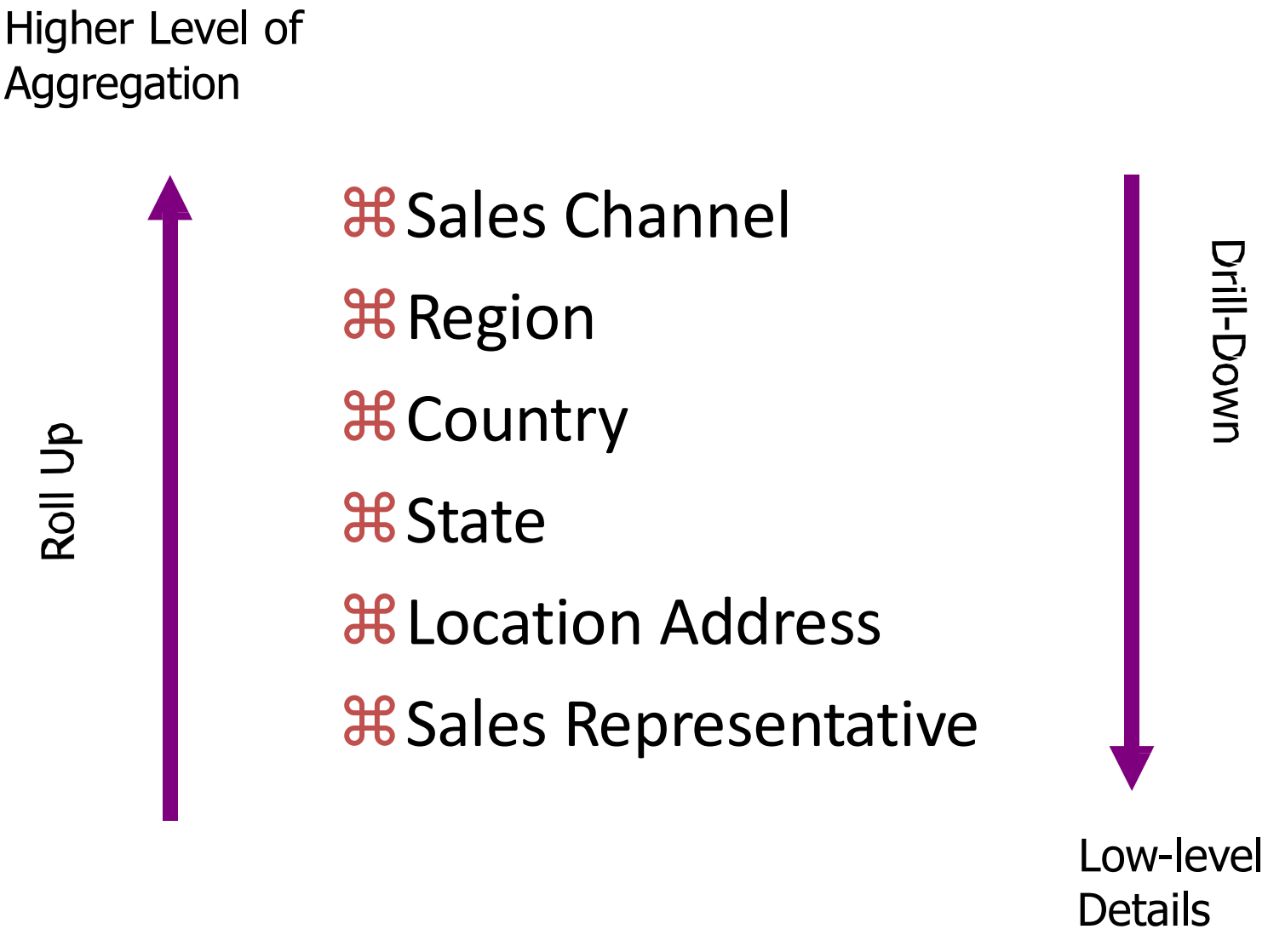
2-D cuboids

3-D (*base*) cuboid

Operations on Data cubes

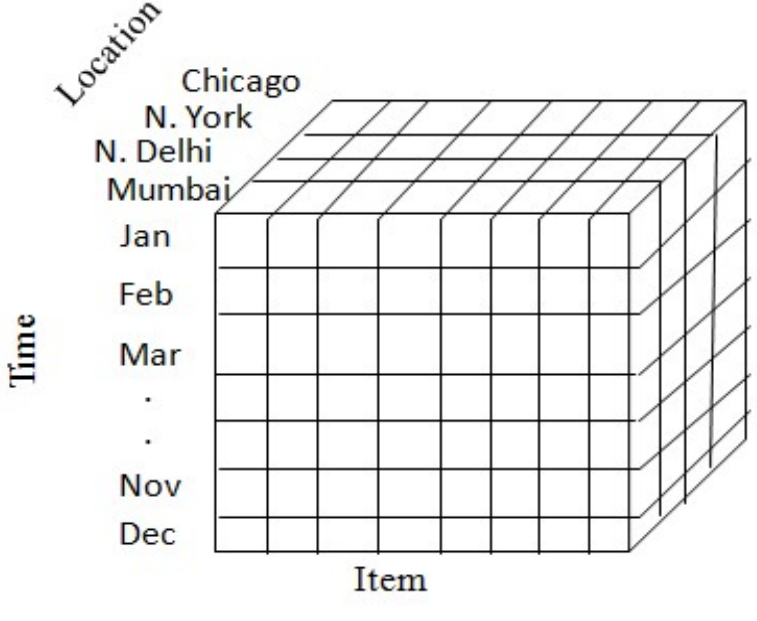
- Drill Down
- Roll Up
- Dice
- Slice
- Pivot

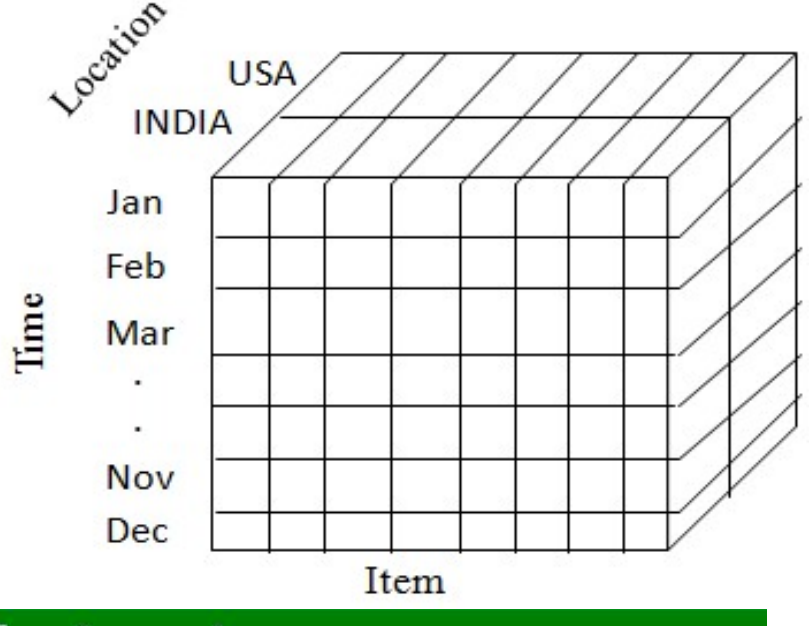
Roll-up and Drill Down



Roll up

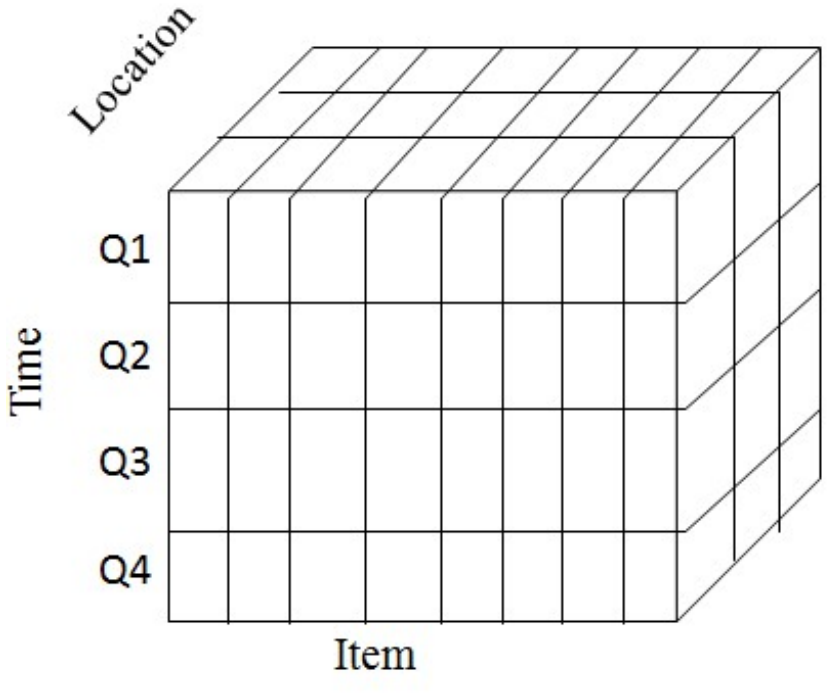
- Roll-up is also known as "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways
  - Reducing dimensions
  - Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.
- Applying roll up operation on LOCATION we have, so we can roll up to its COUNTRY to USA and INDIA only

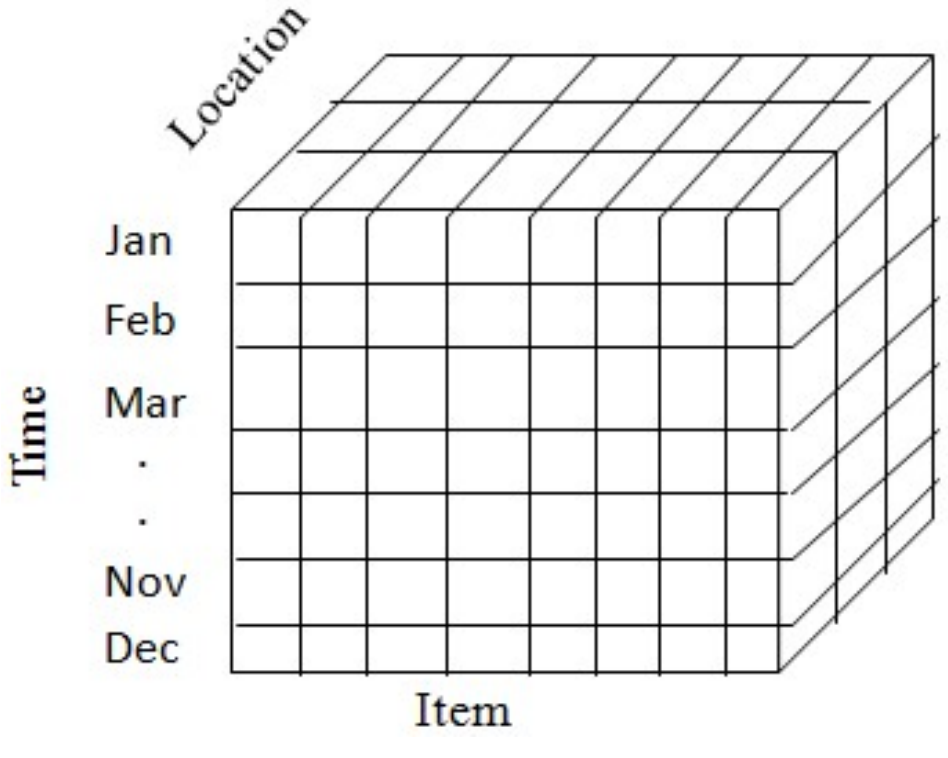




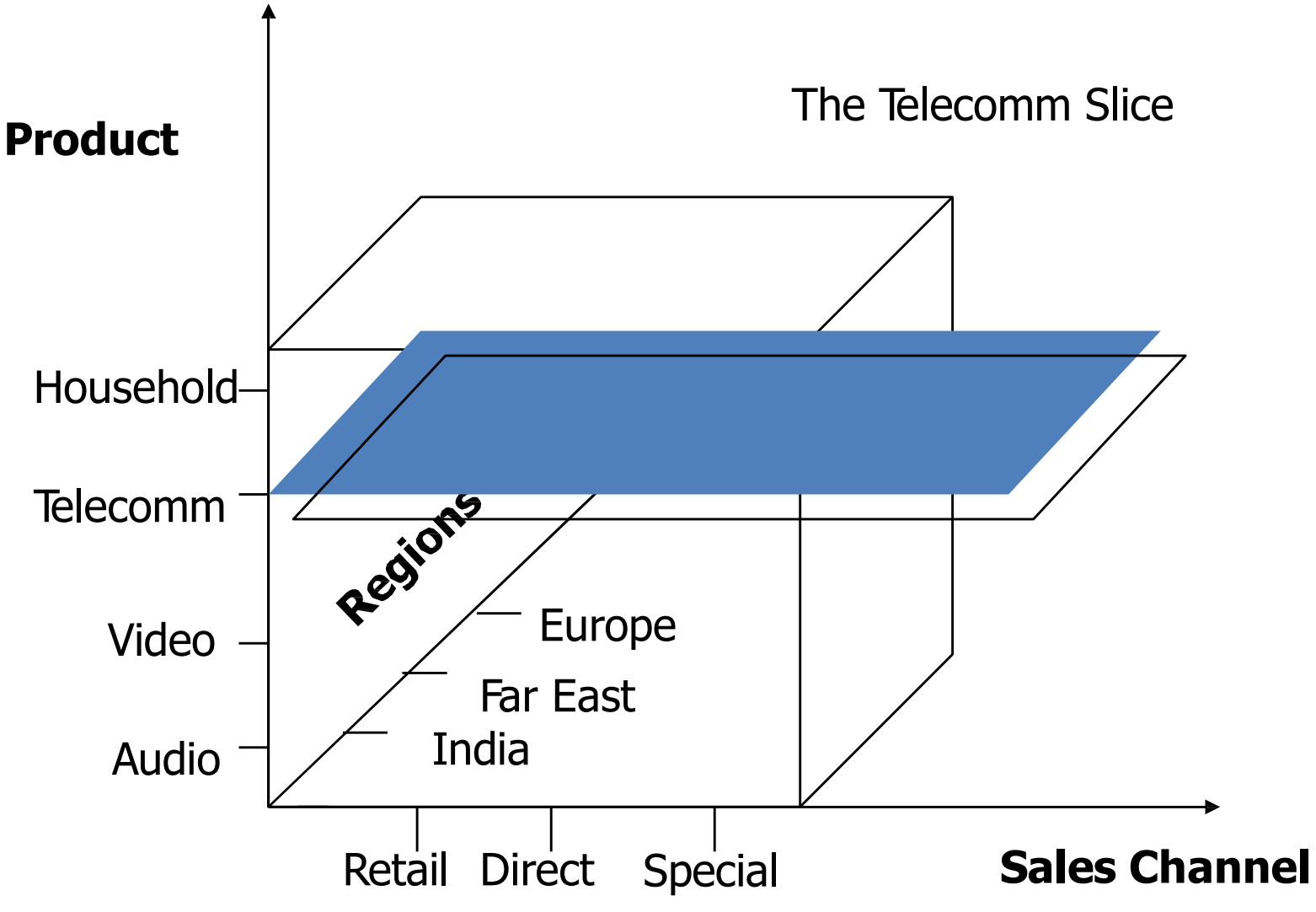
Drill down

- More detail information can be retrieved from this.
- Let us consider that the time axis has to drill down to get more information by moving down the concept hierarchy by adding new dimension
- To get more details in all quarters, we mentioned the months from Jan to Dec



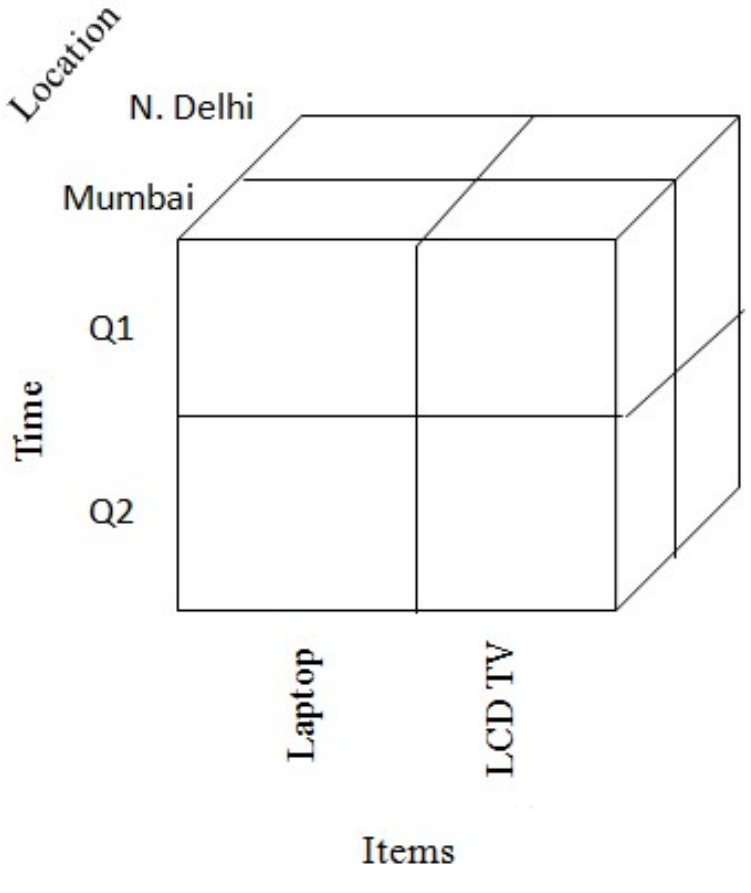


Slicing and Dicing



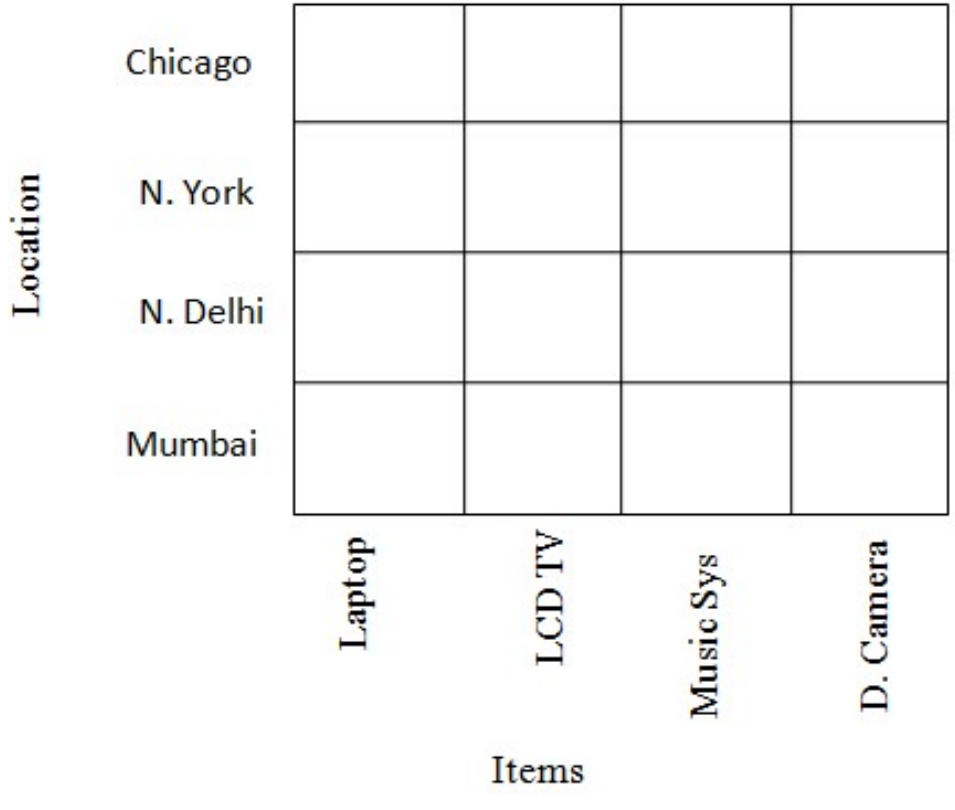
Dice

- It selects a sub-cube from the data cube by selecting two or more dimensions.
- A sub-cube is selected by selecting following dimensions with criteria:
- Location = “Mumbai” or “N. Delhi”
- Time = “Q1” or “Q2”
- Item = “Laptop” or “LCD TV”



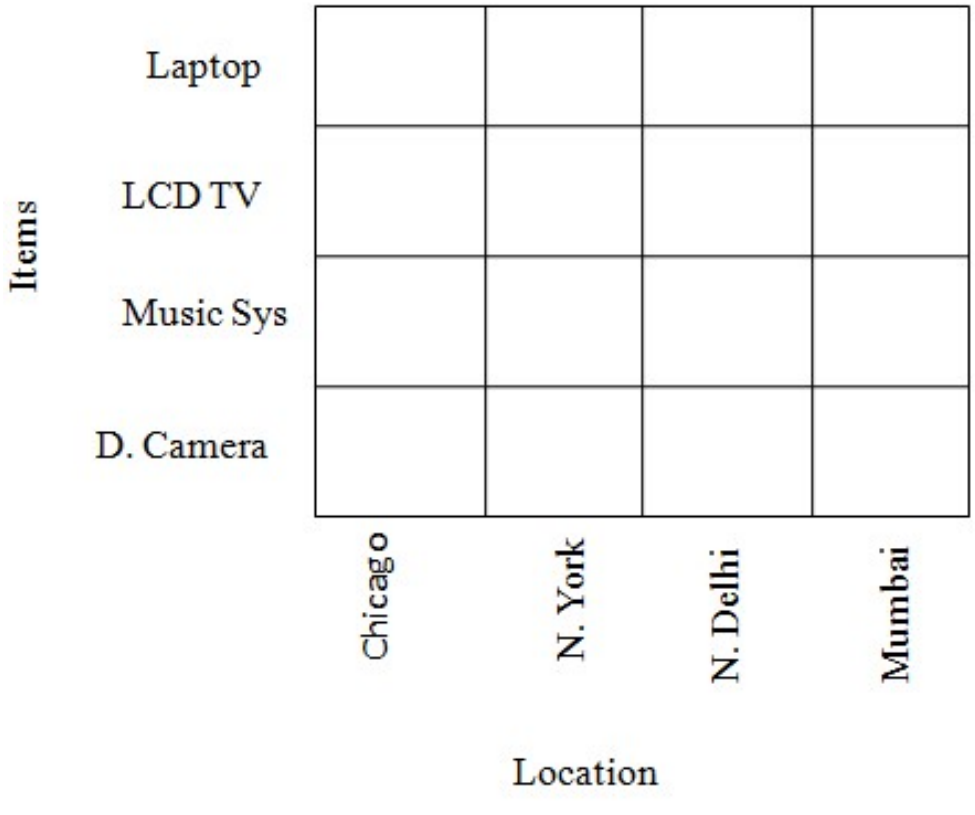
Slice

- It selects a single dimension from the data cube which results in a new sub-cube creation. A Slice is performed on the dimension Time = “Q1”.



Pivot

- It is also known as rotation operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.



Recommended Text and Reference Books

- **Text Book:**
  - J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- **Reference Books:**
  - H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2006.
  - I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2000.
  - D. Hand, H. Mannila and P. Smyth. Principles of Data Mining.Prentice-Hall. 2001.

THANK YOU!