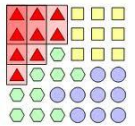
Data Warehousing and Data Mining

Unit 3-4

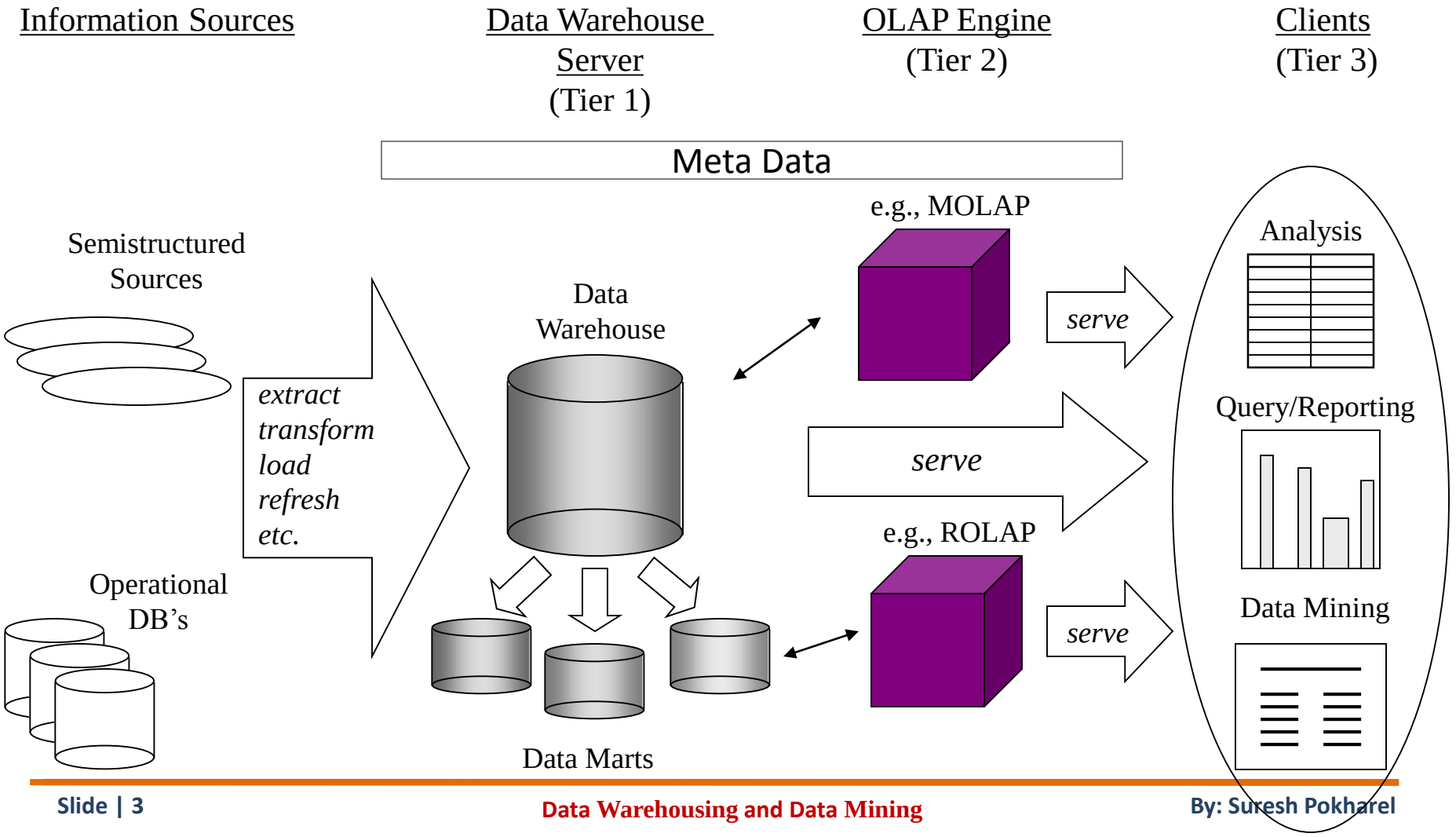
Instructor
Suresh Pokharel

Content

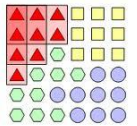


- Data Warehouse
- Architecture
- Distributed and Virtual Data Warehouse,
- Data Warehouse Manager
- OLTP
- Types of OLAP:OLAP,MOLAP, HOLAP

Data Warehouse Architecture

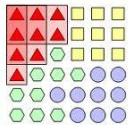


Virtual Data Warehouse



- Create Virtual view of databases as opposed to a physical warehouse.
- Create single "virtual database" from all the data resources.
- Data resource can be local or remote
- Data is not move from sources but users are given direct access to the data
- Access to data may be : simple SQL queries, view definition, data-access middleware
- Possible to access remote sources : major RDBMS
- Client application access data distributed across multiple data sources through a single SQL statement, a single interface
- All data sources are accessed as though they are local users and their applications don't even need to know the physical location of the data

Virtual Data Warehouse



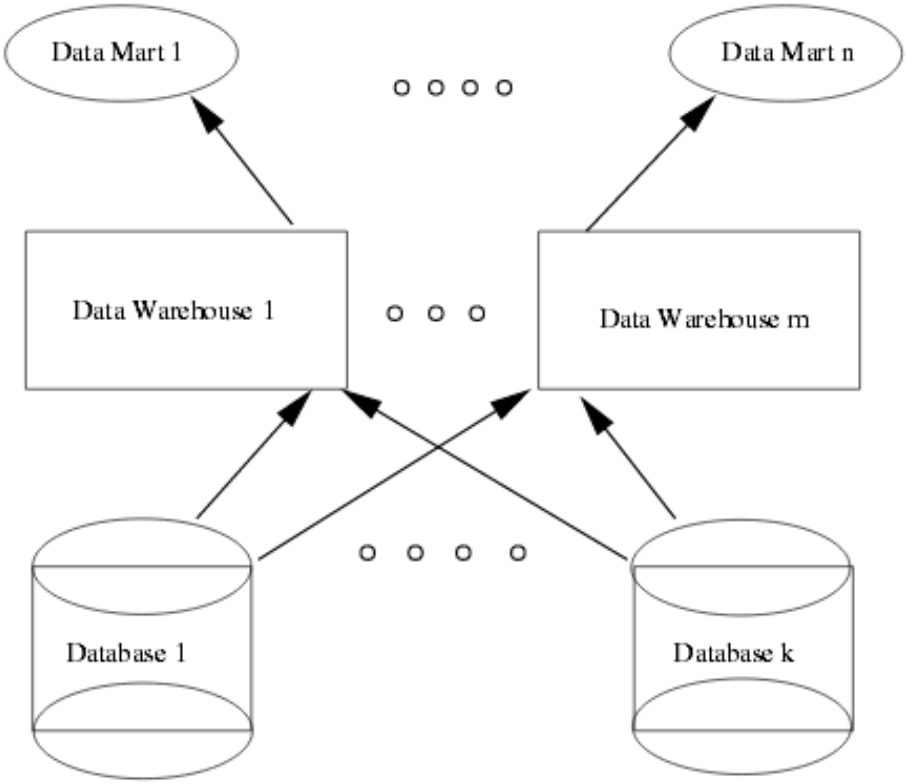
Advantages

- Don’t need to replicate information
- Don’t need to maintain separate physical storage
- Initially easy and fast

Disadvantages

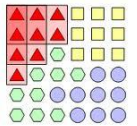
- Live performance may be degraded as some resources will be used for analytical purpose
- Since there is no metadata, no summary data, history, all query must be repeated, creating an additional burden on the system
- No clearing and refreshing process : queries became complex

Distributed Data Warehouse



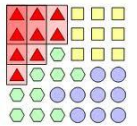
Distributed Data Warehouse Architecture

Distributed Data Warehouse



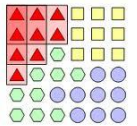
- A Distributed Data Warehouse is similar in most respects to a Central Data Warehouse, except that the data is distributed to separate mini-Data Warehouses (Data Marts) on local or specialized servers
- Business is distributed geographically or over multiple, differing product lines.
- The data warehouse environment will hold a lot of data, and the volume of data will be distributed over multiple processors.

Distributed Data Warehouse



- The data copied into a data warehouse does not change (except to correct errors). The data warehouse is a historical record of the state of an organization. The frequent changes of the source OLTP systems are reflected in the data warehouse by adding new data, not by changing existing data.
- Data warehouses are subject oriented, that is, they focus on measuring entities, such as sales, inventory, and quality. OLTP systems, by contrast, are function oriented and focus on operations such as order fulfillment.
- In data warehouses, data from distinct function-oriented systems is integrated to provide a single view of an operational entity.
- Data warehouses are designed for business users, not database programmers, so they are easy to understand and query.

Approaches to OLAP Servers



OLAP servers

(1) Relational OLAP (ROLAP)

- Relational and specialized relational DBMS to store and manage warehouse data
- OLAP middleware to support missing pieces
- have greater scalability

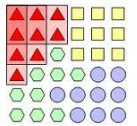
(2) Multidimensional OLAP (MOLAP)

- Array-based storage structures
- Direct access to array data structures
- Fast indexing to pre-computed summarized data

(3) Hybrid OLAP (HOLAP) server

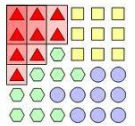
- Combine both ROLAP and MOLAP
- E.g. Microsoft SQL Server 2000

Warehouse Manager



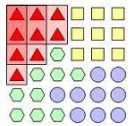
- Responsible for the warehouse management process
- Knowledge about third party system software, C programs, and shell scripts
- Task varies over the size and complexity of warehouse

Warehouse Manager: Tasks



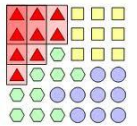
- The controlling process
- Stored procedures or C with SQL
- Backup/Recovery tool
- SQL Scripts

Warehouse Manager: Operations



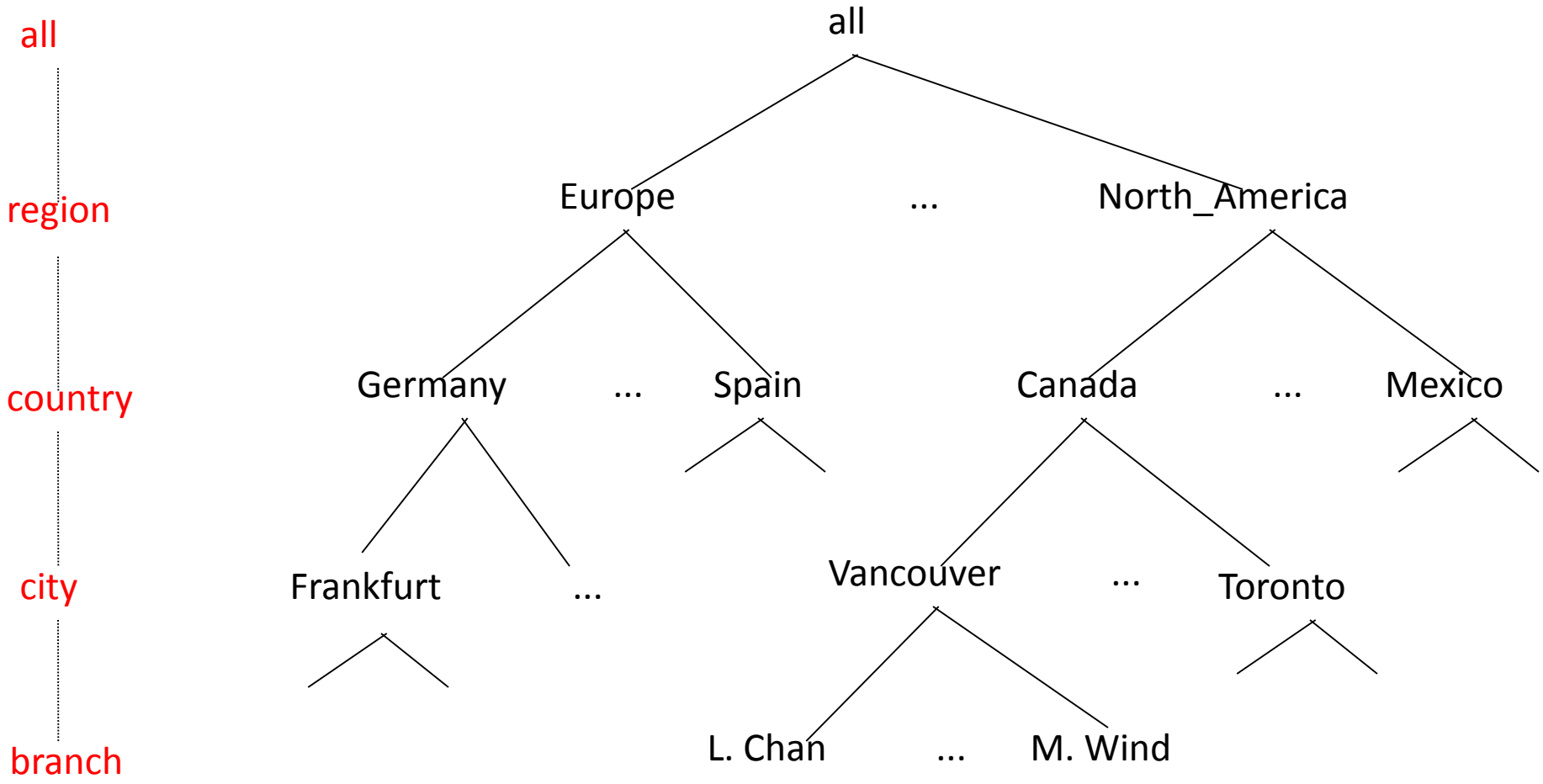
- A warehouse manager analyzes the data to perform consistency and referential integrity checks.
- Creates indexes, business views, partition views against the base data.
- Generates new aggregations and updates existing aggregations. Generates normalizations.
- Transforms and merges the source data into the published data warehouse.
- Backup the data in the data warehouse.
- Archives the data that has reached the end of its captured life.
- A warehouse Manager also analyzes query profiles to determine index and aggregations are appropriate.

DATA WAREHOUSING - RELATIONAL OLAP



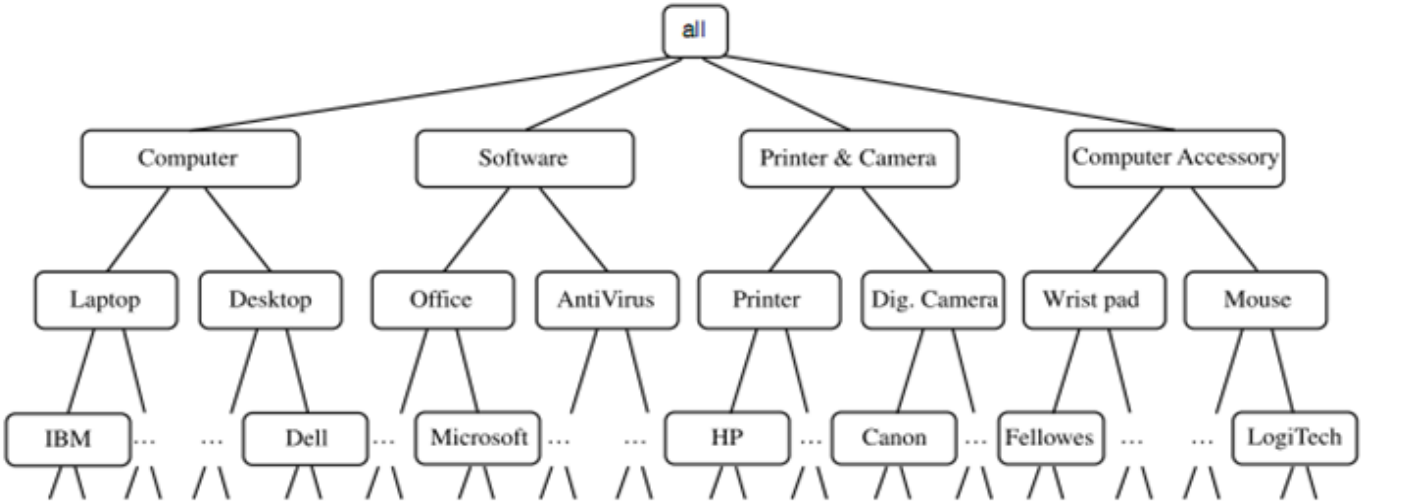
- A warehouse manager analyzes the data to perform consistency and referential integrity checks.
- Creates indexes, business views, partition views against the base data.
- Generates new aggregations and updates existing aggregations. Generates normalizations.
- Transforms and merges the source data into the published data warehouse.
- Backup the data in the data warehouse.
- Archives the data that has reached the end of its captured life.
- A warehouse Manager also analyzes query profiles to determine index and aggregations are appropriate.

A Concept Hierarchy: Dimension (location)

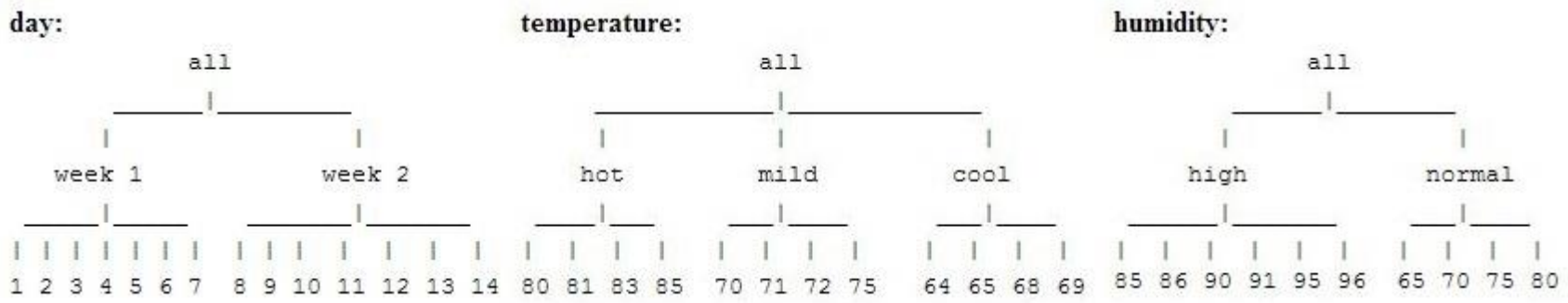


A Concept Hierarchy

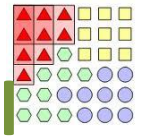
TID	Items Purchased
T100	IBM-ThinkPad-T40/2373, HP-Photosmart-7660
T200	Microsoft-Office-Professional-2003, Microsoft-Plus!-Digital-Media
T300	Logitech-MX700-Cordless-Mouse, Fellowes-Wrist-Rest
T400	Dell-Dimension-XPS, Canon-PowerShot-S400
T500	IBM-ThinkPad-R40/P4M, Symantec-Norton-Antivirus-2003
...	...



Concept of Hierarchy



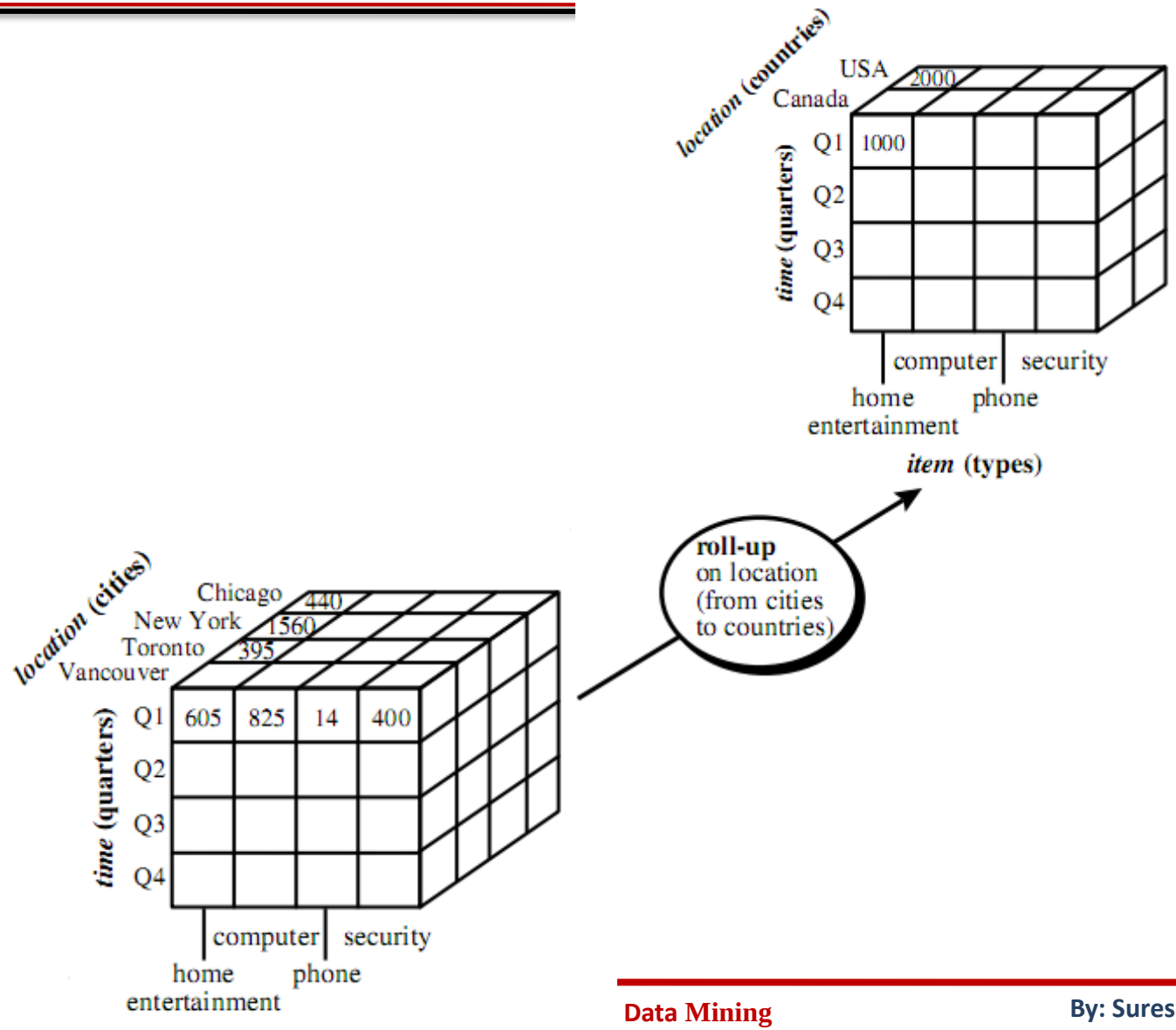
OLAP Operations in Multidimensional Data Model



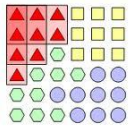
- Roll-up :
- Drill- down :
- Slice and dice :
- Pivot (rotate) :

Reference: http://athena.ecs.csus.edu/~olap/olap/OLAPoperations.php

OLAP Operations : roll-up



OLAP Operations : roll-up



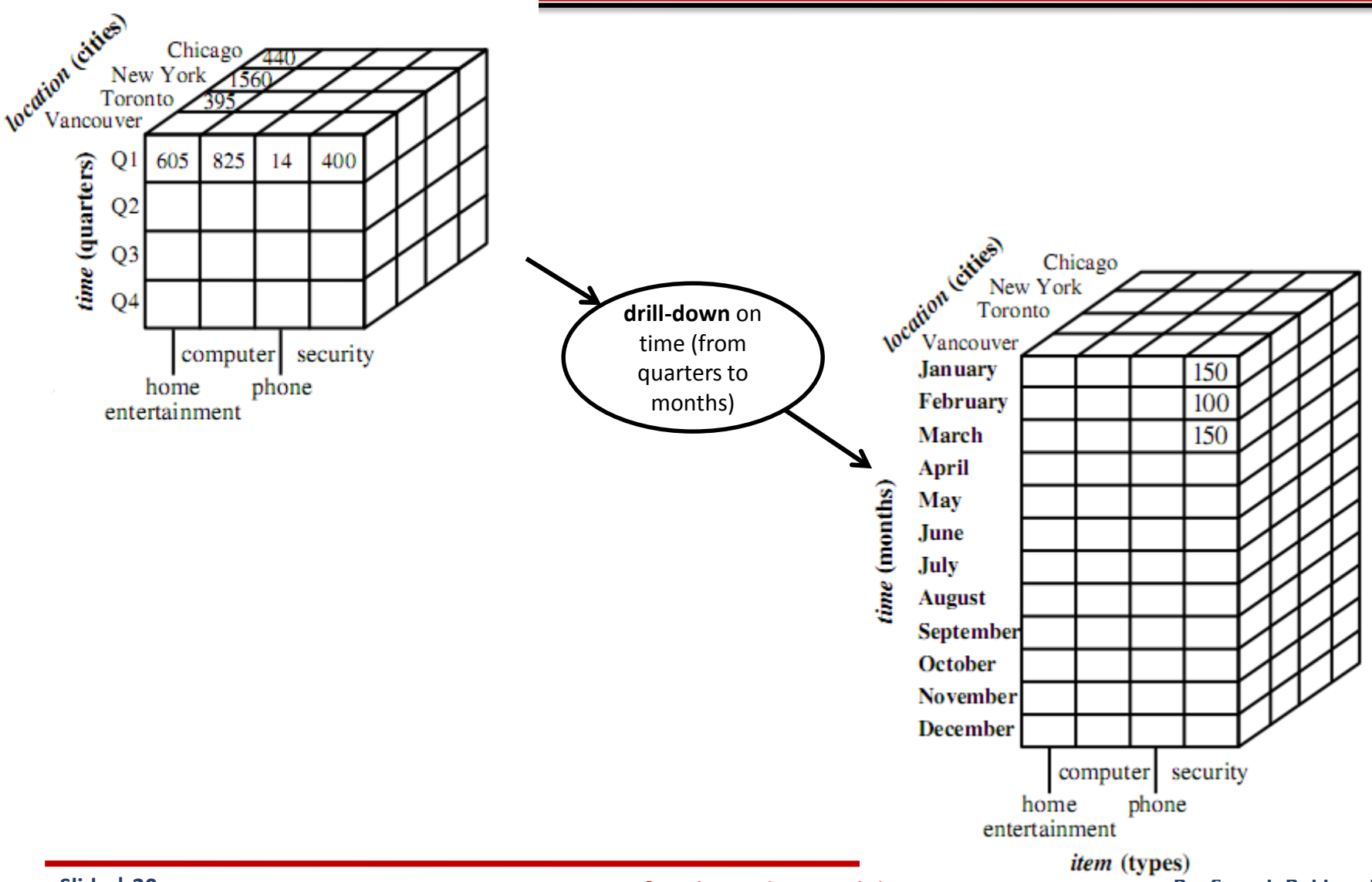
temperature	64	65	68	69	70	71	72	75	80	81	83	85
week 1	1	0	1	0	1	0	0	0	0	0	1	0
week 2	0	0	0	1	0	0	1	2	0	1	0	0

set up levels (hot(80-85), mild(70-75), cold(64-69)) in temperature

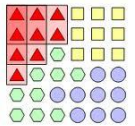
temperature	cool	mild	hot
week 1	2	1	1
week 2	1	3	1

hot-->day-->week

OLAP Operations : drill-down

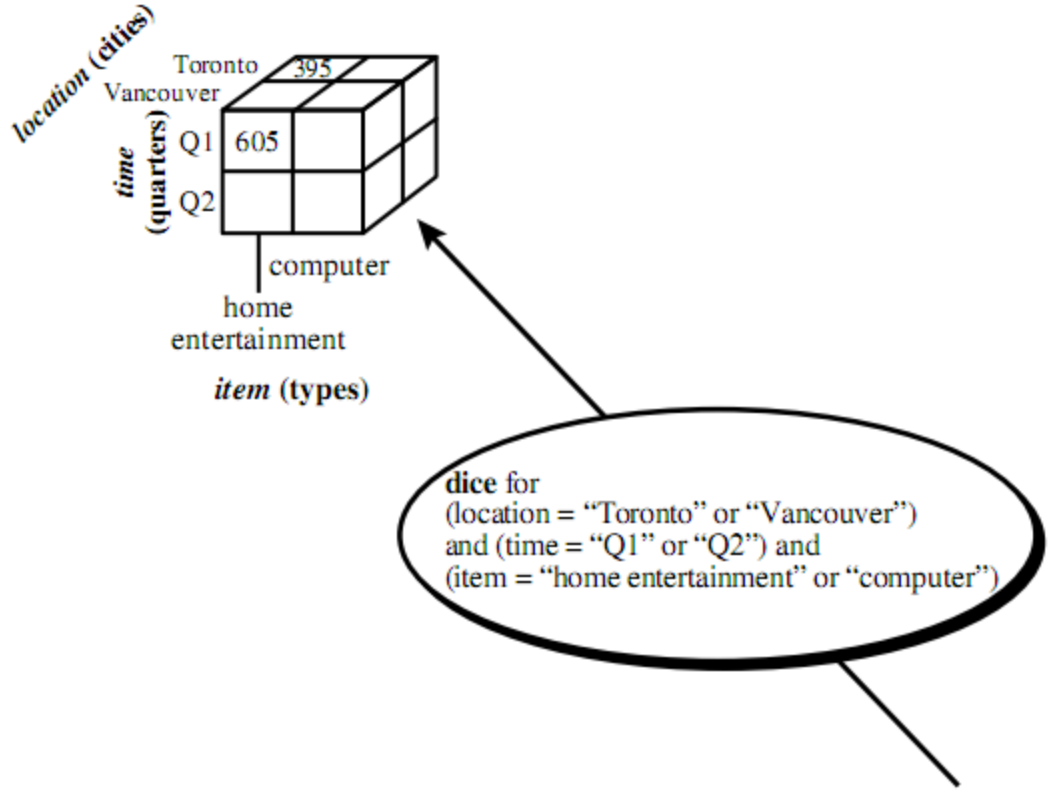


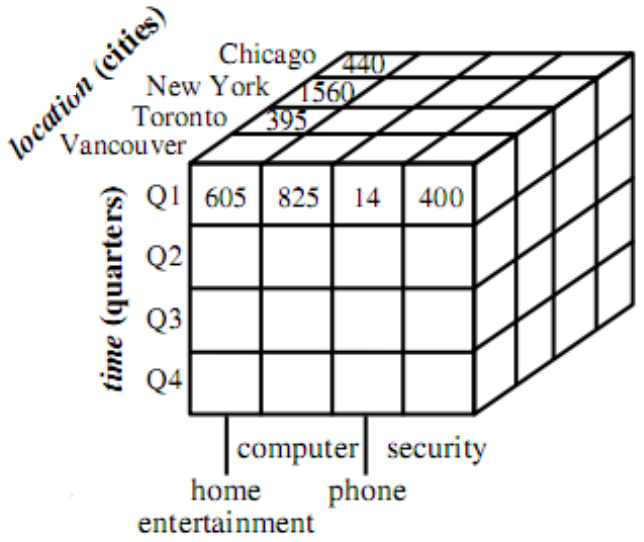
OLAP Operations : drill-down



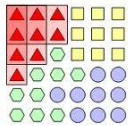
	cool	mild	hot
day 1	0	0	0
day 2	0	0	0
day 3	0	0	1
day 4	0	1	0
day 5	1	0	0
day 6	0	0	0
day 7	1	0	0
day 8	0	0	0
day 9	1	0	0
day 10	0	1	0
day 11	0	1	0
day 12	0	1	0
day 13	0	0	1
day 14	0	0	0

OLAP Operations : dice



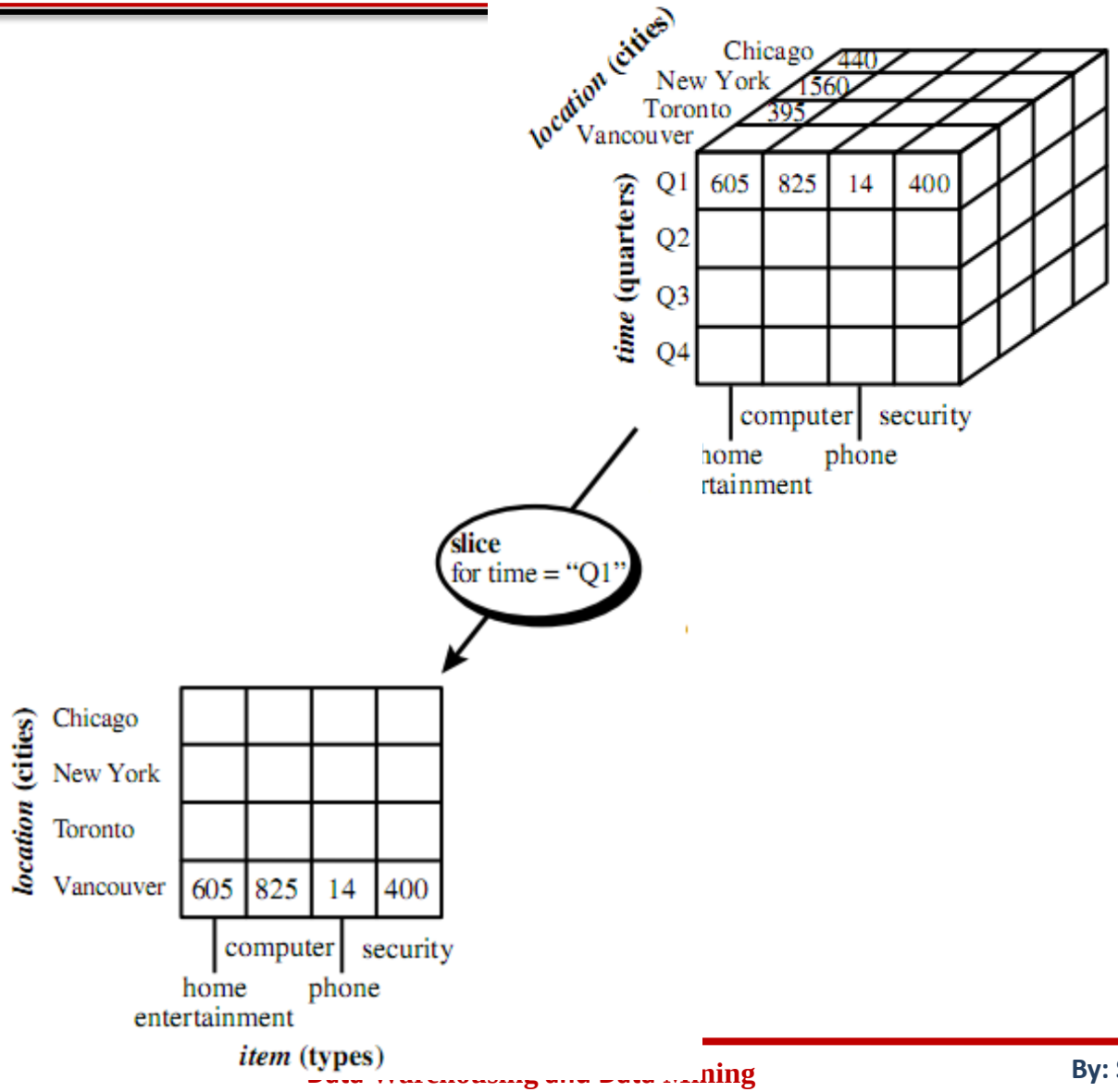


OLAP Operations : dice

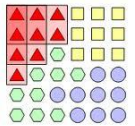


	cool	hot
day 3	0	1
day 4	0	0

OLAP Operations : slice

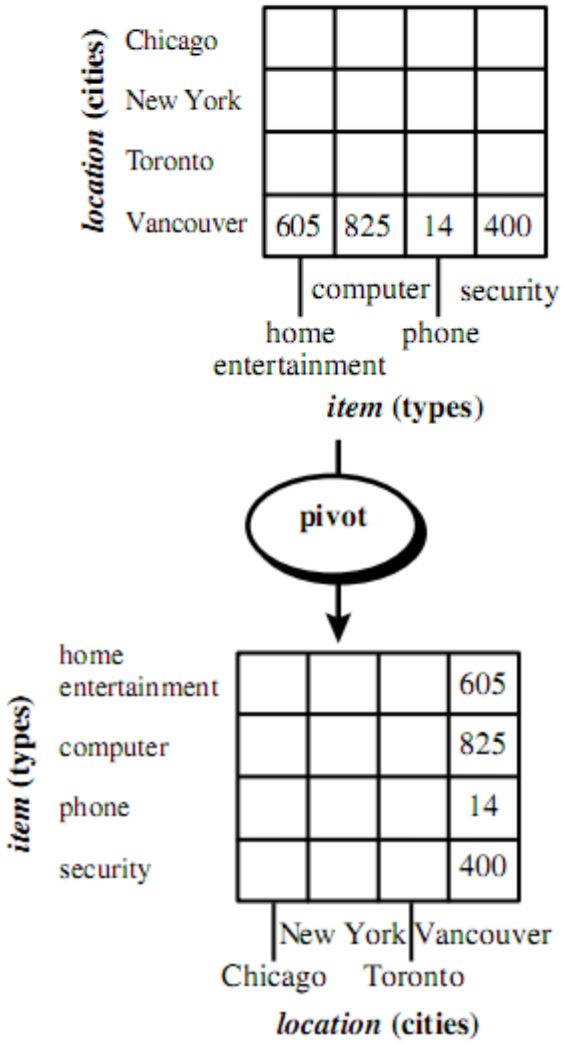


OLAP Operations : slice

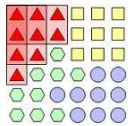


	cool
day 1	0
day 2	0
day 3	0
day 4	0
day 5	1
day 6	0
day 7	1
day 8	0
day 9	1
day 10	0
day 11	0
day 12	0
day 13	0
day 14	0

OLAP Operations : pivot



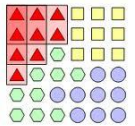
Data Warehouse Background Process



Tasks:

- **Data Extraction** : gather data from multiple sources [Production data, legacy data, external system, metadata]
- **Data Cleaning** : remove noise [age, DOB]
- **Data Transformation** : convert data format [Currency]
- **Load** : sorts, summarize, consolidate, integrated [Batch loading, sequential loading, incremental loading]
- **Refresh**: updates from data sources to the warehouse [data source update => update data warehouse]

Functions of Data Warehouse Tools and Utilities

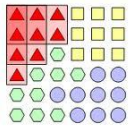


The following are the functions of data warehouse tools and utilities:

- Data Extraction - Involves gathering data from multiple heterogeneous sources.
- Data Cleaning - Involves finding and correcting the errors in data.
- Data Transformation - Involves converting the data from legacy format to warehouse
- format.
- Data Loading - Involves sorting, summarizing, consolidating, checking integrity, and
- building indices and partitions.
- Refreshing - Involves updating from data sources to warehouse.

Ref: https://www.tutorialspoint.com/dwh/pdf/dwh_quick_guide.pdf

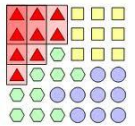
Extract Transform and Load (ETL) Tools



Some popular ETL tools are :

- Informatica - Power Center
- IBM - Websphere DataStage(Formerly known as Ascential DataStage)
- SAP - BusinessObjects Data Integrator
- IBM - Cognos Data Manager (Formerly known as Cognos DecisionStream)
- Microsoft - SQL Server Integration Services
- Oracle - Data Integrator (Formerly known as Sunopsis Data Conductor)
- SAS - Data Integration Studio
- Oracle - Warehouse Builder
- AB Initio
- Information Builders - Data Migrator
- Pentaho - Pentaho Data Integration
- Embarcadero Technologies - DT/Studio
- IKAN - ETL4ALL
- IBM - DB2 Warehouse Edition
- Pervasive - Data Integrator
- ETL Solutions Ltd. - Transformation Manager

DATA WAREHOUSING - TUNING



Difficulties in Data Warehouse Tuning:

Tuning a data warehouse is a difficult procedure due to following reasons:

- Data warehouse never remains constant.
- It is very difficult to predict what query the user is going to post in the future.
- Business requirements change with time.
- Users and their profiles keep changing.
- The user can switch from one group to another.
- The data load on the warehouse also changes with time.

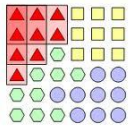
Performance Assessment

Here is a list of objective measures of performance:

- Average query response time
- Scan rates
- Time used per day query
- Memory usage per process
- I/O throughput rates

Ref: Tutorialpoint

DATA WAREHOUSING - TUNING



Data Load Tuning

If there is a delay in transferring the data, or in arrival of data then the entire system is affected badly. Therefore it is very important to tune the data load first.

Integrity Checks

Integrity checking highly affects the performance of the load. Following are the points to remember:

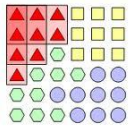
- Integrity checks need to be limited because they require heavy processing power.
- Integrity checks should be applied on the source system to avoid performance degrade of data load.

Tuning Queries

Two kinds of queries in data warehouse:

- Fixed queries (regular reports, Common aggregations): It is good to store the most
- successful execution plan while testing fixed queries
- Ad hoc queries (To understand ad hoc queries, it is important to know the ad hoc users of the data warehouse)

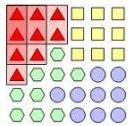
DATA WAREHOUSING - Testing



Three basic levels of testing

- Unit testing
- Integration testing
- System testing

DATA WAREHOUSING - Testing



Unit Testing

- In unit testing, each component is separately tested.
- Each module, i.e., procedure, program, SQL Script, Unix shell is tested.
- This test is performed by the developer.

Integration Testing

- In integration testing, the various modules of the application are brought together and then tested against the number of inputs.
- It is performed to test whether the various components do well after integration.

System testing

- In system testing, the whole data warehouse application is tested together.
- The purpose of system testing is to check whether the entire system works correctly together or not.
- System testing is performed by the testing team.