An Overview of Data Warehousing and OLAP Technology

Karanjit Singh Tiwana

CMPT 843

Intro and Architecture

What is Data Warehouse?

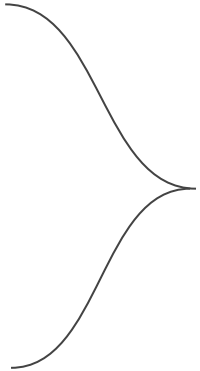
“Subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making”

Where is it used?

- Manufacturing
  - Order shipment
  - Customer Support
- Retail
  - User Profiling
  - Inventory Management
- Financial services
  - Risk Analysis
  - Fraud Detection
- Transportation
  - Fleet Management
- Telecommunications
- many more..

Why do we need it?

- Analysis
- Knowledge Extraction
- Decision Making
- Visualization
- BI
- Easy to understand and Summarize



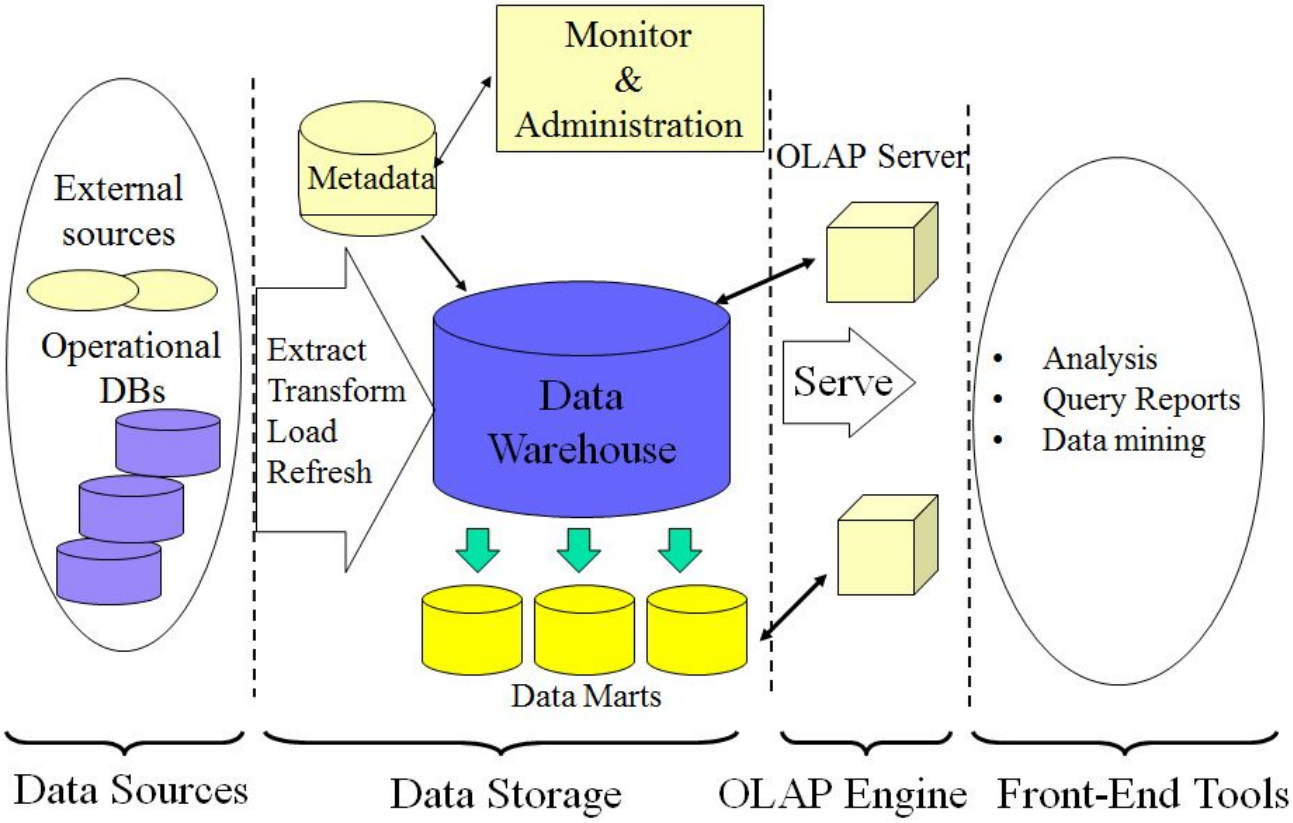
OLAP

Example - tabular vs multidimensional

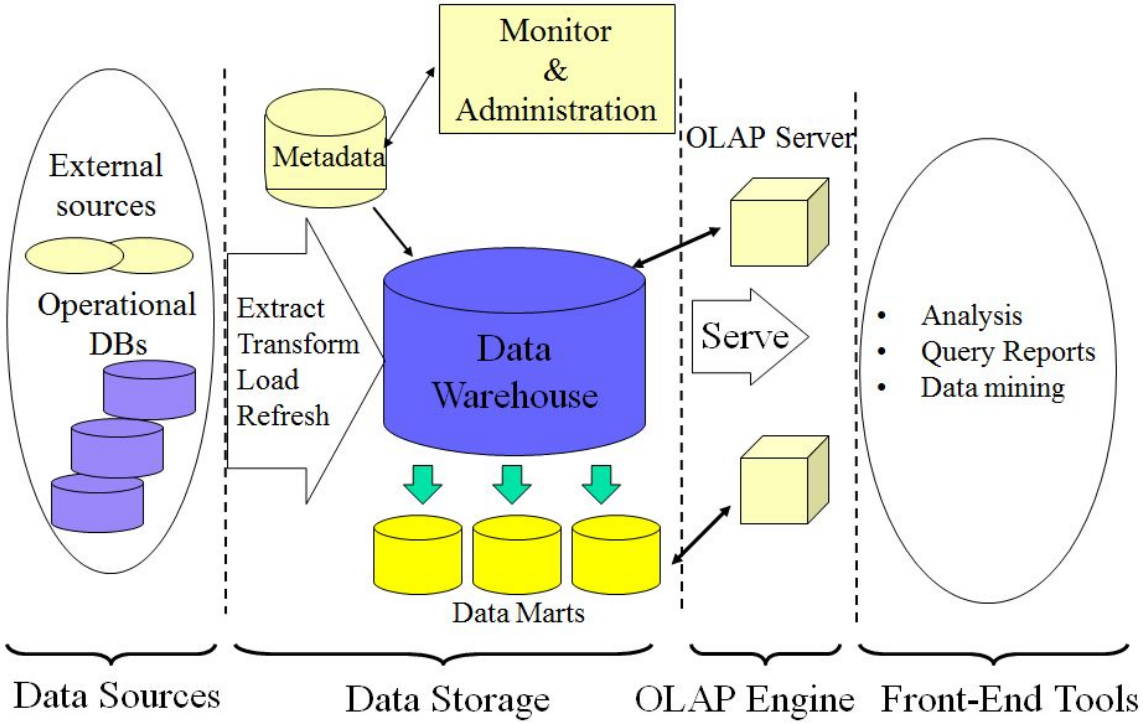
Product	Region	Sales
Nuts	East	50
Nuts	West	60
Nuts	Central	100
Nuts	**Total**	**210**
Screws	East	40
Screws	West	70
Screws	Central	80
Screws	**Total**	**190**
Bolts	East	90
Bolts	West	120
Bolts	Central	140
Bolts	**Total**	**350**
Washers	East	20
Washers	West	10
Washers	Central	30
Washers	**Total**	**60**
Total	**East**	**200**
Total	**West**	**260**
Total	**Central**	**350**
Total	**Total**	**810**

	East	West	Central	Total
Nuts	50	60	100	**210**
Screws	40	70	80	**190**
Bolts	90	120	140	**350**
Washers	20	10	30	**60**
Total	**200**	**260**	**350**	**810**

Architecture

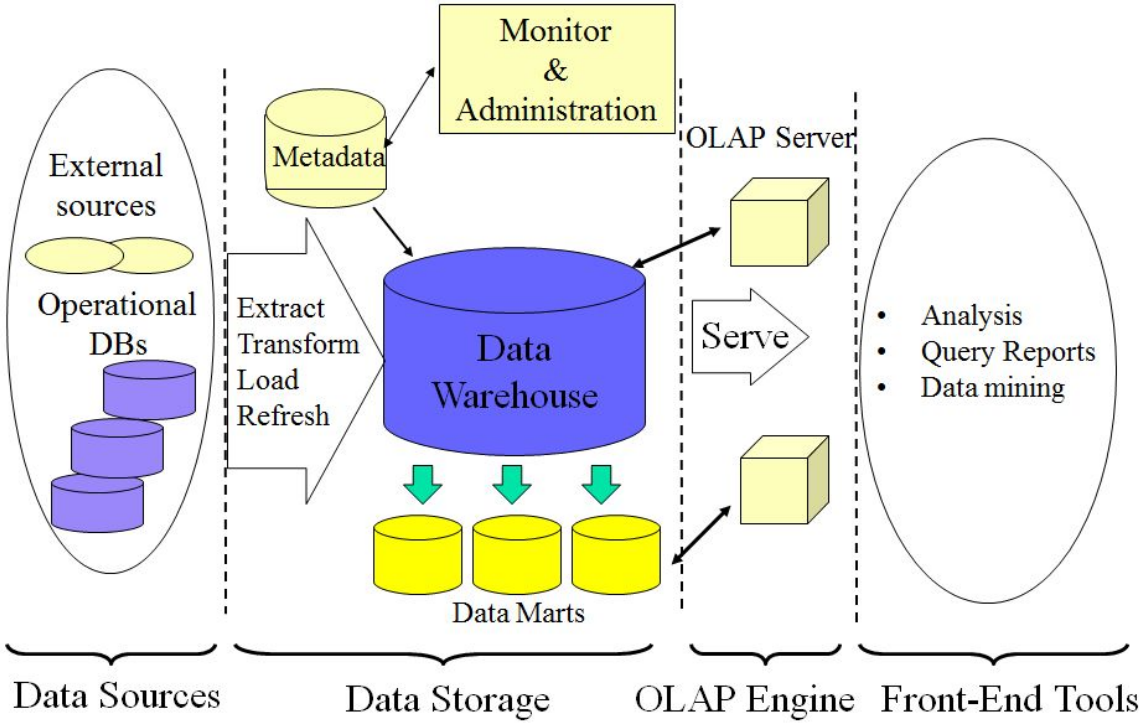


Monitor and Administrator



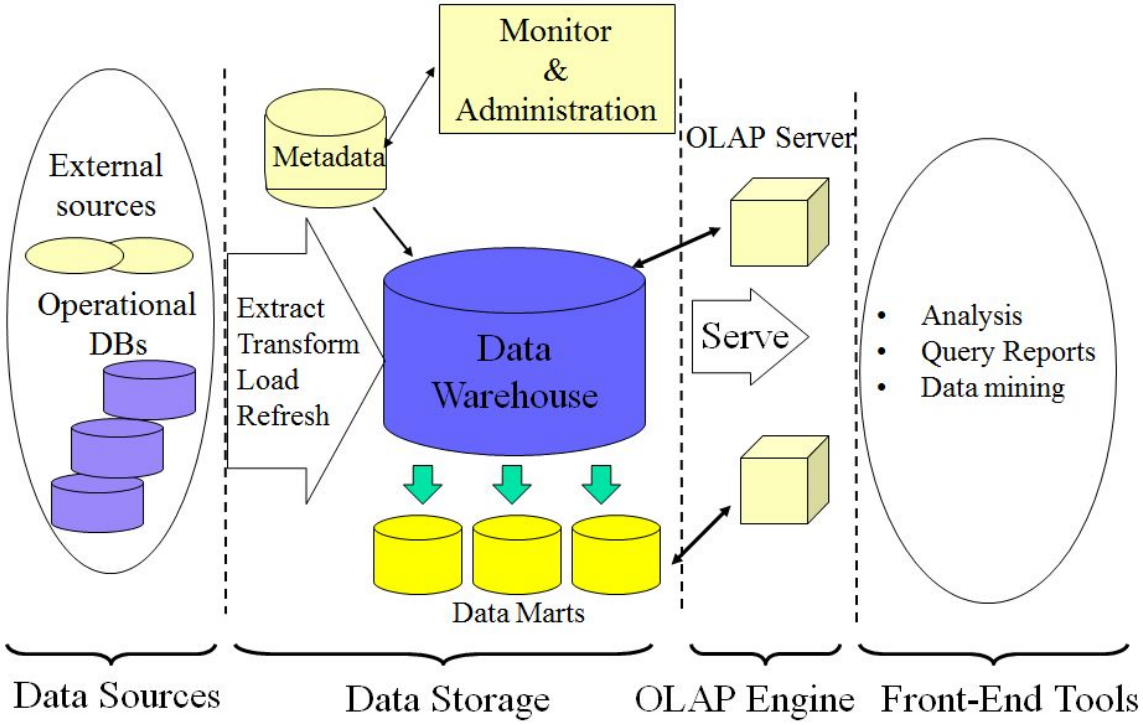
- Analyzes the data to perform consistency and referential integrity checks.
- Creates indexes, business views, partition views against the base data.
- Generates new aggregations and updates existing aggregations. Generates normalizations.

Metadata



- Metadata is the road-map to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

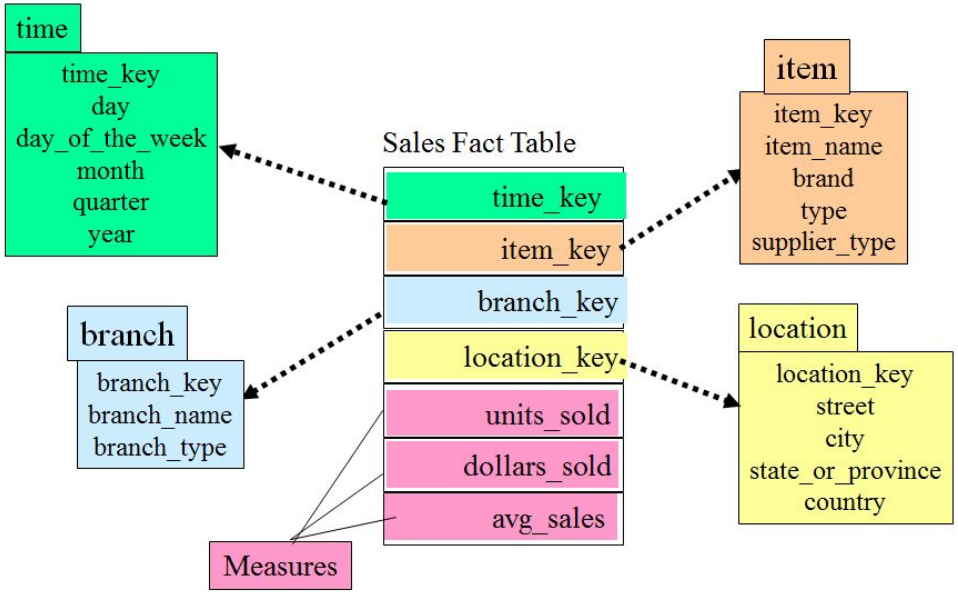
Data Marts



The data mart is a subset of the data warehouse and is usually oriented to a specific business line or team.

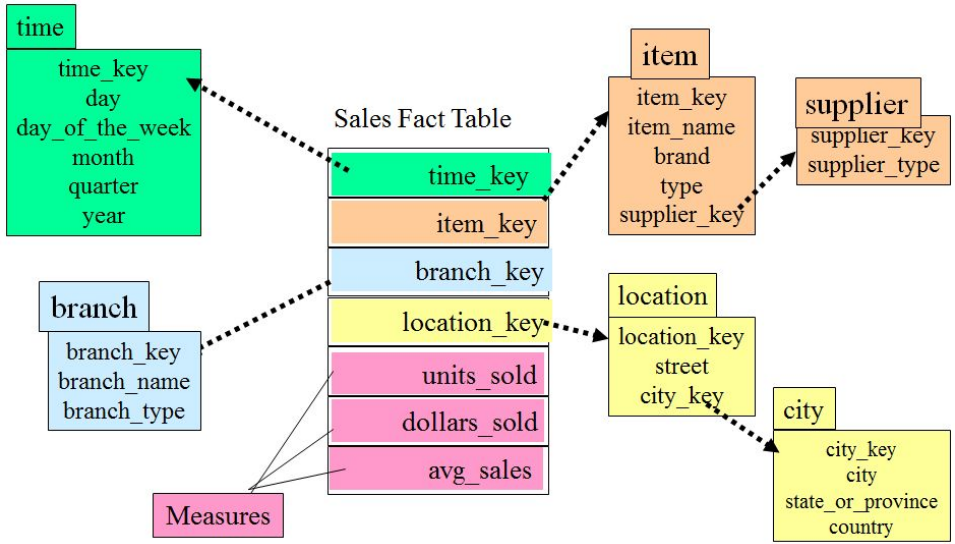
Schema Representation

Star Schema



- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Snowflake Schema



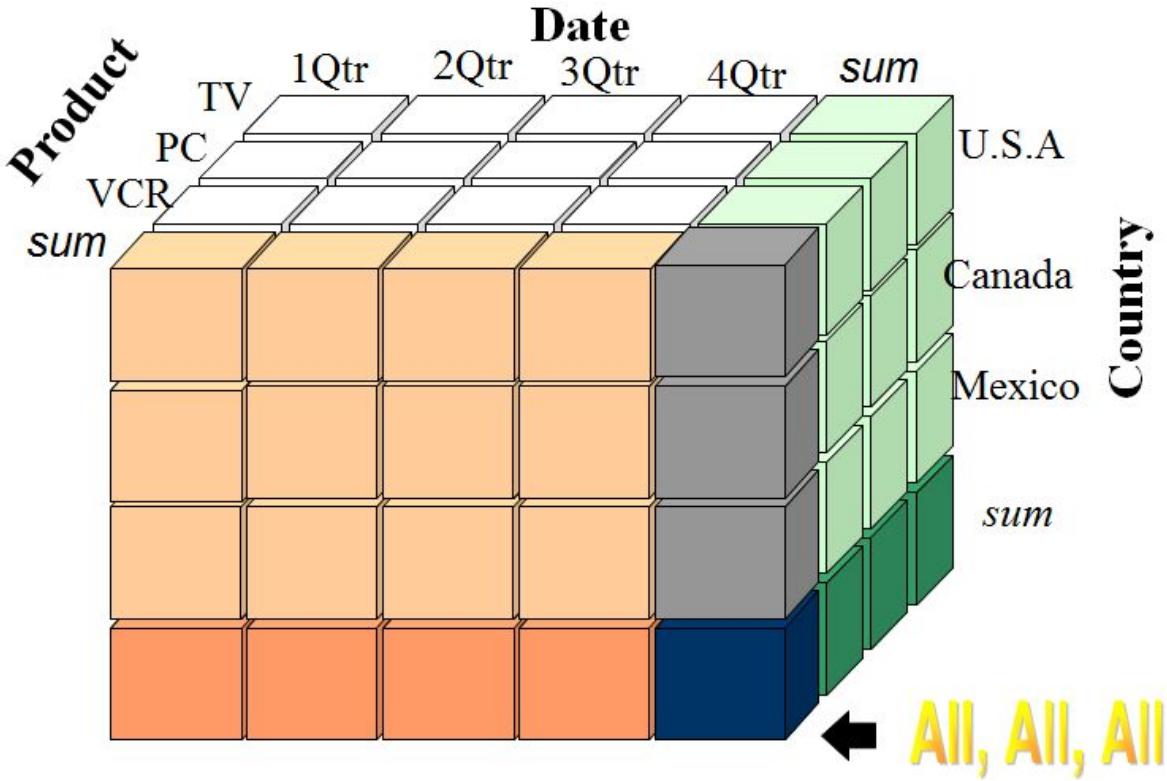
- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

OLAP

What is OLAP?

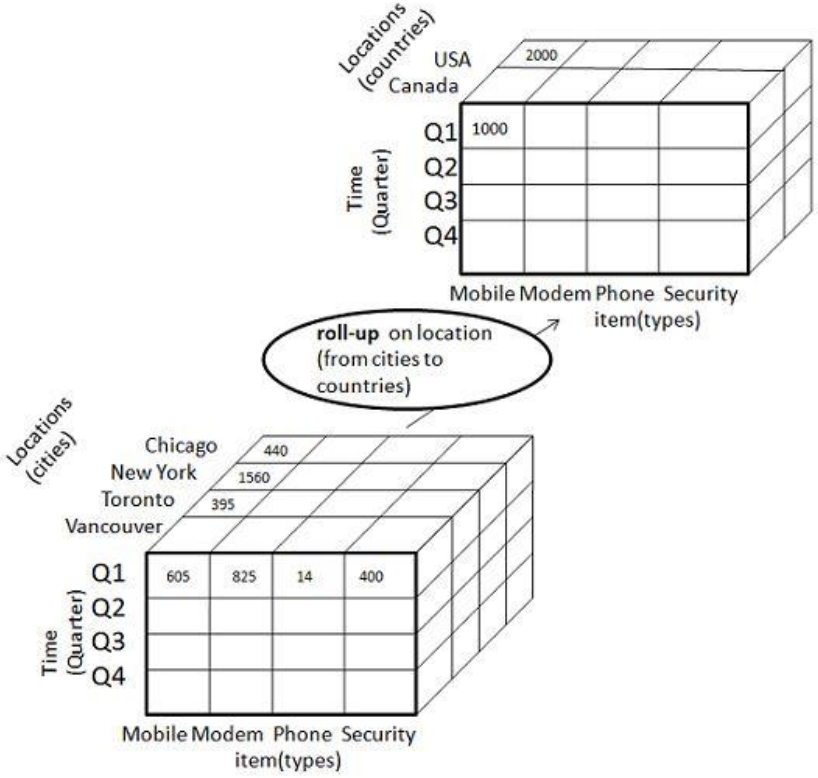
- Online analytical processing, or OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly in computing
- Databases configured for OLAP use a multidimensional data model, allowing for complex analytical and ad hoc queries with a rapid execution time.
- Holds summaries, aggregations and creates hierarchy

Example - Data Cubes



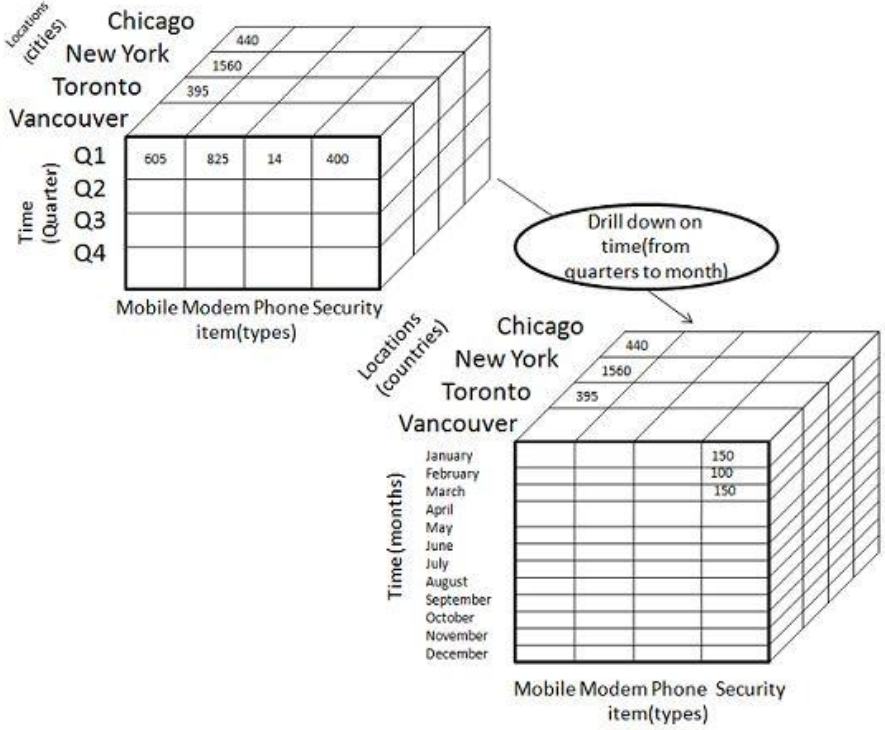
Operations

Roll-up



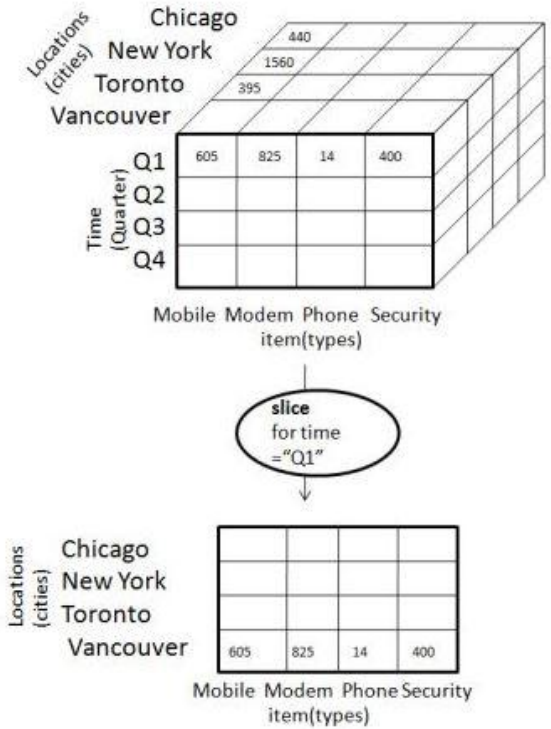
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.

Drill-down



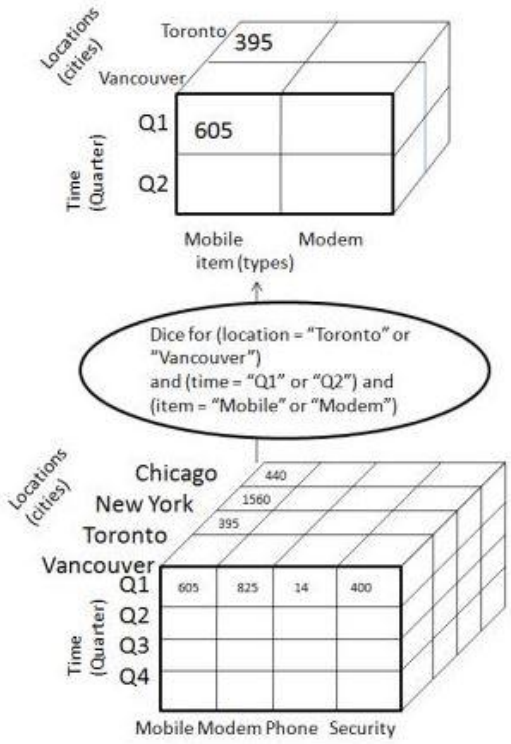
- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.

Slice



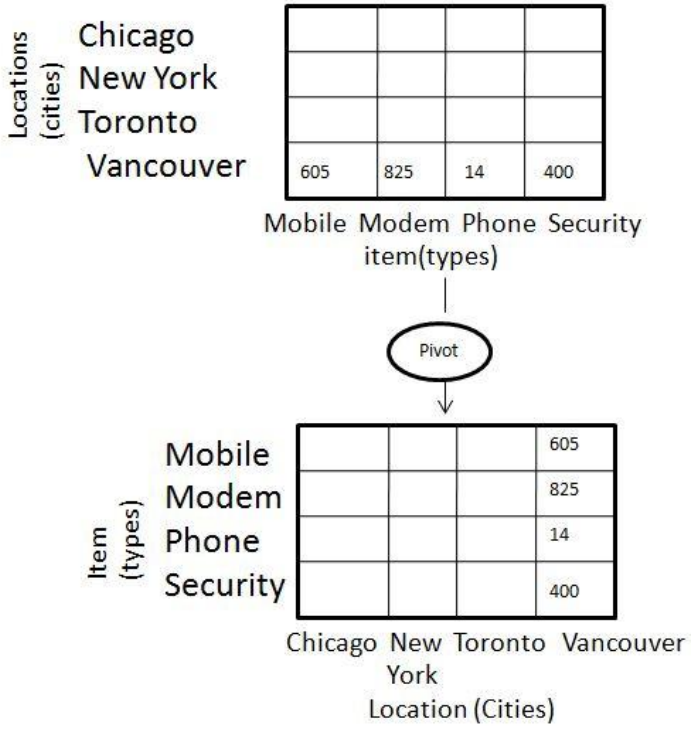
- The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.
- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

Dice



- Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.
- The dice operation on the cube based on the following selection criteria involves three dimensions.
  - (location = "Toronto" or "Vancouver")
  - (time = "Q1" or "Q2")
  - (item =" Mobile" or "Modem")

Pivot



The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.

Server Techniques

BitMap index

record number	*name*	*gender*	*address*	*income_level*
0	John	m	Perryridge	L1
1	Diana	f	Brooklyn	L2
2	Mary	f	Jonestown	L1
3	Peter	m	Brooklyn	L4
4	Kathy	f	Perryridge	L3

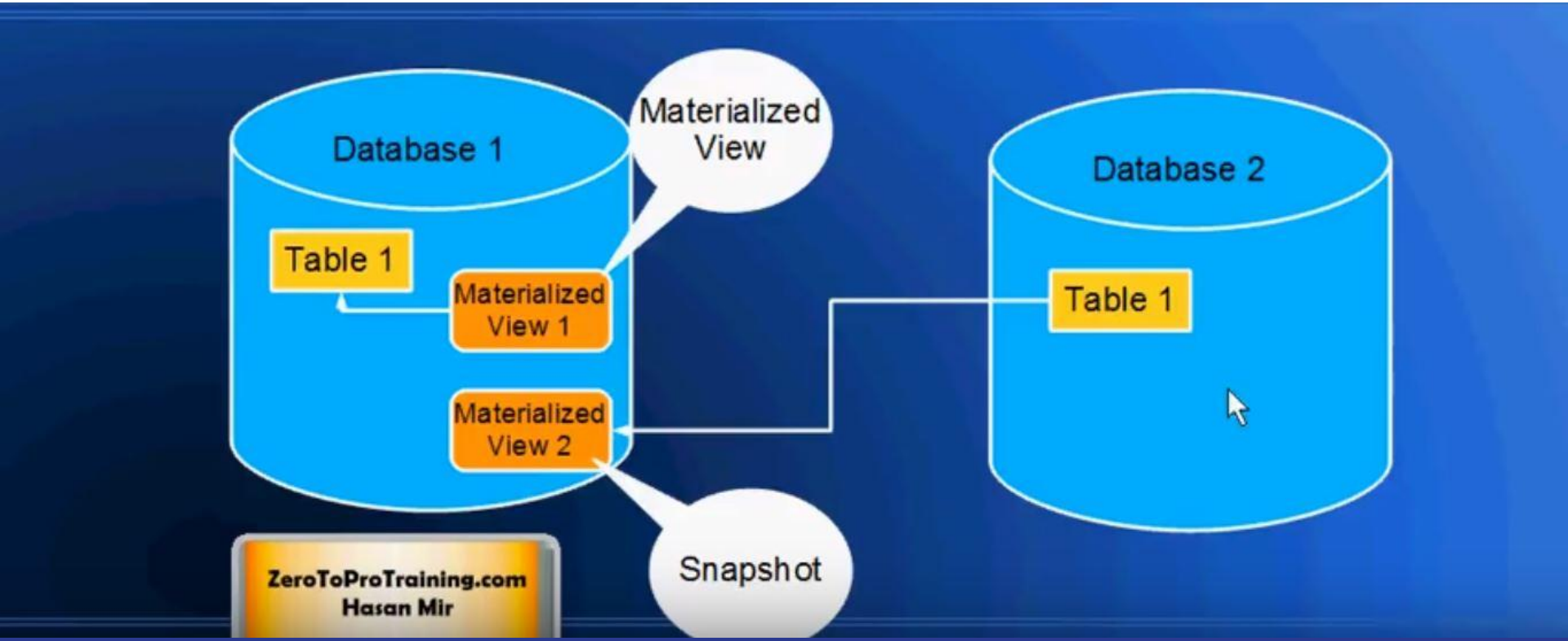
Bitmaps for *gender*

m	10010
f	01101

Bitmaps for *income_level*

L1	10100
L2	01000
L3	00001
L4	00010
L5	00000

Materialized Views



Adv:

- Store Subset of the data
- Join of multiple tables
- Summarization of Table data

Dis- Adv:

- Not up-to-date data

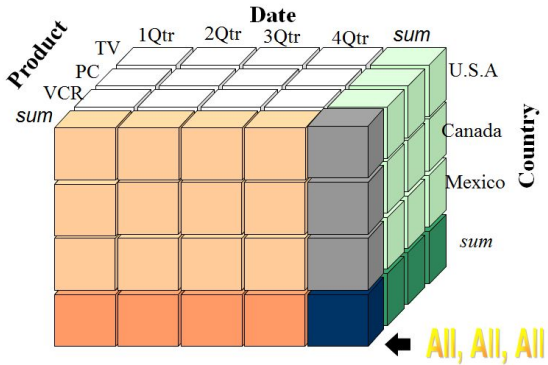
Server Architecture

ROLAP - Relational Online Analytical Processing

- manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality.
- Array of inbuilt functionalities because of RDBMS backend

MOLAP - Multidimensional Online Analytical Processing

- Data is stored in array-based structure
- Aggregations are calculated when cube is generated



MOLAP vs ROLAP

MOLAP	ROLAP
Information retrieval is fast.	Information retrieval is comparatively slow.
Uses sparse array to store data-sets.	Uses relational table.
MOLAP is best suited for inexperienced users, since it is very easy to use.	ROLAP is best suited for experienced users.

Questions?

Fact Constellation