Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

JUNAID GIRKAR
60004190057
TE COMPS A4

DWM
ASSIGNMENT 6

Q1. Illustrate metadata type in data warehouse with suitable examples?

ANS: Metadata in data warehouse is similar to data dictionary or the data catalogue in DBMS. The data dictionary contains data about the data in the data warehouse. It contains logical data structures, information about files and address, information, metrics to analyze Data warehouse and security authorizations, etc.

Metadata in a data warehouse fall into three major categories :

1) **OPERATIONAL METADATA**:

As data for the data warehouse comes from several operational systems of the enterprise. These source systems contain different data structures. The data elements selected for the data warehouse have various field lengths and data types. In selecting information from source systems for data warehouses, we divide records, combine factors of documents from different source files and deal with multiple coding schemes and field lengths. When we deliver information to end users we must be able to tie that back to the source data sets. Operational metadata contains all information about operational data sources

Example:

Operational metadata contains information about the source of data in the data warehouse. Considering the end-user wants to know from where the data has been retrieved, the operational metadata will give the source operational system amongst all operational systems and other operational data source information related to it.

2) **EXTRACTION AND TRANSFORMATION METADATA**:

Extraction and Transformation metadata contain data about the extraction of data from source systems, namely the extraction frequencies, extraction methods and business rules for the data extraction. Also, this category of metadata contains information about all data transformations that take place in the data staging .

Example:

Consider, some operational systems had gender stored in the form ‘M’ & ‘F’ and some had them stored in the form ‘D’ & ‘I’. At data staging, all were converted to ‘M’ & ‘F’. This information about transformation is stored in the extraction and transformation metadata. It also stores the frequency of data extraction. E.g: 1 month, 6 month, yearly, etc. It also stores which method was used for data extraction. E.g. Immediate / Deferred; and other extraction and transformation related information.

3) **END-USER METADATA**:

The end user metadata is the navigational map of the data warehouse. It enables the end users to find information from the data warehouse. The end user metadata allows the users to use their own business terminology and look for information in those ways in which they normally think of the business.

Example:

The end user metadata stores indexes of the data stored in the data warehouse. Through these indexes the end-user can find or search the data of his / her need or interest.

Q2. Differentiate between OLTP and OLAP

Distinguishing Term	OLTP	OLAP
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to	Instead of regular backups, some environments may consider simply reloading the

	entail significant monetary loss and legal liability	OLTP data as a recovery method
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support

Q3. Describe and exemplify Type 2 and Type 3 changes to a data warehouse

ANS: **Type 2 change: Preservation of History**

This type keeps track of historical data by adding multiple rows of given key columns in the dimension table. We capture attribute change by adding a new row with the surrogate key to the dimension table.

Consider a customer table storing details of the customer. It has a 'marital status' attribute. If a customer eventually was 'single', it may be possible that over time his status may change to 'married'. If change to marital status is important for business analysis, the history must be preserved. Suppose the customer was married on 6th June 2001 then all records before 06/06/2001 must have marital status = 'single' and after 06/06/2001 must have marital status = 'married' referring to the same customer. This type of change partitions history in the data warehouse. Every change has to be preserved.

Attribute	**Before 06/06/01**	**After 06/06/01**
Customer Key (SK)	10001	24125
Customer Name	Harry	Harry
Customer Code (PK)	C124	C124
Marital Status	Single	Married
Address	22 Bleecker Street	22 Bleecker Street
State	New York	New York
Zip Code	10012	10012

Method for applying type 2 change:-

- Add a new row with new value of changed attribute
- Effective date field may be included in dimension table
- There are no changes to the original row
- Key of original row is not affected
- New row is inserted with a new surrogate key.

Type 3 change: Tentative Soft Revision :-

This type keeps the limited history of changed data in the form of a new column. History preservation is limited to the number of columns designed for storing historical data. It keeps a limited history about changed data. It is the least common type of change.

Consider, a department is contemplating a realignment of territorial alignments for salespersons. Before making realignment, they want to count data in 2 ways: based on current alignment and based on new alignment. This tentative change in a type 3 change. Example: Salesperson Jack Smith is being moved from Goa to Kerala with the ability to trace his orders in both territories. This type is used to compare the performances across transitions.

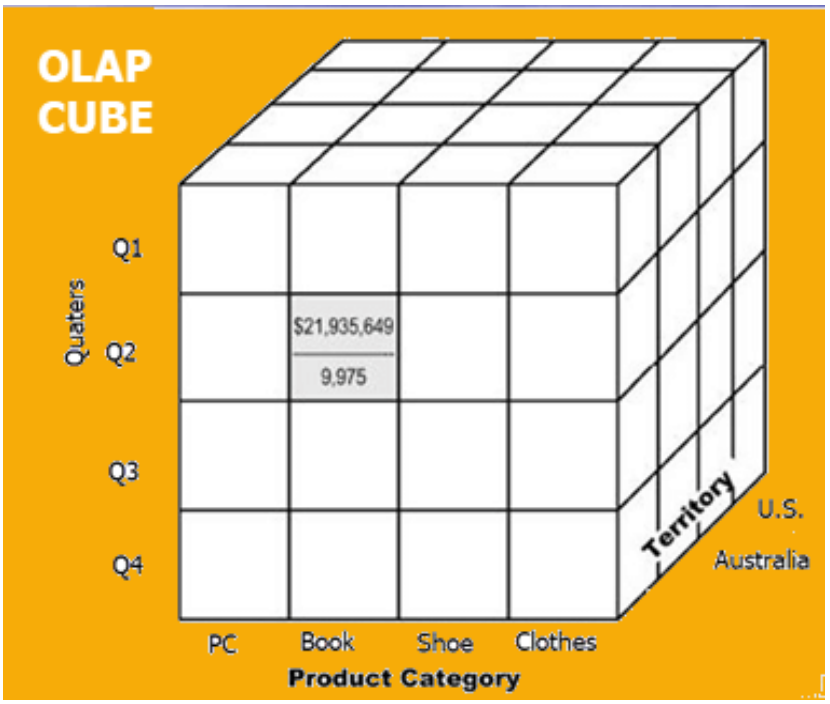
Attribute	**Before**	**After**
Salesperson Key	S240	S240
Salesperson Name	Jack Smith	Jack SMith
Old Territory		Goa
Current Territory	Goa	Kerala
Effective Date	26-11-1995	15-12-2002
Region	South	South

Method for applying Type 3 change:-

- Add an old field for affected attribute
- Push existing value from current field to old field
- Keep new value in current field
- May add a current effective date field
- Key of row is not affected
- No new row is added
- Existing queries switch to current value
- Queries requiring old value must be revised
- But for one soft change at a time
- If there is a succession of changes, more sophisticated techniques must be devised.

Q4. Explain the OLAP operations for Retail stores with the help of a diagram.

ANS:

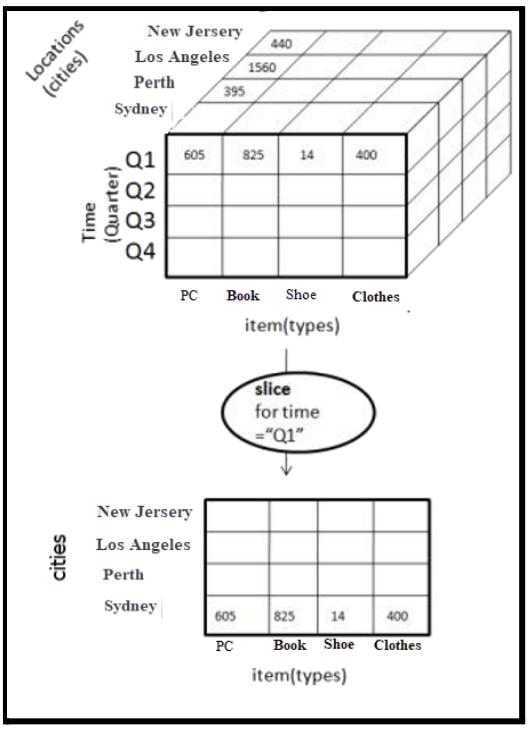


Data cube for retail store

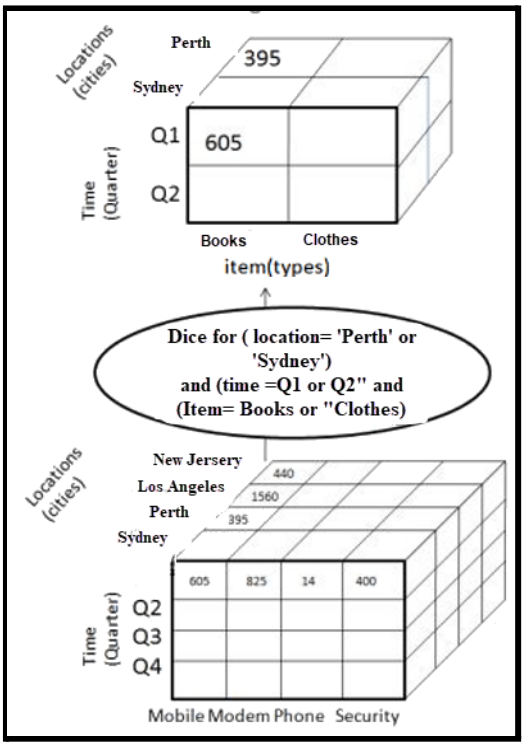
1) Perth and Sydney in Australia and New Jersey and Los Angeles in the US are territories of the retail store.
2) PC, Book, Shoe and Clothes are the product categories for the retail store
3) Q1, Q2, Q3, Q4 are quarters of the year

OLAP Operations:

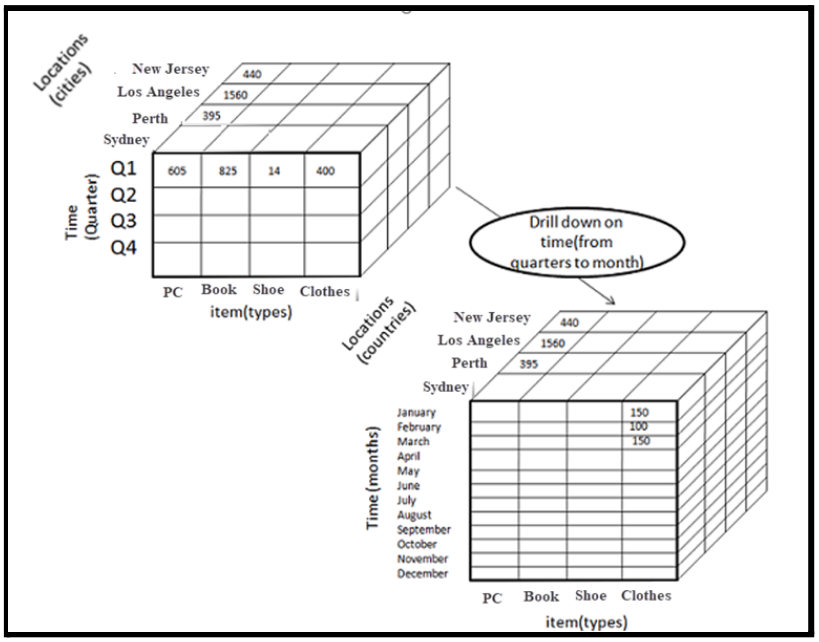
1) **SLICE:-** Creating a subset of multidimensional array corresponding to a single value for one or more members of dimensions not in subset, Consider a slice for time = 'Q1'



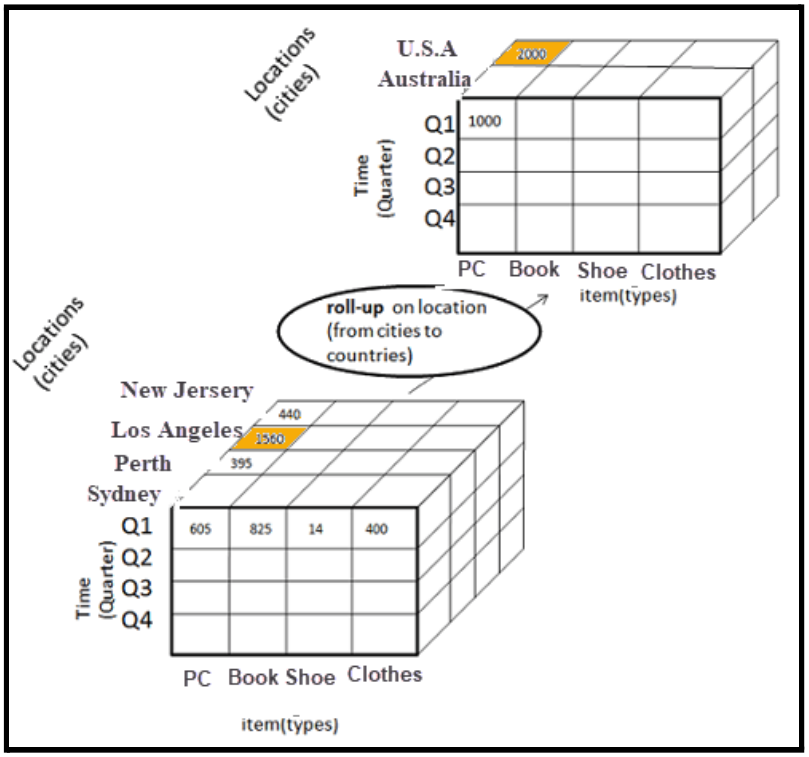
2) **DICE:-** Creates a subcube, by applying slice on more than 2 dimensions of data cube. Consider, dice for (location = 'Perth' or 'Sydney') and (time = 'Q1 or Q2') and (item = 'Books' or 'Clothes')



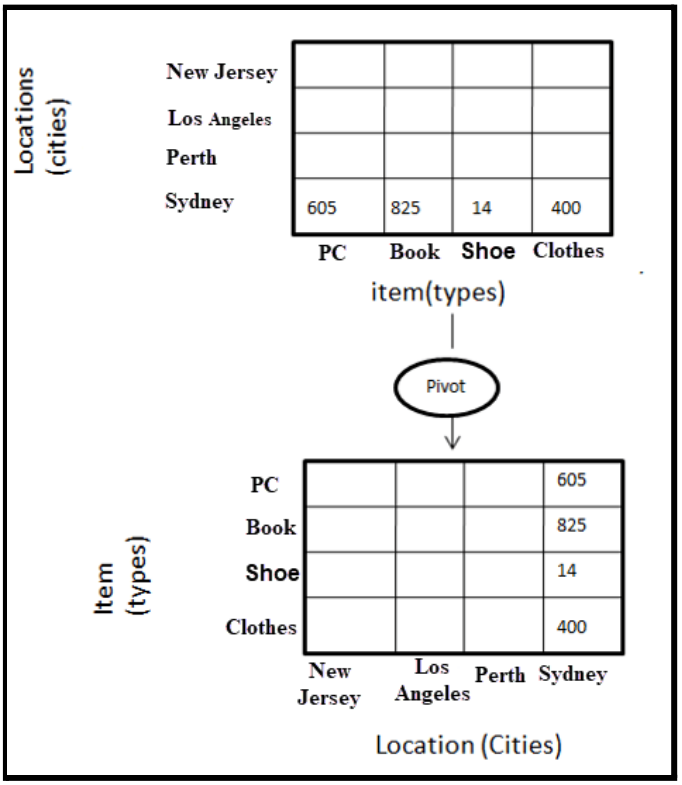
3) **DRILL DOWN:-** Less detailed data is converted to highly detailed data. Moving down in the concept hierarchy. Drill down on time (Quarters to Months)



4) **ROLL UP:-** More detailed converted to less detailed. Aggregation of cube by moving up in concept hierarchy. Consider roll up on locations (from cities to countries)



5) **PIVOT :-** To change the dimensional orientation of a report or page display. Pivot on slice (time = 'Q1')



Q5. Differentiate between

i. Data warehouse versus Data Marts

Data Warehouse	**Data Marts**
Data warehouse is a centralized system (Union of data marts)	Data Mart is a decentralized system (Single business process)
Lightly denormalization takes place here	Highly denormalization takes place here
Structured for corporate view of data	Structured for departmental view of data
Building a warehouse is difficult and time consuming	Building a data mart is easy and take less time
It is flexible and has a long life span	It is not flexible and has short life than Data Warehouse
Data oriented in nature	Project oriented in nature
Size is vast	Size is smaller than data warehouse
Data is in detailed form	Data is in summarized form

ii. Data warehouse versus Data Lake

Data Warehouse	**Data Lake**
Inputs are structured processed data	Inputs are structures, unstructured and raw data
Data needs to be cleaned and processed before storing	Data is not necessarily cleaned and processed
Schema-on-write (Schema cleaned before writing in database)	Schema-on-read (Schema created only while reading the data)
Fixed configuration	Flexible configuration
Cost scales with volume	Low costs, high volume
Used by business users for reports	Used by data scientists to form models
Not so efficient	Efficient handling of unstructured data
Updated periodically	Rapidly updated
E.g: Supermarket Data warehouse	E.g: Supermarket DL with customer sentiments, advertising result (unstructured data)

iii. Top-down versus Bottom-up approach

Top Down	**Bottom Up**
Data Marts are created from Data Warehouse	Data Warehouse is created from Data Marts
Inherently architected	Inherently incremental
Single central storage of information about the content	Departmental information stored
Centralized rules and control	Departmental rules and control
Contains redundant information	Redundancy can be removed
May see quick results if implemented with repetitions	Less risk of failure, favorable return on investment and proof of techniques.