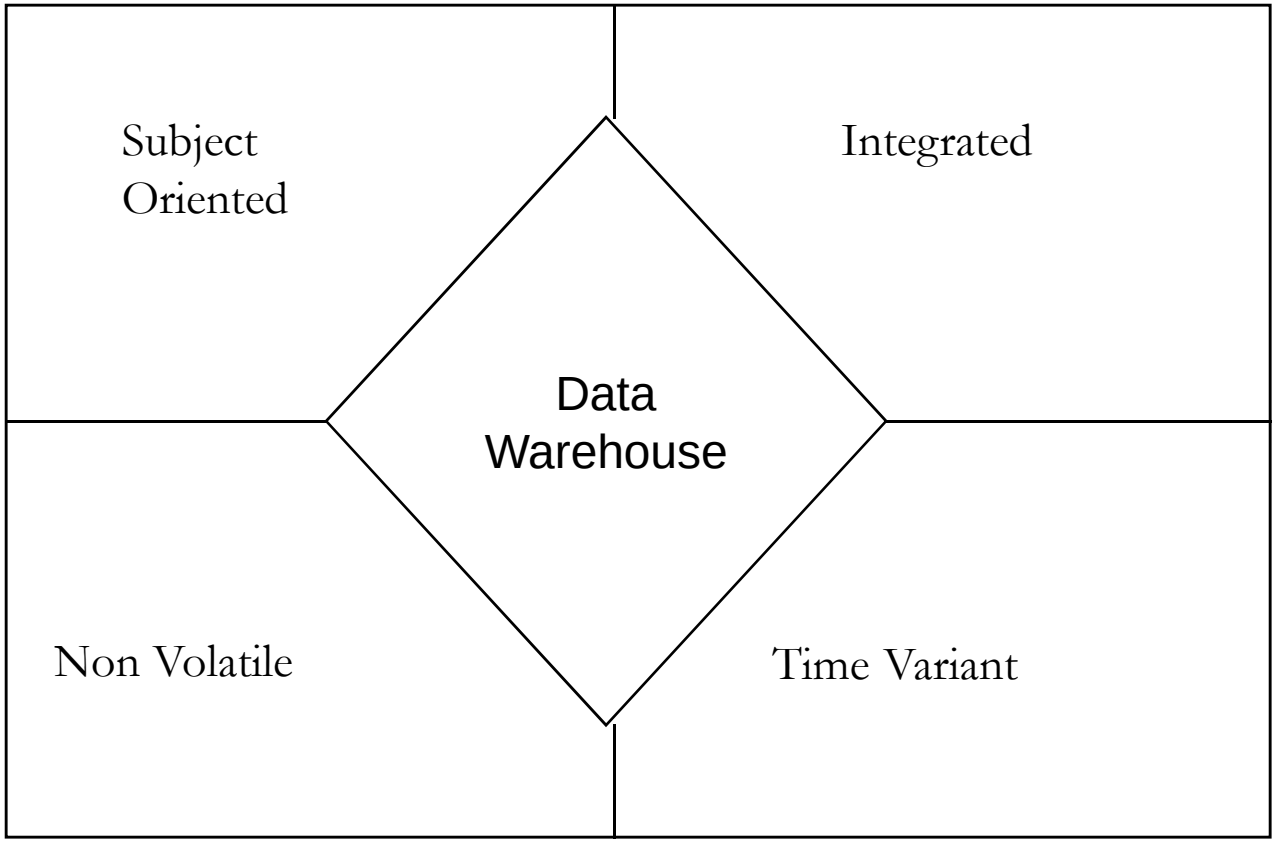
Definition of a Data Warehouse

" A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process"

William H Inmon

Data Warehouse Properties



- **Subject –oriented** :- a data warehouse is organized around major subjects focuses on the modeling and analysis of data for decision makers.
- Provide simple and concise view around particular subject.

- **Integrated:-** A data warehouse is usually constructed, by integrating multiple heterogeneous sources, such as relational db, flat files, and on-line transaction records.
- Data cleaning and data integration techniques are applied

- **Time variant :-**data are stored to provide info. From a historical prospective .
- Every key structure in the data warehouse contain, either implicitly or explicitly an element of time.

- **Non-volatile:-** A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.
- Due to this separation , a data warehouse does not require transaction processing , recovery, and concurrency control mechanism.
- It usually requires only two operations in data accessing , initial loading and access of data.

How organizations using info from data warehouse

- Use this info to support business decision making activities including

1.customer focus, which include the analysis of customer buying patterns

2. repositioning products and managing product portfolios.

3. analyzing operations and looking for sources of profit

4. managing customer relationships, making environment corrections and managing the cost of corporate assets.

operational db/systems and data warehouse

- The major task of on-line transaction database system is to perform on-line transaction and query processing.
- These system are called on-line transaction processing (OLTP) system.
- They cover the most of day to day operations of an orga. Such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

- Data ware house system on the other hand serve users or knowledge workers in the role of data analysis and decision making.
- Such system can organize and present data in various formats in order to accommodate the diverse need of the different users.
- These system are known as on-line analytical processing system(OLAP)

Difference in features of OLTP/OLAP

- Users and system orientation
- Data content
- Database design
- View
- Access patterns:

Difference in features of OLTP/OLAP

- Users and system orientation:-
  - an OLTP is customer oriented and is used for transaction and query processing by clerk ,clients, and info. Technology professionals.
  - And OLAP is market oriented and is used for data analysis by knowledge workers, including managers, executives and analysts.

Data content:- an OLTP system manages current data that, are too detailed to be easily used for decision making.

An OLAP system manage large amounts of historical data, provide facilities for summarization and aggregation and store and manage info. at different levels of granularity.

So it make the data easier to use in informed decision making.

Database design :- An OLTP system usually adopt an entity-relationship data model and application oriented database design.

An OLAP system adopt either star or snowflake model and subject oriented database design.

- View :- an OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organization.
- In contrast an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization.
- OLAP system also deal with info. That originates from different org. Integrating info. From many data stores.

- Access patterns: the access patterns of OLTP system consist of many short, atomic transaction.
- Such a system require concurrency control and recovery mechanism In most cases , access to OLAP system are mostly read only operations(because most data warehouse store historical rather than up to date info. Although many could be complex queries.

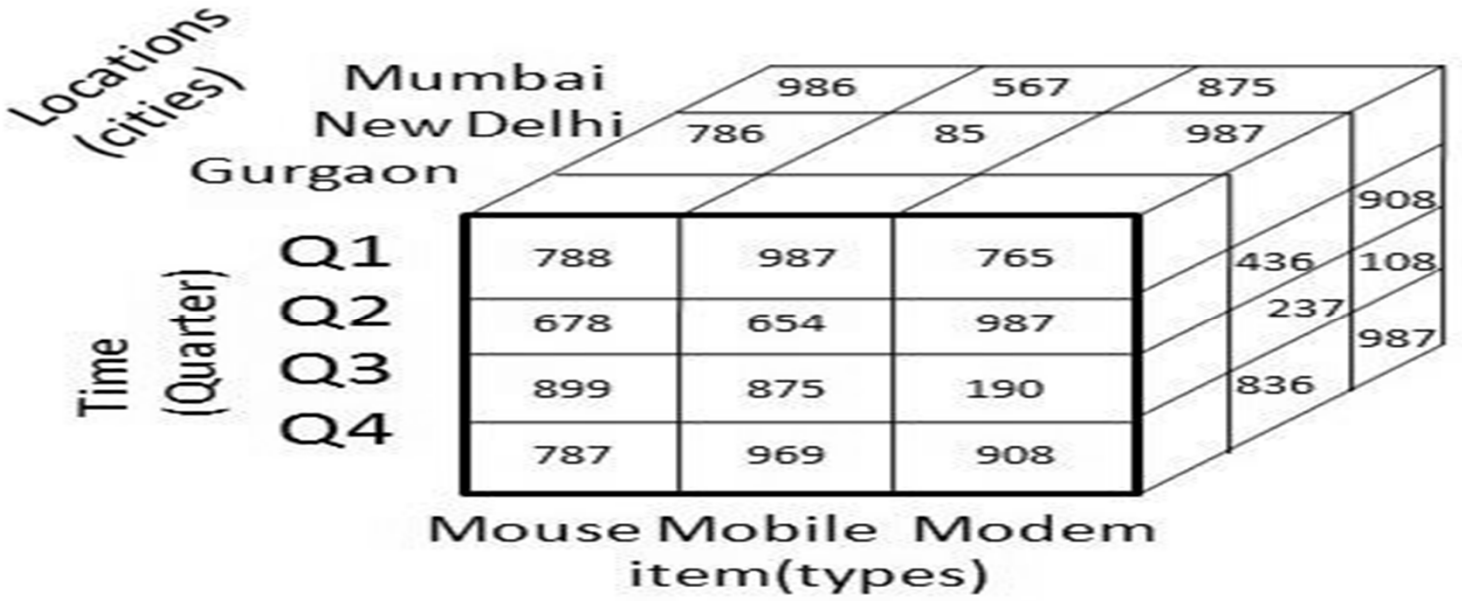
Features	OLTP	OLAP
Characteristic	Operational processing	Informational processing
Orientation	Transaction	Analysis
Users	Clerk,DBA,database professional	Knowledge workers, (manager,executive…)
Function	Day-to-day operations	Long-term info
Db design	ER based, application oriented	Star/snowflake, object oriented
Data	current., granteed upto date	Historical
Summarization	Primitive, highly detailed	Summarized
View	Detailed	Summarized
Unit of work	Short, simple transaction	Complex query
Access	Red/write	Mostly read
Focus	Data in	Information out

Operation	Index/ hash on primary key	Lot of scan
No. of records accessed	Tens	Millions
No of users	Thousands	Hundered
DB size	100MB to GB	100 GB to TB
Priority	High performance, high availability	High flexibility, end user autonomy
Metric	Transaction throughput	Query throughput, response time

Multidimensional data model

- Data warehouses and OLAP tools are based on a multi dimensional data model.
- This model views data in the form of data cube.
- A data cube allow data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

Time	Location="Gurgaon"			Location="New Delhi"			Location="Mumbai"		
	Item			Item			Item		
	Mouse	Mobile	Modem	Mouse	Mobile	Modem	Mouse	Mobile	Modem
Q1	788	987	765	786	85	987	986	567	875
Q2	678	654	987	659	786	436	980	876	908
Q3	899	875	190	983	909	237	987	100	1089
Q4	787	969	908	537	567	836	837	926	987



Stars, snowflakes, and constellations

- The entity relationship data model is commonly used in the design of relational db. where a database schema consists of a set of entities and the relationships between them.
- A data warehouse require a concise, subject –oriented schema that facilitates on line data analysis

- Most popular data model for a data warehouse is a multidimensional model.
- Such models can exist in the form of
  - star schema
  - snowflake schema
  - fact constellation schema

Star schema

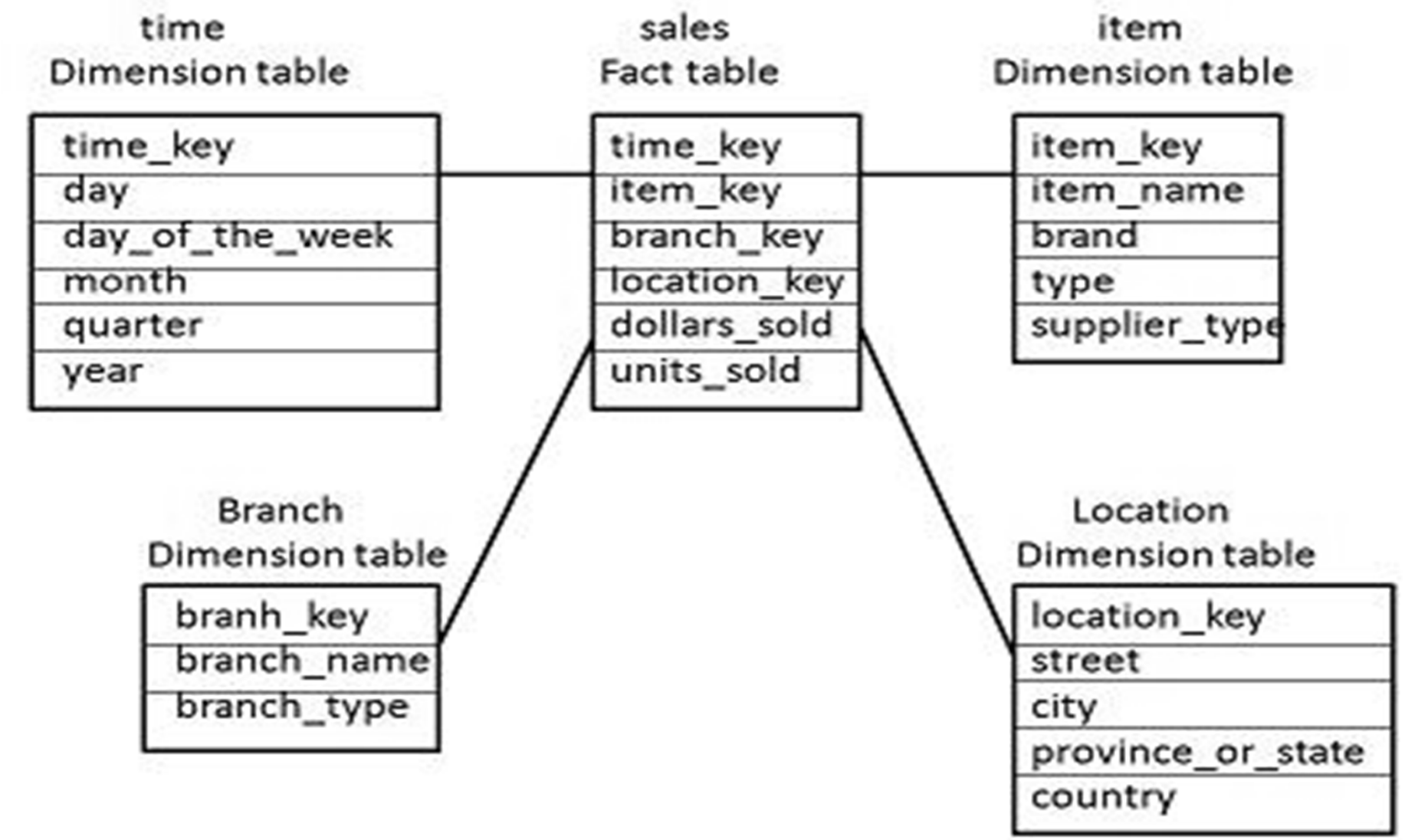
- The most common modeling paradigm is the star schema , in which the data warehouse contain

1. large central table (fact table) containing the bulk of data with no redundancy .

2. a set of smaller attendant tables(dimension tables) one for each dimension.

The schema graph resembles a starburst , with the dimension tables displayed in a radial pattern around the central fact table.

Star schema of d/w for sales



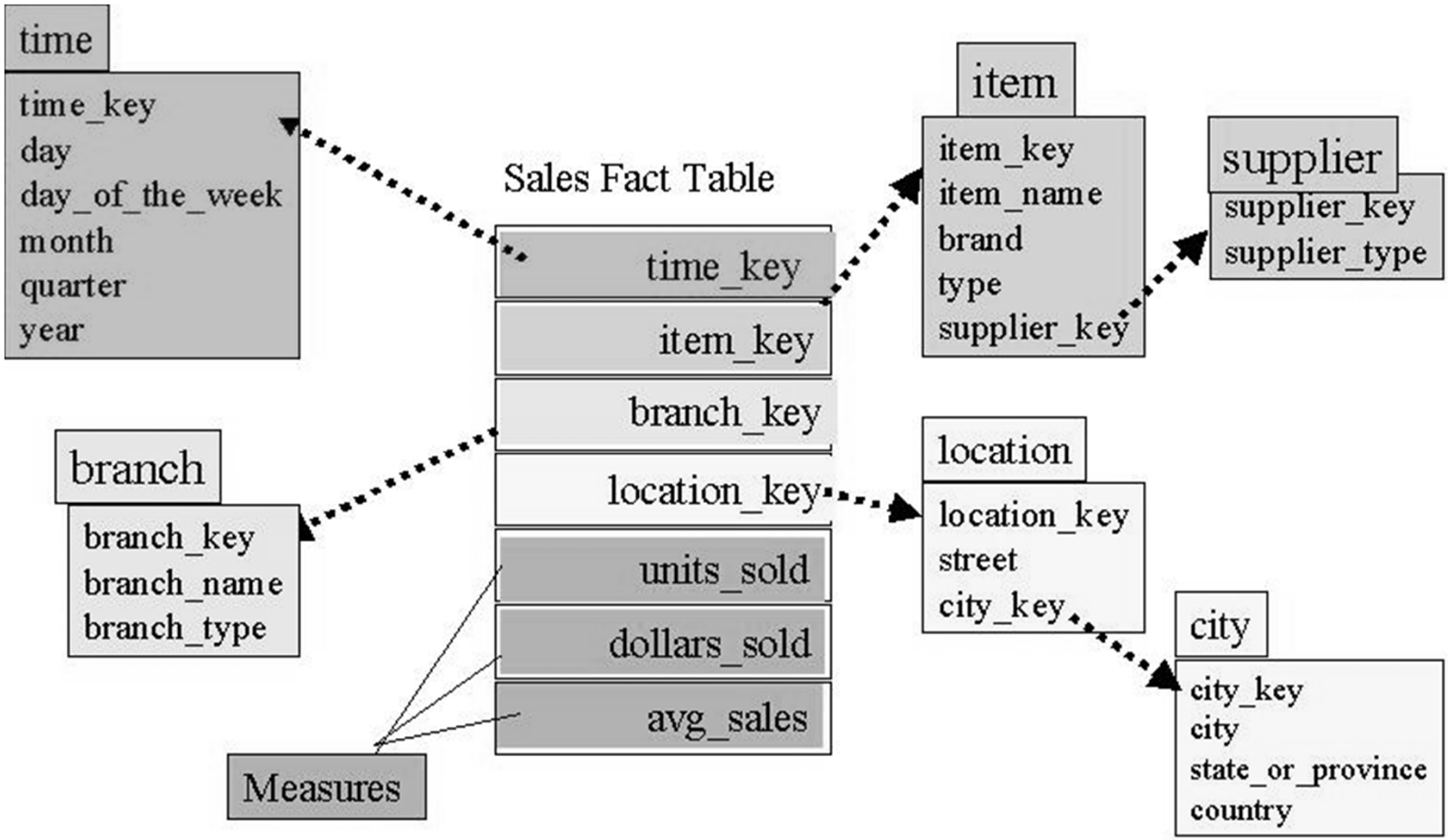
Snowflake schema

- The snowflake schema is a variant of star schema model, where some dimensions tables are normalized, thereby further spitting the data into additional tables.
- The resulting schema graph forms a shape similar to a snowflake.
- The major difference between the snowflake and star schema models is that the dimensions tables of snowflake

- Model may be kept in normalized form to reduce redundancies. Such table is easy to maintain and save storage space.
- The snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. Consequently the system performance may be adversely impacted.

- Snowflake schema reduce redundancy, it is not as popular star schema in data warehouse design.

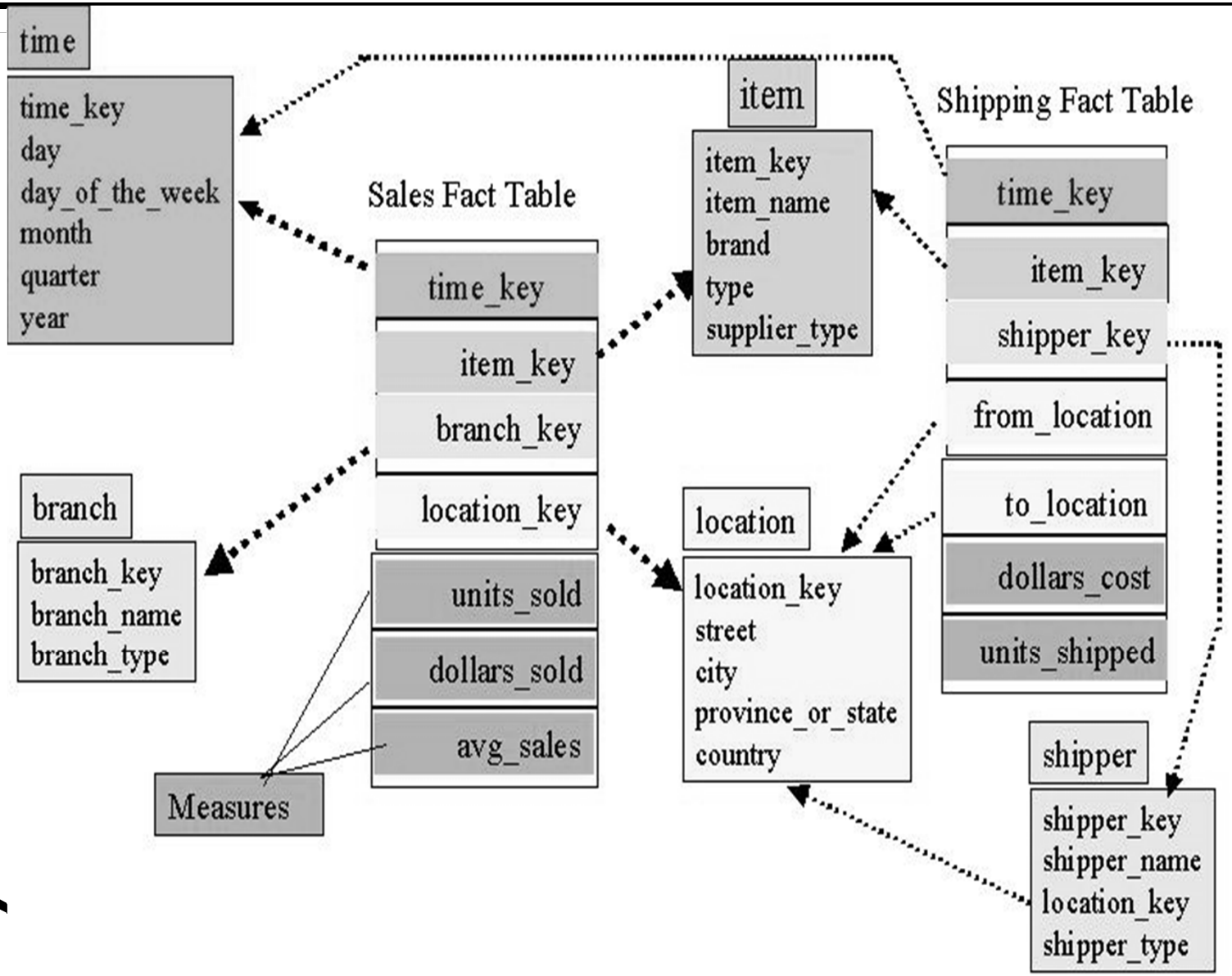
Snowflake schema of d/w for sales



Fact constellation

- Collection of stars ,also known as galaxy schema or fact constellation.
- This schema specifies two fact tables , sales and shipping.
- A fact constellation schema allows dimension tables to be shared between fact tables.

- For eg: the dimension tables for time, item and location are shared b/t both sales and shipping fact tables.


time
time_key
day
day_of_the_week
month
quarter
year
Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales
item
item_key
item_name
brand
type
supplier_type
Shipping Fact Table
time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped
branch
branch_key
branch_name
branch_type
location
location_key
street
city
province_or_state
country
shipper
shipper_key
shipper_name
location_key
shipper_type
Measures

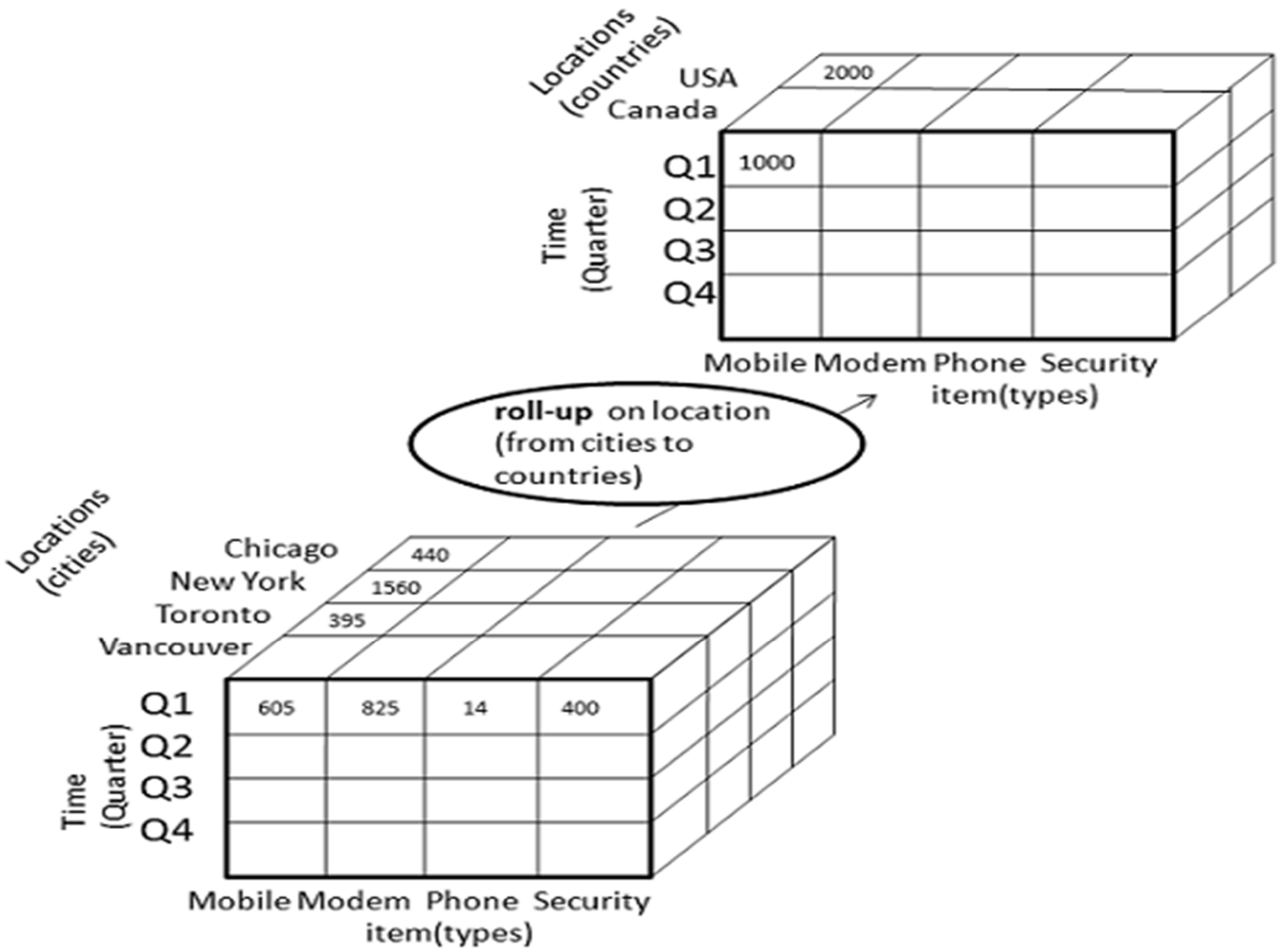
ology (BCA)

OLAP operations in the multidimensional data model

- A no. of OLAP operations allowing the users interactive querying and analysis of data at hand.
  - Roll-up
  - Drill –down
  - Slice and dice
  - drill-across
  - Dill-through

Roll-up

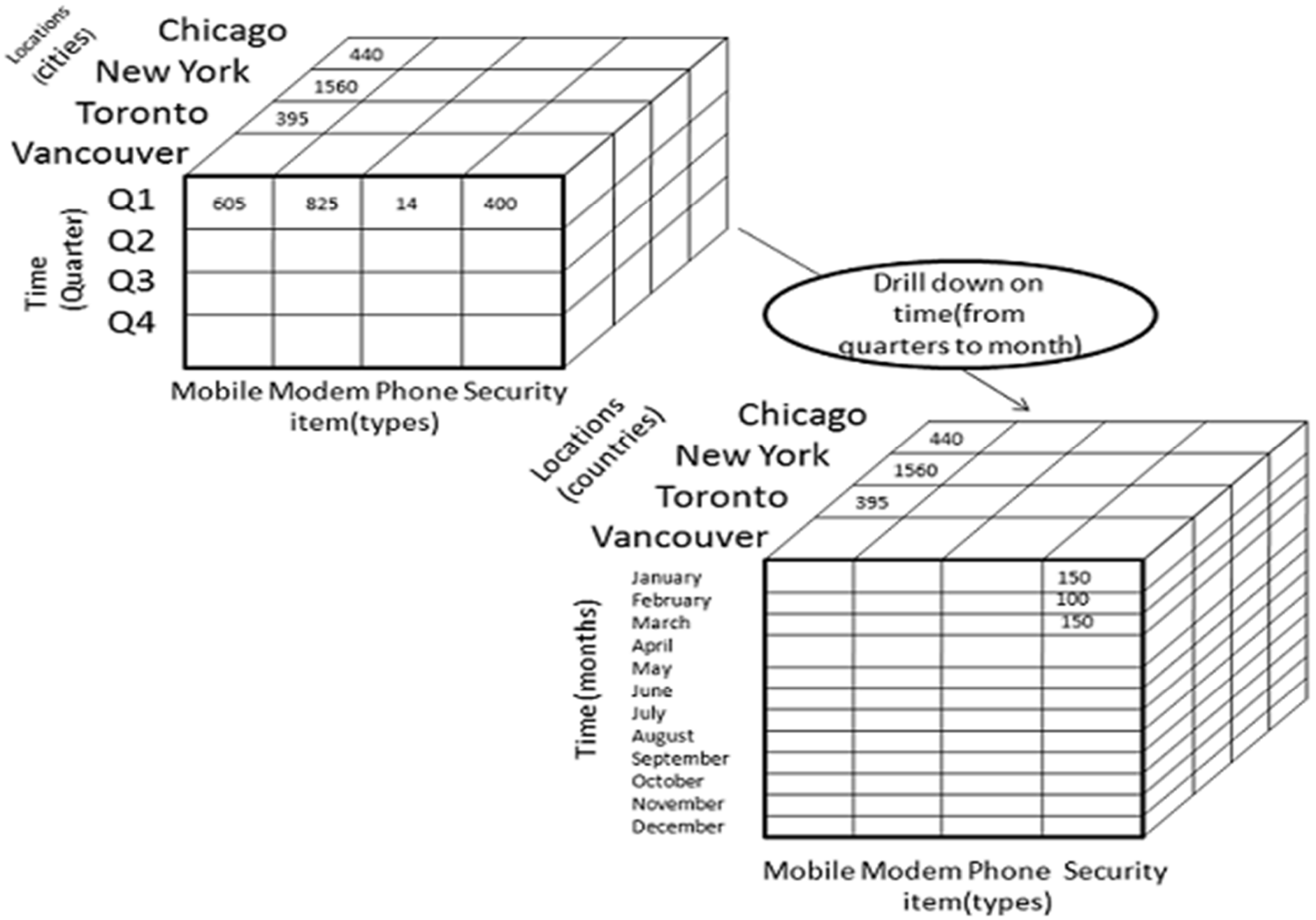
- The roll-up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.
- In fig. shows the result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for location.
- street<city<province or state<country.
- The roll up operation shown aggregates the data by ascending the location hierarchy from the level city to the level of country.


Locations (countries)
USA
Canada
2000
Q1
Q2
Q3
Q4
1000
Time (Quarter)
Mobile Modem Phone Security
item(types)
roll-up on location (from cities to countries)
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Time (Quarter)
Mobile Modem Phone Security
item(types)

Drill -down

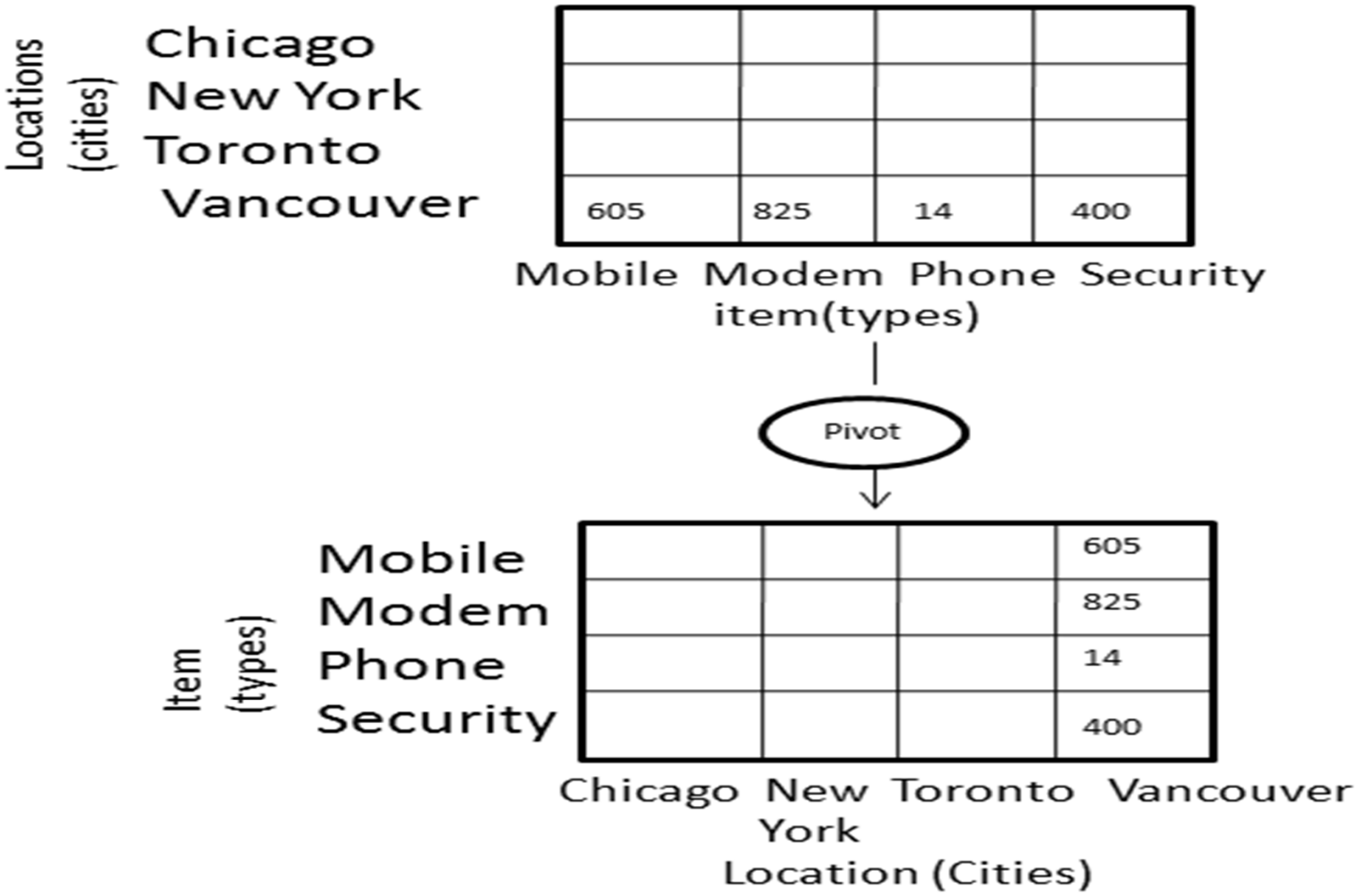
- Drill –down is the reverse of roll-up. It navigate from less detailed data to more detailed data.
- The fig. shows the result of a drill dowm operation performed on the central cube by stepping down a concept heirarchy for time defined as
- " day<month<quarter<year "
- Drill-down occurs by descending the time from level of quarter to the more detailed level of month.

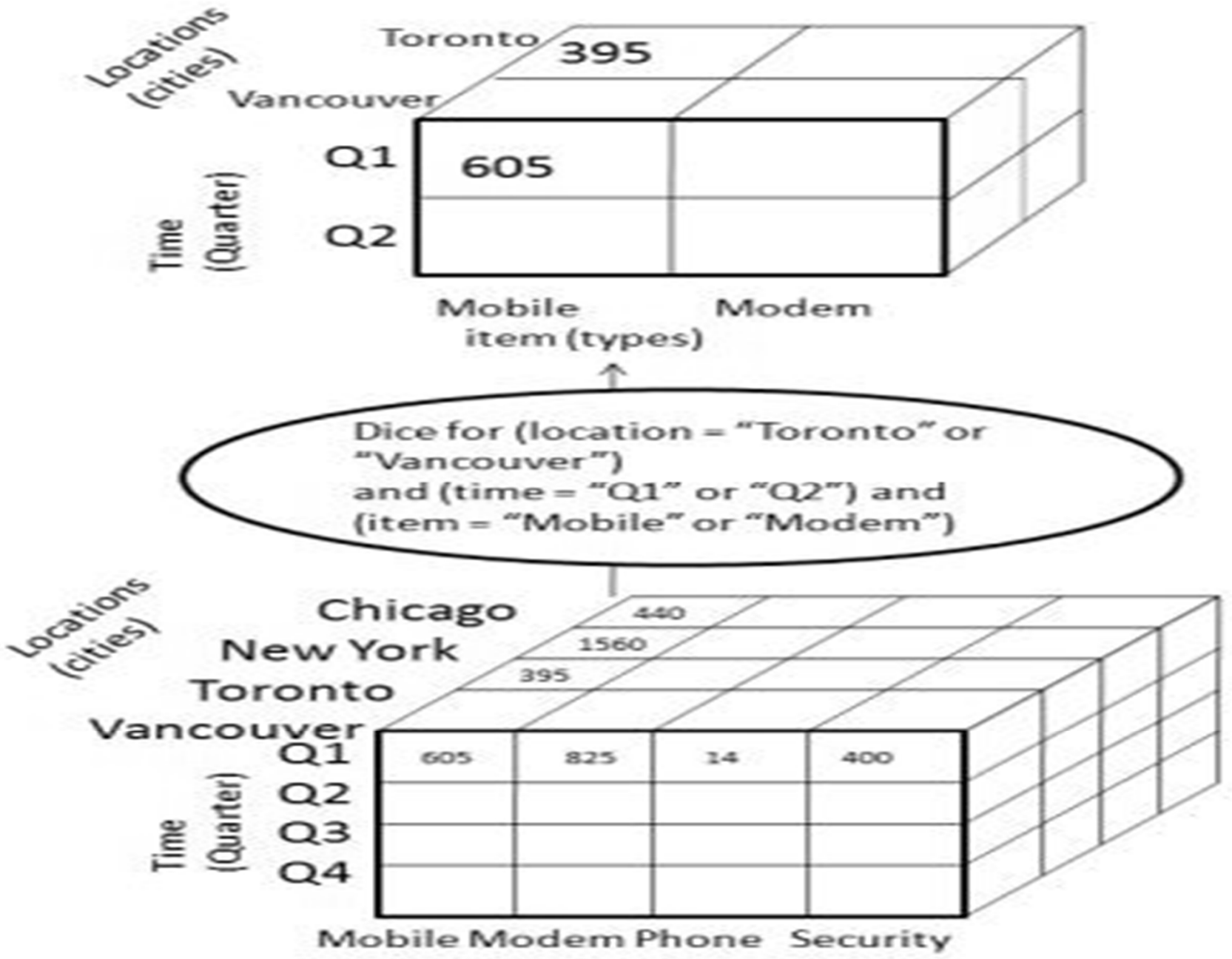
- The resulting data cube details the total sales per month rather than summarizing them by quarter.


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
Time (Quarter)
605
825
14
400
Mobile Modem Phone Security
item(types)
Drill down on time(from quarters to month)
Locations (countries)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
Mobile Modem Phone Security
item(types)

Slice and dice

- The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.
- Fig. Shows a slice operation where the sales data are selected from the central cube for the dimension time using the criterion time=“ Q1”.
- The dice operation defines a subcube by performing a selection on two or more dimensions.


Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile
Modem
Phone
Security
item(types)
Pivot
Item (types)
Mobile
Modem
Phone
Security
605
825
14
400
Chicago
New York
Toronto
Vancouver
Location (Cities)


Locations (cities)
Toronto
395
Vancouver
Q1
605
Q2
Time (Quarter)
Mobile
Modem
item (types)
Dice for (location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "Mobile" or "Modem")
Locations (cities)
Chicago
440
New York
1560
Toronto
395
Vancouver
Q1
605
825
14
400
Q2
Q3
Q4
Time (Quarter)
Mobile
Modem
Phone
Security

- Fig shows a dice operation on the central cube based on the following selection criteria that involve three dimension.
- (location="toronto " or "vancouver") and (time ="home entertainment " or " computer")
- Pivot(rotate): pivot is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

Other OLAP operations

- Additional drilling operations
- For eg: drill-across execute queries involving (i e across) more than one fact table.
- Dill-through uses relational sql facilities to drill through the bottom level of data cube down to its back end relational table.

Data warehouse architecture

- Design of a warehouse : A business analysis framework
- What can business analysts gain from having a data warehouse?

1. - d/w provide a competitive advantage by presenting relevant information from which to measure performance and make critical adjustment in order to help win over competitors.

2. Enhance business productivity because it is able to quickly and effectively gather information that accurately describe the organization.

3. A d/w facilitate customer relationship management because it provide a consistent view of customers and items across all lines of business, all department, and all markets.

4. A d/w may bring cost reduction by tracking trends, patterns, and exceptions over long periods in a consistent and reliable manner.

- To design effective DW we need to understand and analyze business needs and construct a business analysis framework.
- It can be viewed as construction of large and complex building.
- Owner -
- Architect
- Builder

- To design an effective d/w we need to understand and analyze business needs and construct a business analysis framework.
- 4 diff. view regarding the design of a data ware house

    1. top down

    2. data source view

    3. data warehouse view

    4. business query view

Top view

- Allow the selection of the relevant information necessary for the data warehouse.
- This inf. Matches the current and future business needs.
- The data source view:- exposes the information being captured , stored, and managed by operational systems.
- This information may be documented at various levels of detail and accuracy from individual data source tables to integrated data s/w tables.

The data source view

- Exposes the information being captured , stored, and managed by operational systems.
- This information may be documented at various levels of detail and accuracy from individual data source tables to integrated data s/w tables.

Data warehouse view

- Data warehouse view:- include fact tables and dimension tables.
- It represent the info. that is stored inside the d/w , including pre calculated totals and counts, as well as inf. regarding the source , date and time of origin added to provide historical context.
- Business query view:- is the perspective of data in data warehouse from the view point of the end user.

Business query view

- Business query view:- is the perspective of data in data warehouse from the view point of the end user.

D/w design process contain the following steps.

1. choose a business process to model.

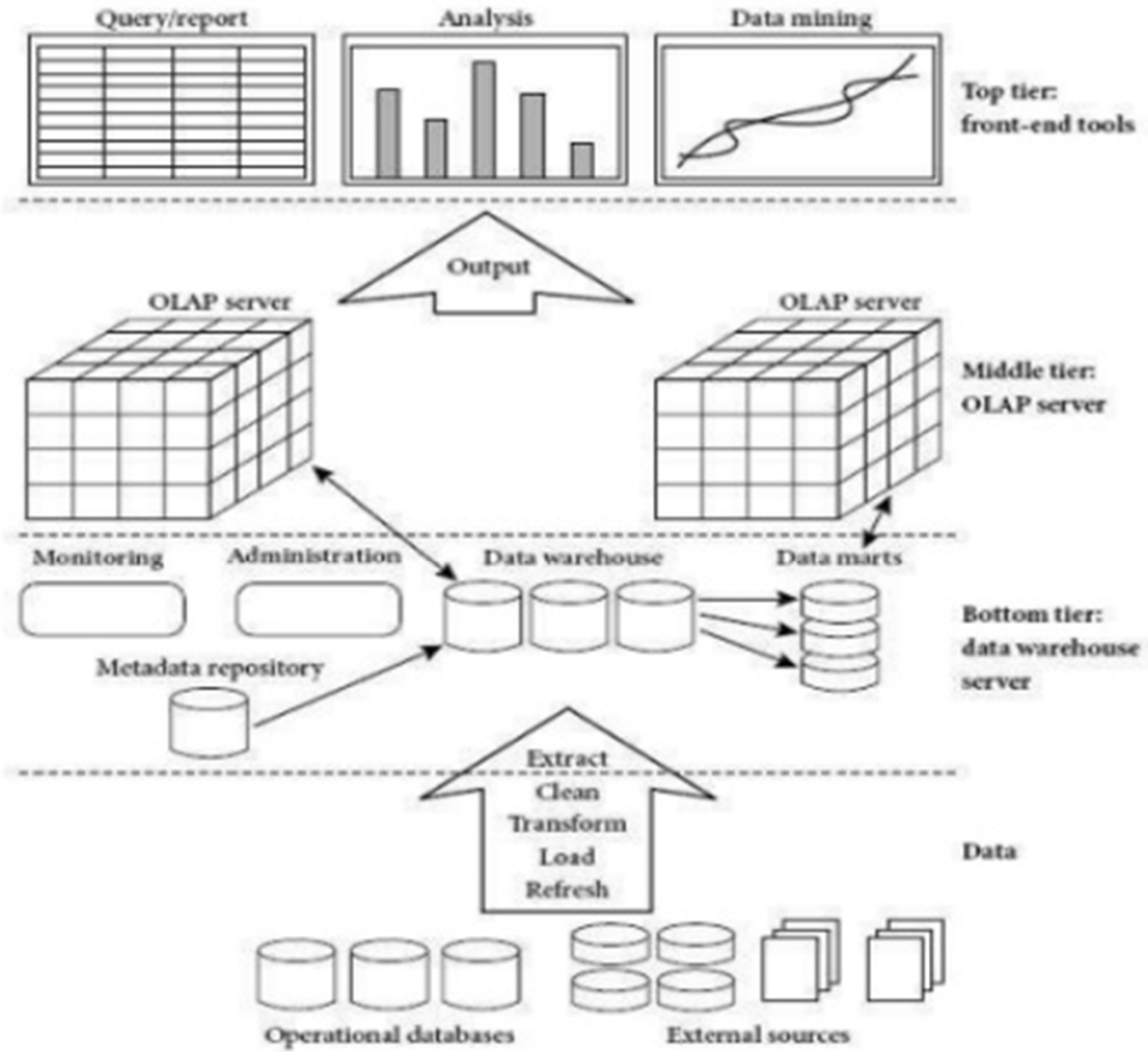
2. choose the grain of the business process.

3. choose the dimensions that will apply to each fact table record.

Three-tier DW architecture

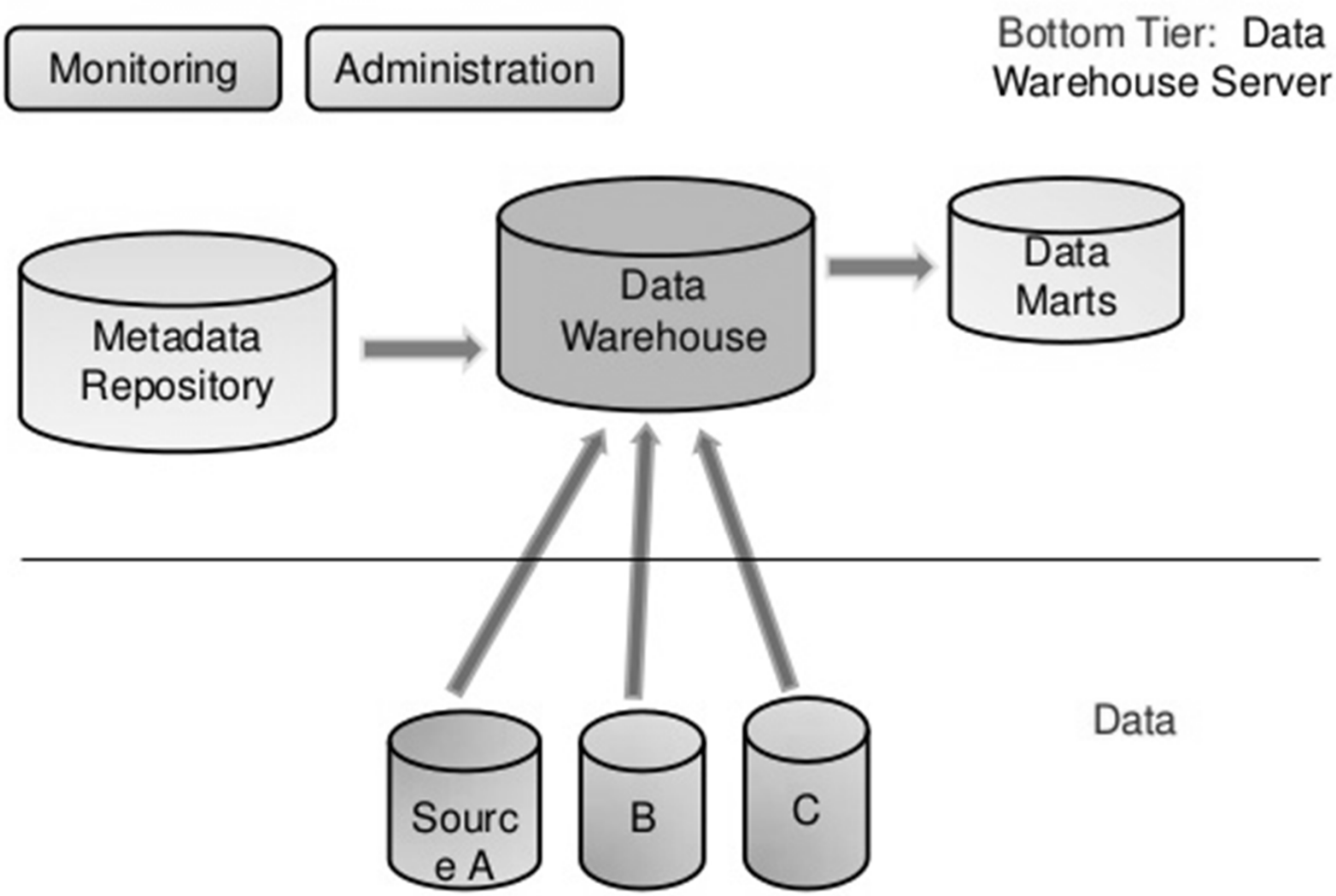
Data ware house adopt 3 tier

1. bottom tier
2. middle tier
3. top tier architecture



3.12 A three-tier data warehousing architecture.

CA)


Monitoring
Administration
Bottom Tier: Data Warehouse Server
Metadata Repository
Data Warehouse
Data Marts
Sourc e A
B
C
Data

Monitoring & Administration:

- Data Refreshment
- Data source synchronization
- Disaster recovery
- Managing access control and security
- Manage data growth, database performance
- Controlling the number & range of queries
- Limiting the size of data warehouse

Bottom tier

- Bottom tier is ware house server that is almost always a relational db system.
- Back end tools and utilities are used to feed data into bottom tier from operational db or external sources .
- These tool and utilities perform data extraction, data cleaning, and transformation.
- The data are extracted using application program interface known as gateways.
- A gateway is supported by underlying DBMS and allow client program to generate SQL code to be executed at server.(ODBC,OLEDB,JDBC)

Middle Tier: OLAP Server

- It presents the users a multidimensional data from data warehouse or data marts.
- Typically implemented using two models:

ROLAP Model

Present data in relational tables

MOLAP Model

Present data in array based structures means map directly to data cube array structure.

MIDDLE TIER

- The middle tier is an OLAP server that is typically implemented using either
  - 1.relational OLAP(ROLAP) extended relational DBMS maps operations on multidimensional data to standard relational operations.
  - 2. multidimensional OLAP(MOLAP) a special server that directly implements multidimensional data and operations.

Top Tier: Front end tools

It is front end client layer.

- Query and reporting tools

  Reporting Tools: → Production reporting tools
  → Report writers

  Managed query tools: Point and click creation of SQL used in customer mailing list.

- Analysis tools : Prepare charts based on analysis

- Data mining Tools: mining knowledge, discover hidden piece of information, new correlations, useful pattern

Top tier

- Is the front end client view which contain query and reporting tool, analysis tools and /or data mining tools.
  - From the architecture point of view there are 3 data warehouse model
  - 1. enterprise ware house
  - 2. data mart
  - 3. virtual ware house

From the architecture point of view there are 3 data warehouse model

1. Enterprise ware house
2. Data mart
3. Virtual ware house

Enterprise warehouse

- An enterprise warehouse collects all of information about subject spanning the entire organization.
- It provide cooperate wide data integration.
- Detailed as well as summarize data.
- Data range a few gigabyte to 100 gigabyte, terabyte or beyond

Data mart

- Data mart contain subset of cooperate wide data that is of value to a specific group of users.
- Scope is confined to specific selected subjects – customer, item, sales
- Data mart usually implemented on low cost departmental servers that are unix/LINUX or windows based.
- Depending on source of data data mart can be categorized into 1. independent data mart 2. dependent data mart

Virtual ware house

- A virtual ware house is a set of views over operational data base. A virtual warehouse is easy to build but requires excess capacity on operational d/b servers.

Data ware house application

- Tools for data warehouse can be categorized into access and retrieval tools, database reporting tools, data analysis tools and data mining tools.
- Information processing:- support querying, basic statistical analysis, and reporting crosstab, chart or graph.
- A current trend in data warehouse information is to construct low –cost web based accessing tools that are then integrated with web browsers.

- Analytical processing:- support basic OLAP operations , including slice and dice , drill down, roll up and pivoting.
- It generally operates on historical data in both summarized and detailed forms.
- The major strength of online analytical processing over info. Processing is the multidimensional data analysis of data ware house data.

- Data mining:- support knowledge by finding hidden pattern and associations, constructing analytical models, performing classification and prediction and presenting the mining result using visualization tools.