# Data Warehousing

# Data Warehousing

## Introduction

- Data warehousing is the process of constructing and using a data warehouse.
- A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries,and decision making.
- Data warehousing involves data cleaning, data integration,and data consolidations.

# Data Warehousing

## Introduction

- DBMS in industry are pervasive throughout industry with relational database management being the dominant systems.
- These types of systems are called **online transaction processing systems-OLTP.**
  - **OLTP** are the normal business databases e.g. inventory control, invoicing, etc. and are designed to maximise the transaction processing capacity
  - They are designed to handle high transaction throughput
  - transactions typically making small changes to the operational data
  - operational data is data that the organisation requires to handle in its day-to-day operations
- OLTP systems size can range from small databases being mbs to large databases that can require terabytes or even petabytes of storage.

# Why Data Warehousing?

- **Support decision-making**: Corporate decision makers require access to all the organisation's data both current and past data(historic) to provide comprehensive analysis of the organisation, its business, its requirements and its trends.
  - Thus Organizations are focusing on ways to use operational data to support decision-making as a means of regaining competitive advantage.
- To facilitate this type of analysis, a **data warehouse** is created to hold data drawn from several data sources, maintained by different operating systems, together with historical and summary transformations.
  - Based on extended database technology, a data warehouse provides the management with a data store.

# Data Warehousing

## Definition

- A data warehouse is a **subject-oriented**, **integrated**, **time-variant** and **non-volatile** collection of data in support of management's decision-making process' (Inmon 1993)
  - Its a repository of information, or archive information, gathered from multiple sources stored under a unified schema.

# Data Warehousing
## Definition ~ Explained

- **Subject oriented**: it is organised around the major subjects of the enterprise (e.g. customers, products, sales) rather than the major application areas (e.g. customer invoicing, stock control, product sales)
  - A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.
- **Integrated:** coming together of the source data from different enterprise-wide applications systems. Often inconsistent, e.g. different formats
  - A data warehouse integrates data from multiple data sources.
    - For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

# Data Warehousing
## Definition ~ Explained

- **Time-variant:** data in the warehouse is only accurate at some point in time or over some time interval
  - Historical data is kept in a data warehouse.
    - For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse.
  - This contrasts with a transactions system, where often only the most recent data is kept.
    - For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.
- **Non-volatile:** data is not updated in real time but is refreshed from operational systems on a regular basis. New data is always added, rather than replaced.
  - Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

# Data Warehousing

Challenges in creating a Data warehouse :

- Data is scattered in many types of incompatible structure
- Lack of documentation prevent integration older legacy systems with newer systems
- Internet software like searching engine needs to be improved
- Accurate and accessible metadata across multiple organizations is hard to get

# Data Warehousing

- Data warehouses are designed to support adhoc query processing, therefore they are organised according to the requirements of potential questions and supports long term strategic decision making
- It is often the case that OLTP systems provide the data for data warehouses.
  - However the data held in OLTP systems can be inconsistent, fragmented, contain duplicate or missing entries.
  - This must be cleaned up before it can be used in a data warehouse.

# OLPT versus Data Warehouse

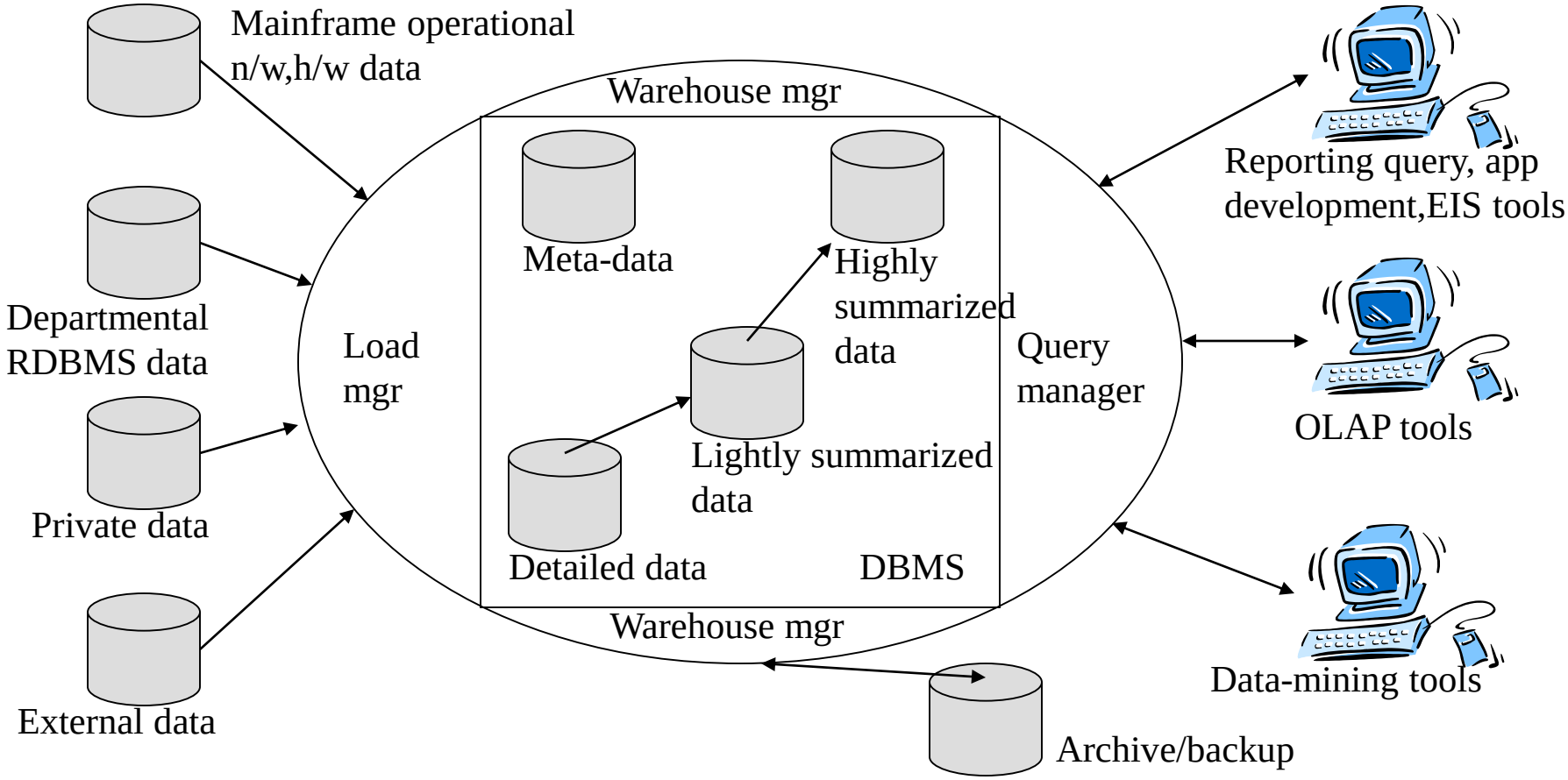
| OLTP | Data Warehouse |
|---|---|
| Holds current data | Holds historic data |
| Stores detailed data | Detailed, lightly/highly summarised data |
| Data is dynamic | Data largely static |
| Repetitive processing | Ad hoc querying, unstructured and heuristic processing |
| High transaction throughput | Medium-low level transaction throughput |
| Predictable usage patterns | Unpredictable usage patterns |
| Transaction driven | Analysis driven |
| Application oriented | Subject oriented |
| Supports day-to-day decisions | Strategic decisions |
| Large number of clerical/operational users | Lower number of managerial users |

Source: Connolly and Begg p1153

# Benefits

- Potential high returns on investment
  - Organisations normally must commit a huge amount of investment and resources in developing the data warehouse, but the potential returns on that investment due to **increasing productivity** and the **competitive advantage** that gives can be very large
- Competitive advantage
  - Competitive advantage is gained by giving access to this data by management decision makers for forecasting trends etc.
- Increased productivity of corporate decision-makers
  - By transforming data into a meaningful information, a data warehouse allows managers to perform more substantive, accurate and consistent analysis

# Typical Architecture of Data Warehouse




Source: Connolly and Begg p1157

# Typical Architecture of Data Warehouse

A typical architecture of a data warehouse consists of the following:

- **Data sources**
  - vary from mainframes to departmental databases to external data.
  - There are lots of different sources and different data types.
- **Load manager** (or frontend)
  - performs the extraction and loading of data into the warehouse.

# Typical Architecture of Data Warehouse

- **Warehouse manager**
  - performs all the operations associated with the management of the data in the warehouse.
  - Operations include:
    - ensuring consistency of data,
    - indexes and views,
    - denormalising,
    - aggregating data, backing up and archiving.

# Typical Architecture of Data Warehouse

- **Query manager** (backend)
  - manages the user queries.
  - Complexity depends on flexibility of end-user access tools.
  - Can include directing queries to tables, scheduling execution of queries, generating query profiles to assist warehouse manager in managing indexes and views.

# Typical Architecture of Data Warehouse

## Types of Data

- **Detailed data:**
  - This is the actual data which has been pulled in from the various sources.
  - Normally stored offline and aggregated into next level of data.

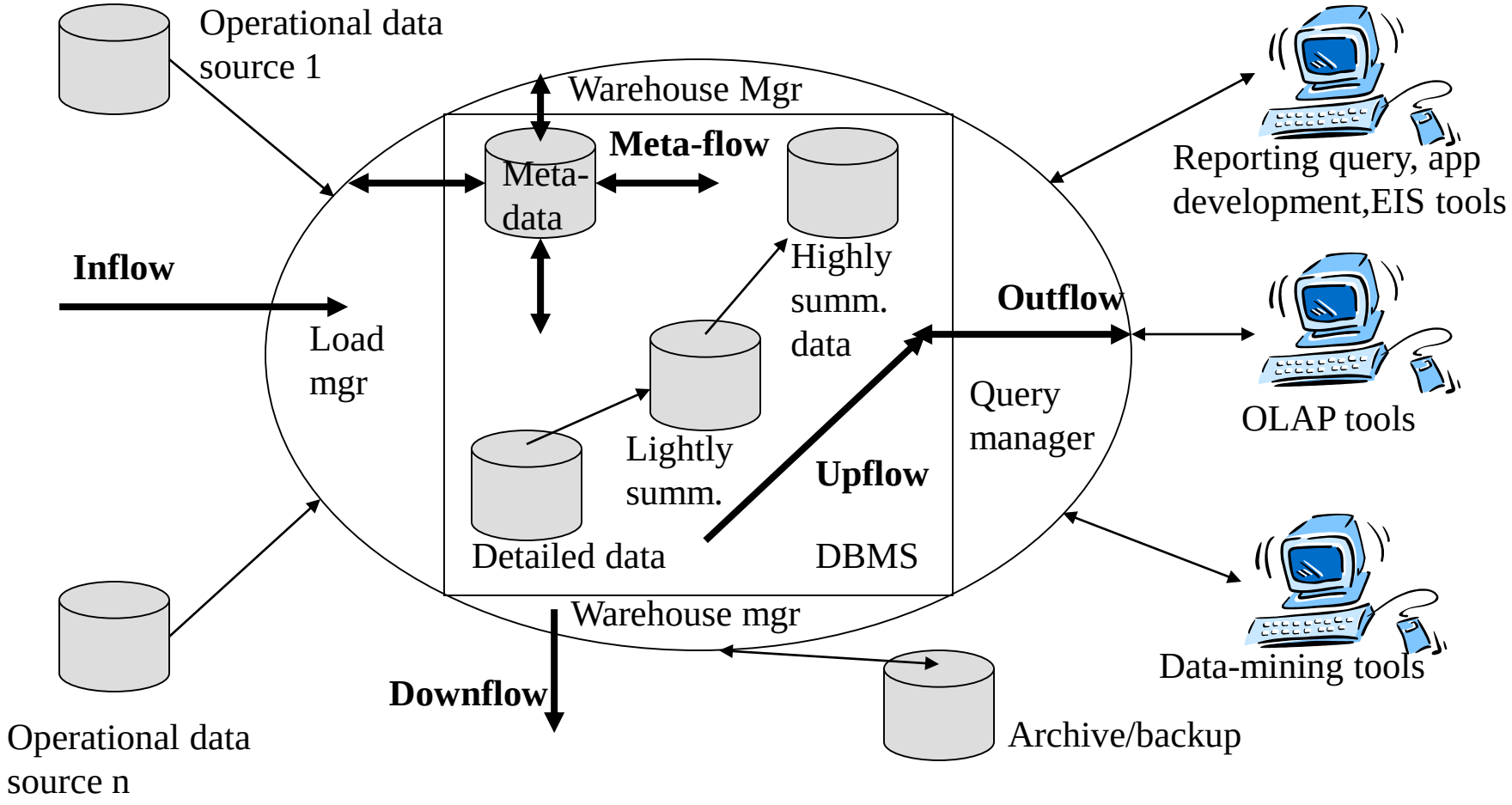
# Typical Architecture of Data Warehouse

- **Lightly/highly summarised data:**
  - Summarised data tends to create various views of the detailed data, to answer specific queries.
    - It is the aggregated data generated by the warehouse manager.
  - It needs to be summarised because there is such a large amount of data
    - Purpose is to speed up queries.
  - Because these views can change, there also needs to be meta-data.
    - This is subject to change on an on-going basis depending on the types of queries.

# Typical Architecture of Data Warehouse

- **Meta-data:**
  - Is a description of data in warehouse.
  - Changes according to structure of data in warehouse.
- **Archive/backup:**
  - Since the data warehouse will always grow, some of the older data can be archived, in a way that it can still be included in queries if required.

# Information Flows




Source Connolly and Begg p1162

# Information Flow Processes

- Five primary information flows
  - **Inflow**
    - Extraction~cleans dirty data,
    - Cleansing ~ restructures data to suit new requirements
    - loading of data from source systems into warehouse ~ ensure source is consistent with data already in the warehouse
  - **Upflow**
    - adding value to data in warehouse through summarizing data into more convenient views, packaging data into more useful formats and distributing data to increase availability/accessibility

# Information Flow Processes

- **Downflow**
  - Archiving (transfer data of limited value) and backing up data in warehouse mostly to be used to restore following crash
- **Outflow**
  - making data available to end users
- **Metaflow**
  - This is the process which moves metadata responding to changing needs, i.e. updating metadata accordingly

# Problems of Data Warehousing

These are the typical problems with a data warehouse, most of them arise from the problems of integrating the source data(see pages 1050-1052, Connolly & Begg)

1. Underestimation of resources for data loading
   - 80% of development time is spent on data loading
2. Hidden problems with source systems
   - Problems with source systems, e.g. nulls allow incomplete data, needs to be fixed
3. Required data not captured - OLTP systems may not store data needed – so may need to alter the OLTP systems
4. Increased end-user demands – once users become aware of capabilities – need better tools

# Problems of Data Warehousing

5. Data homogenization - Homogenization can lessen value of data – similarities v. differences in data
6. High demand for resources e.g. disk space, large no. of indexes etc.
7. Data ownership – the data accessible to all users
8. High maintenance – need for reorganisation of business processes or change to data warehouse
9. Long duration projects – i.e. can take 3 years to build – data marts support only one department so may be quicker
10. Complexity of integration – there is need to integrate all tools to ensure benefits to the organization

# Data Warehouse Design

- Data must be designed to allow ad-hoc queries to be answered with acceptable performance constraints
  - The types of queries needed to be performed are different to those in an OLTP system as they are more factual, analytical and temporal.
- Queries usually require access to factual data generated by business transactions
  - e.g. find the average number of properties rented out with a monthly rent greater than £700 at each branch office over the last six months

# Data Warehouse Design

- Uses Dimensionality Modelling
  - Normal modelling techniques (E-R model) are not suitable as the relationships between the data can sometimes be too complex,
  - We therefore use dimensionality modelling - a logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access.

# Dimensionality Modelling

- Similar to E-R modelling but with constraints/restrictions
  - composed of one **fact** table with a composite primary key
  - **dimension** tables have a simple primary key which corresponds exactly to one foreign key in the fact table
  - uses **surrogate** keys based on integer values e.g unique sequential number usually not derived from any other data in the database
    - They are usually created when the table does not have any natural primary key
  - Can efficiently and easily support ad-hoc end-user queries

# Star Schema

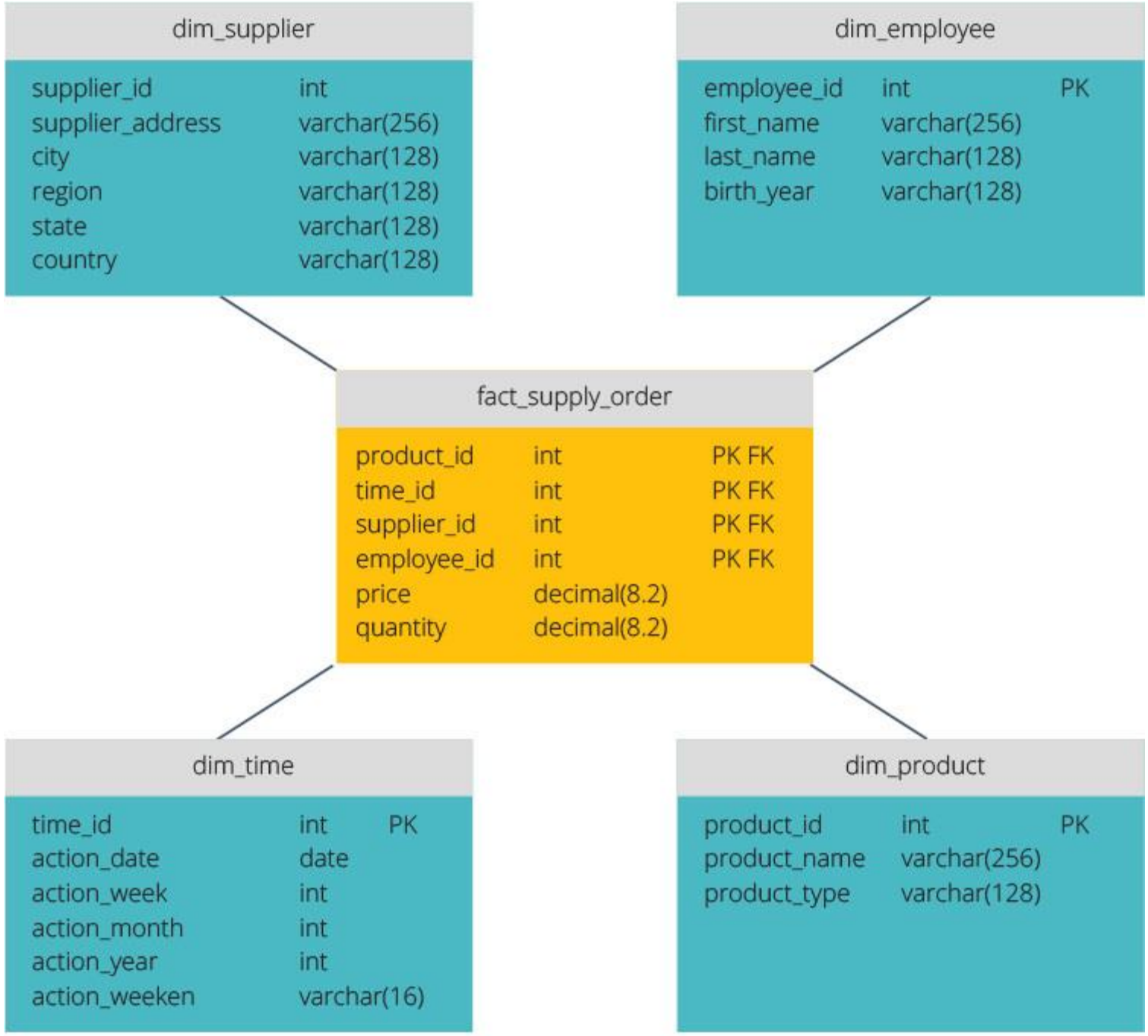
- This is the most common dimensional model
- A star schema contains two types of tables - a fact table surrounded by dimension tables
- **Fact tables**
  - contains FK for each dimension table
  - large relative to dimension tables
  - read-only
- **Dimension tables**
  - Dimensional tables contain reference data – i.e. the data which supports the fact.
  - query performance speeded up by denormalising into a single dimension table. Denormalised minimises the number of joins

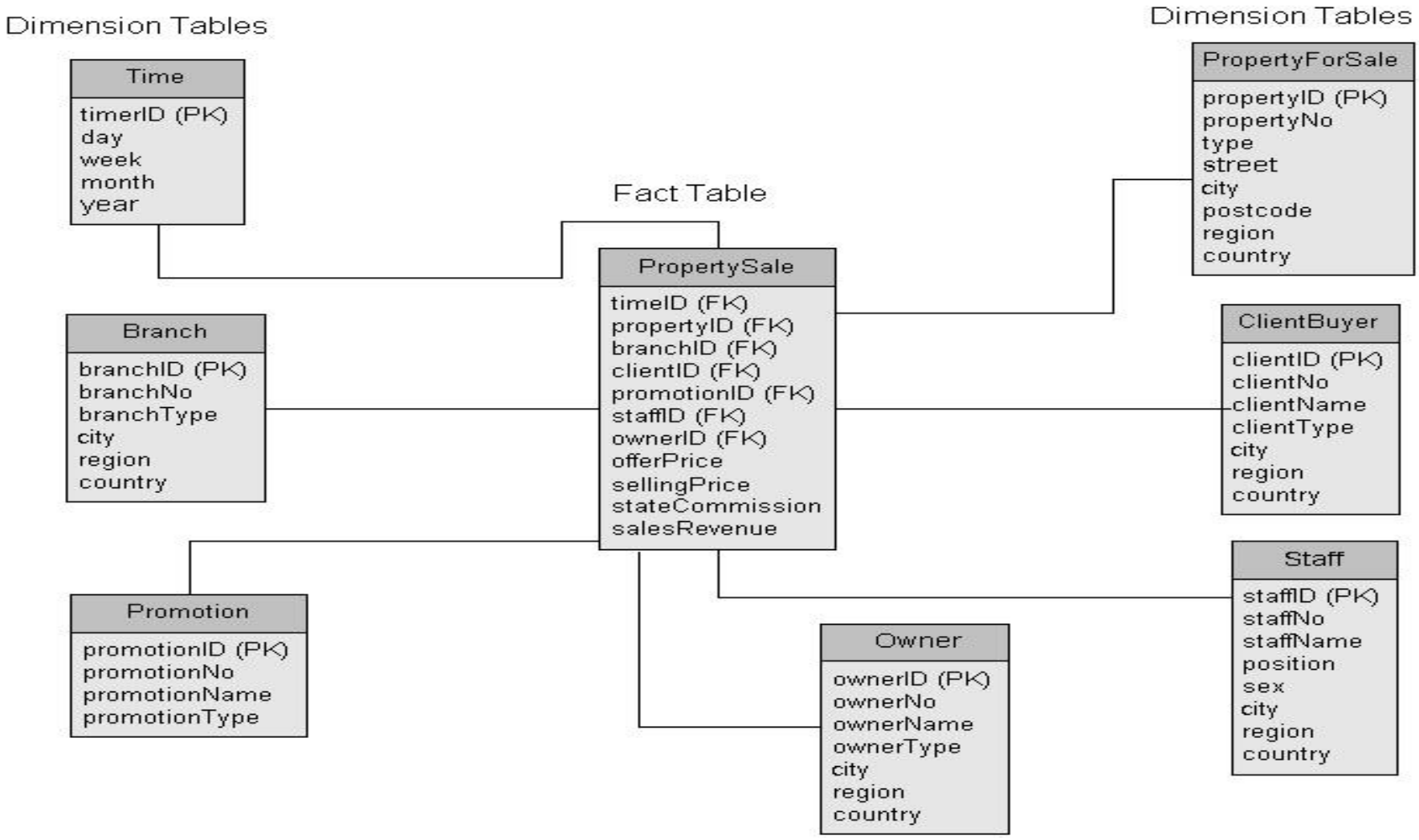
# Star Schema

- The star schema has a centralized data repository, stored in a fact table.
- The schema splits the fact table into a series of denormalized dimension tables.
- The fact table contains aggregated data to be used for reporting purposes while the dimension table describes the stored data.
- Denormalized designs are less complex because the data is grouped.

# Star Schema

- The fact table uses only one link to join to each dimension table.
- The star schema's simpler design makes it much easier to write complex queries.


dim_supplier
supplier_id int
supplier_address varchar(256)
city varchar(128)
region varchar(128)
state varchar(128)
country varchar(128)
dim_employee
employee_id int PK
first_name varchar(256)
last_name varchar(128)
birth_year varchar(128)
fact_supply_order
product_id int PK FK
time_id int PK FK
supplier_id int PK FK
employee_id int PK FK
price decimal(8.2)
quantity decimal(8.2)
dim_time
time_id int PK
action_date date
action_week int
action_month int
action_year int
action_weeken varchar(16)
dim_product
product_id int PK
product_name varchar(256)
product_type varchar(128)




Source: Connolly and Begg

# Star Schemas

**Explanation:**

- This is the Star Schema version.
  - We have one table in the centre which contains all the links to the dimension tables, which contains the data.
  - The fact table is just like a M:N relationship in a relational database.
  - Note that there can be more than one fact table in a star schema.

# Snowflake Schema

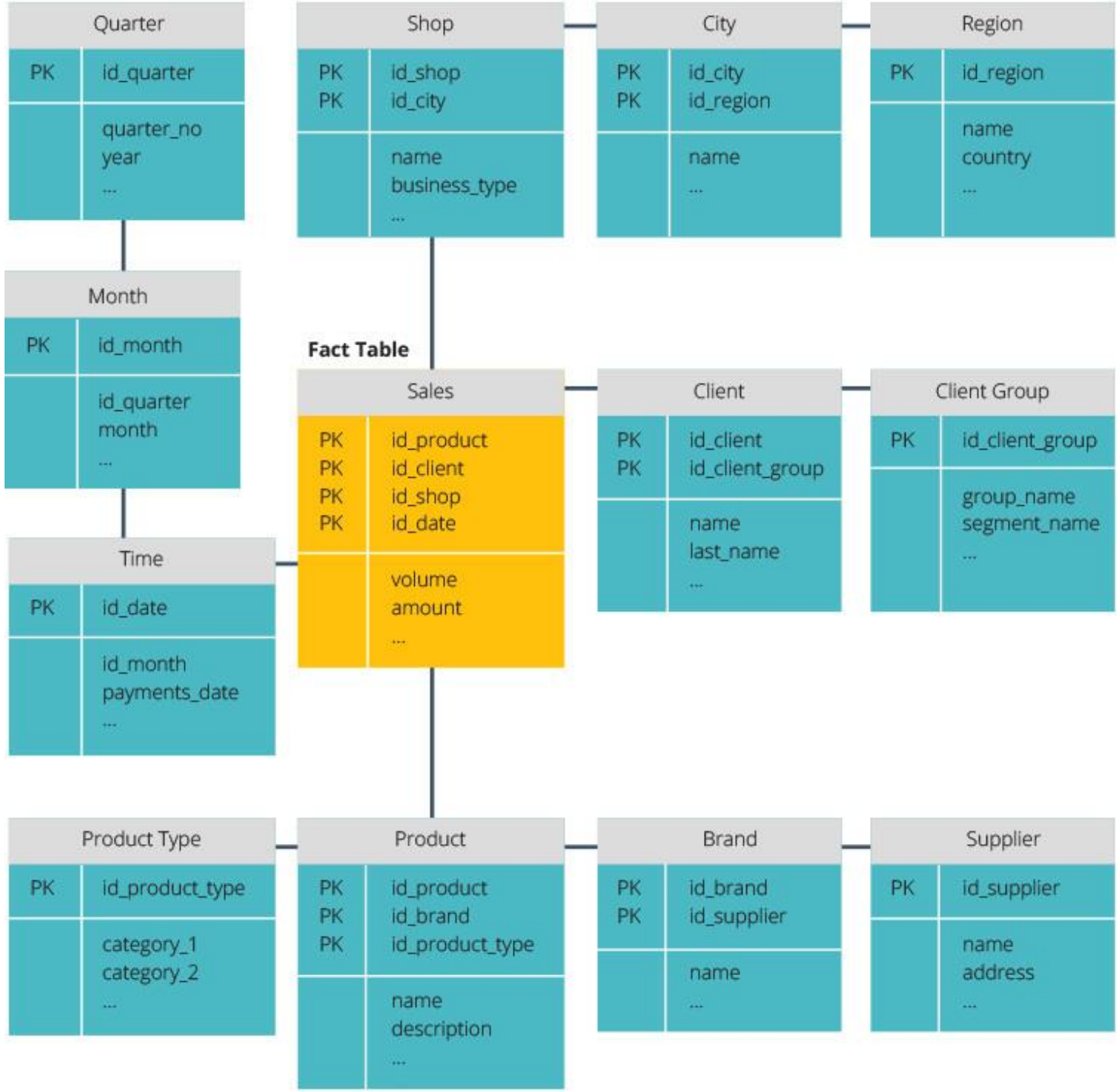
- The snowflake schema is different because it normalizes the data.
- Normalization means efficiently organizing the data so that all data dependencies are defined, and each table contains minimal redundancies.
- Single dimension tables thus branch out into separate dimension tables.
- The snowflake schema uses less disk space and better preserves data integrity.

# Snowflake Schema

- The main disadvantage with snowflake schema is the complexity of queries required to access data—each query must dig deep to get to the relevant data because there are multiple joins.

***Key Points on Snowflake schemas***

- variant of star schema
- each dimension can have its own dimensions
- Unlike Star schema which is denormalised, Snowflake schemas contain no denormalised data, the data is normalised.


Quarter
PK id_quarter
quarter_no
year
...
Shop
PK id_shop
PK id_city
name
business_type
...
City
PK id_city
PK id_region
name
...
Region
PK id_region
name
country
...
Month
PK id_month
id_quarter
month
...
Fact Table
Sales
PK id_product
PK id_client
PK id_shop
PK id_date
volume
amount
...
Client
PK id_client
PK id_client_group
name
last_name
...
Client Group
PK id_client_group
group_name
segment_name
...
Time
PK id_date
id_month
payments_date
...
Product Type
PK id_product_type
category_1
category_2
...
Product
PK id_product
PK id_brand
PK id_product_type
name
description
...
Brand
PK id_brand
PK id_supplier
name
...
Supplier
PK id_supplier
name
address
...

# Starflake schema

- hybrid structure
- contains mixture of (denormalised) star and (normalised) snowflake schemas

# Data Warehouse Models

- In a traditional architecture there are three common data warehouse models:
  - Virtual warehouse
  - Data mart
  - Enterprise data warehouse

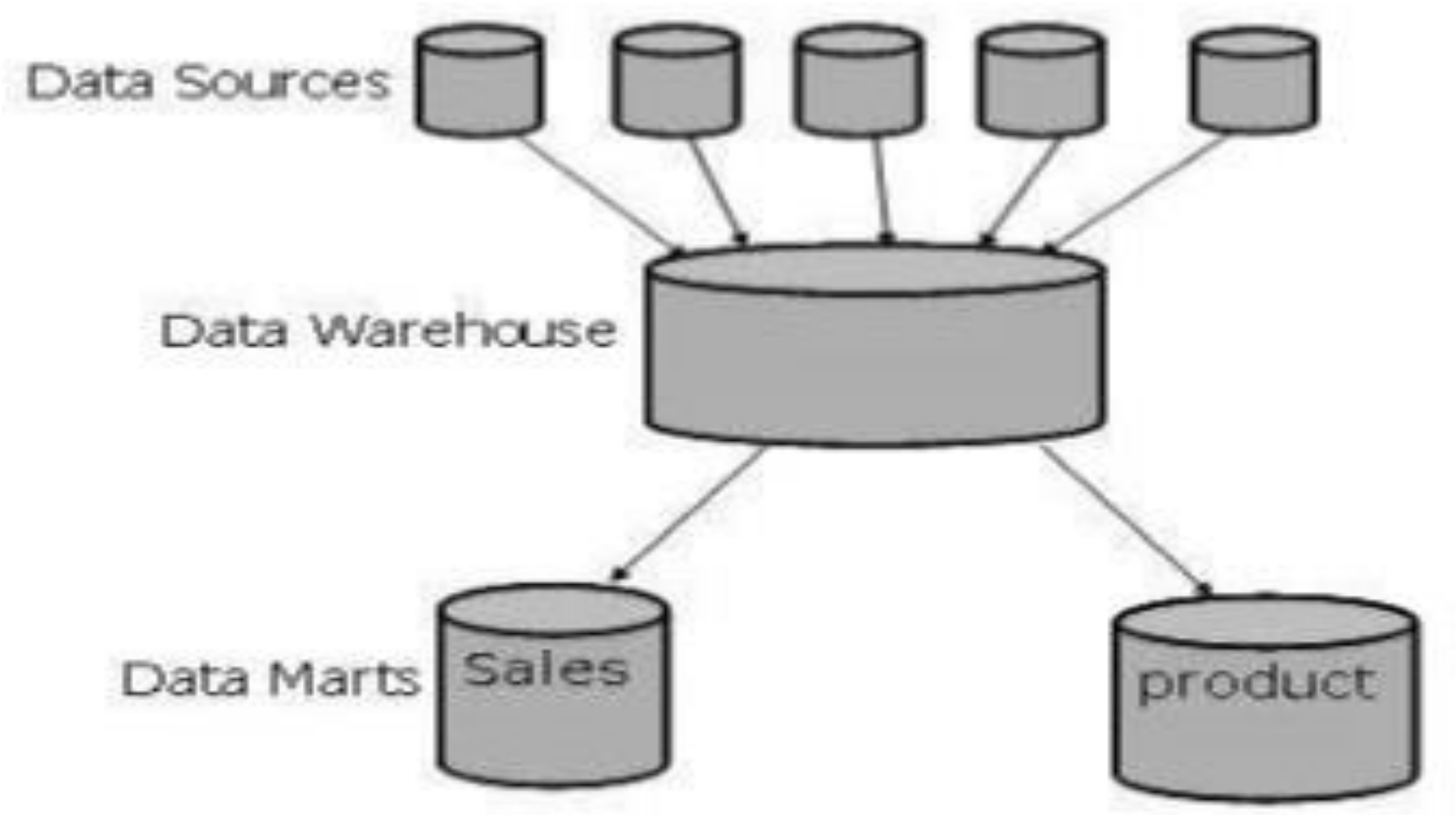
# Virtual warehouse

- A **virtual data warehouse** is a set of separate databases, which can be queried together, so a user can effectively access all the **data** as if it was stored in one **data warehouse**.
- In this **data warehouse** model, **data** is aggregated from a range of source systems relevant to a specific business area, such as sales or finance.
- views over operational dbs
- Materialize some summary views for efficient query processing
- Easier to build

# Data Mart

- Data Mart is a subset of data warehouse (or organization-wide data) that supports the requirements of a particular department or business function or specific group of people.
  - In other words,adata mart contains only those data that is specific to a particular group.
  - For example,the marketing data mart may contain only data related to items, customers,and sales.
- Data marts are confined to subjects.

# Graphical representation of a data mart



# Data Warehouse Vs Data Marts

- Data Marts:
  - Departmental subsets that focus on selected subjects: Marketing data mart: customer, products, sales.
  - Faster roll out, but complex integration in the long run.

# Enterprise warehouse

- An enterprise data warehouse model prescribes that the data warehouse contain aggregated data that spans the entire organization.
- This model sees the data warehouse as the heart of the enterprise's information system, with integrated data from all business units.
  - collects all information about subjects (customers, products, sales, assets, personnel) that span the entire organization.

– Requires extensive business modeling

– May take years to design and build

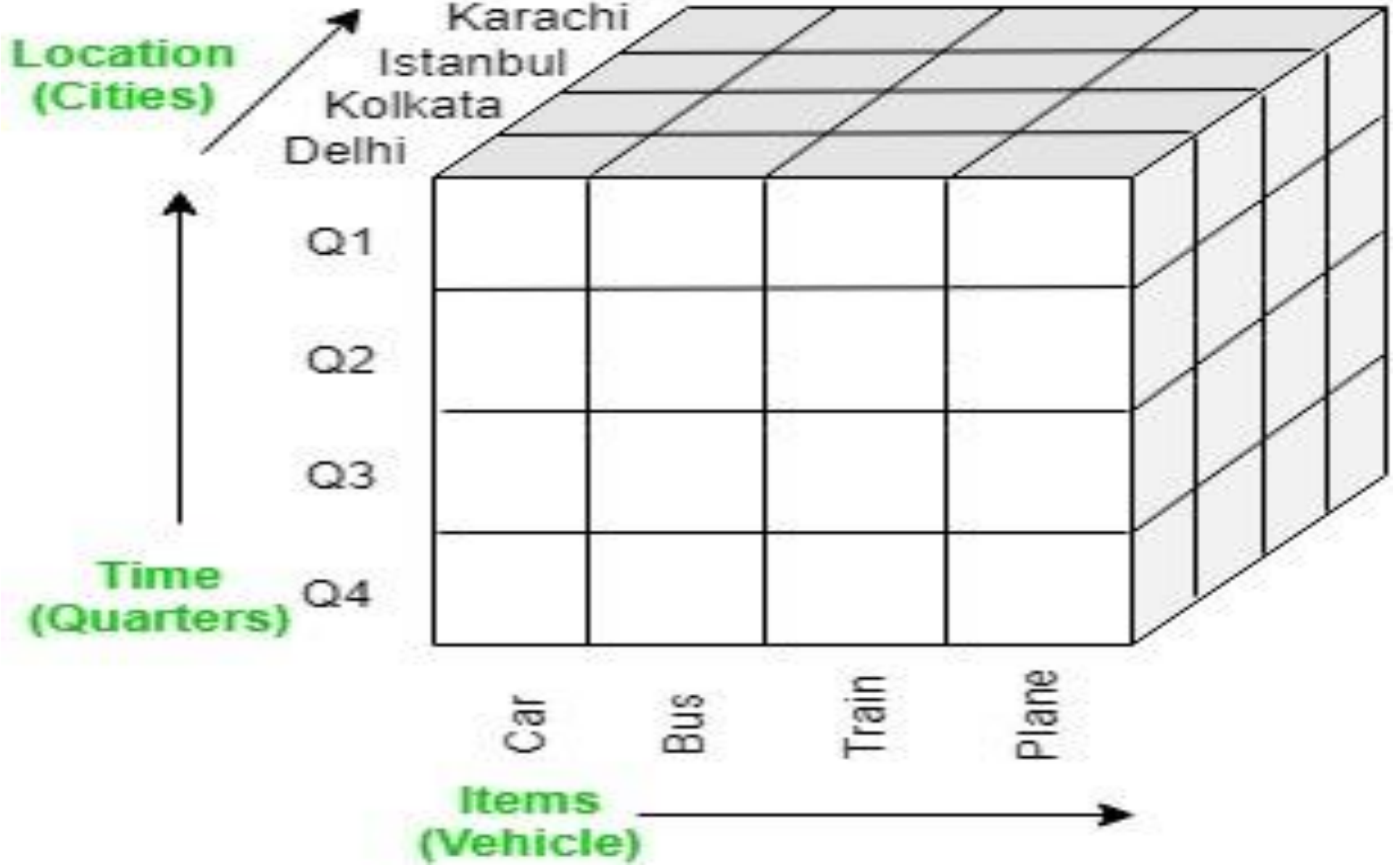
# Data Analysis Tools

- There are two types of tools that are commonly used for data analysis:
  - Online Analytical Processing (OLAP)
  - Data mining
- Each tool is described in the next slides

# Online Analytical Processing Server (OLAP)

- **OLAP (Online Analytical Processing)** Server is a software technology that allows users to analyze information from multiple database systems at the same time.
  - It is based on multidimensional data model and allows the user to query on multi-dimensional data (eg. Delhi -> 2018 -> Sales data).
- OLAP databases are divided into one or more cubes that are designed in such a way that creating and viewing reports become easy.
  - OLAP Cube is also called the **hypercube**.
  - OLAP allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

# OLAP


Location
(Cities)
Karachi
Istanbul
Kolkata
Delhi
Q1
Q2
Q3
Q4
Time
(Quarters)
Car
Bus
Train
Plane
Items
(Vehicle)

# How OLAP works

- A Data warehouse would extract information from multiple data sources and formats like text files, excel sheet, multimedia files, etc.
- The extracted data is cleaned and transformed.
- Data is loaded into an OLAP server (or OLAP cube) where information is pre-calculated in advance for further analysis.

# OLAP Operations

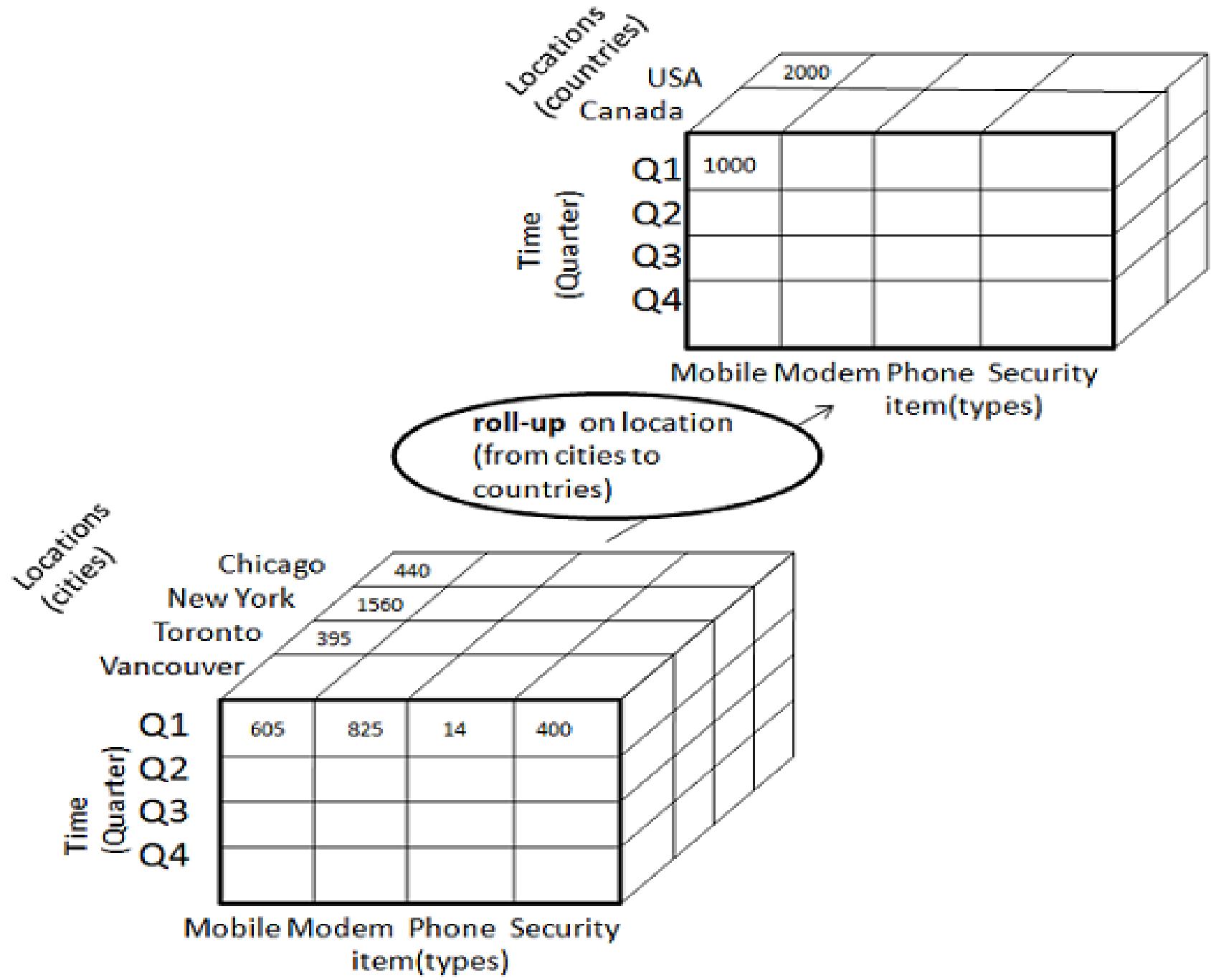
- Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.
- Here is the list of OLAP operations −
  - Roll-up
  - Drill-down
  - Slice and dice
  - Pivot (rotate)

# OLAP Operations

1. **Roll-up**

- Roll-up performs aggregation on a data cube in any of the following ways –
  - By climbing up a concept hierarchy for a dimension
  - By dimension reduction
- The following diagram illustrates how roll-up works.

# Roll-up


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
Q1
Q2
Q3
Q4
605
825
14
400
Time (Quarter)
Mobile Modem Phone Security
item(types)

# OLAP Operations

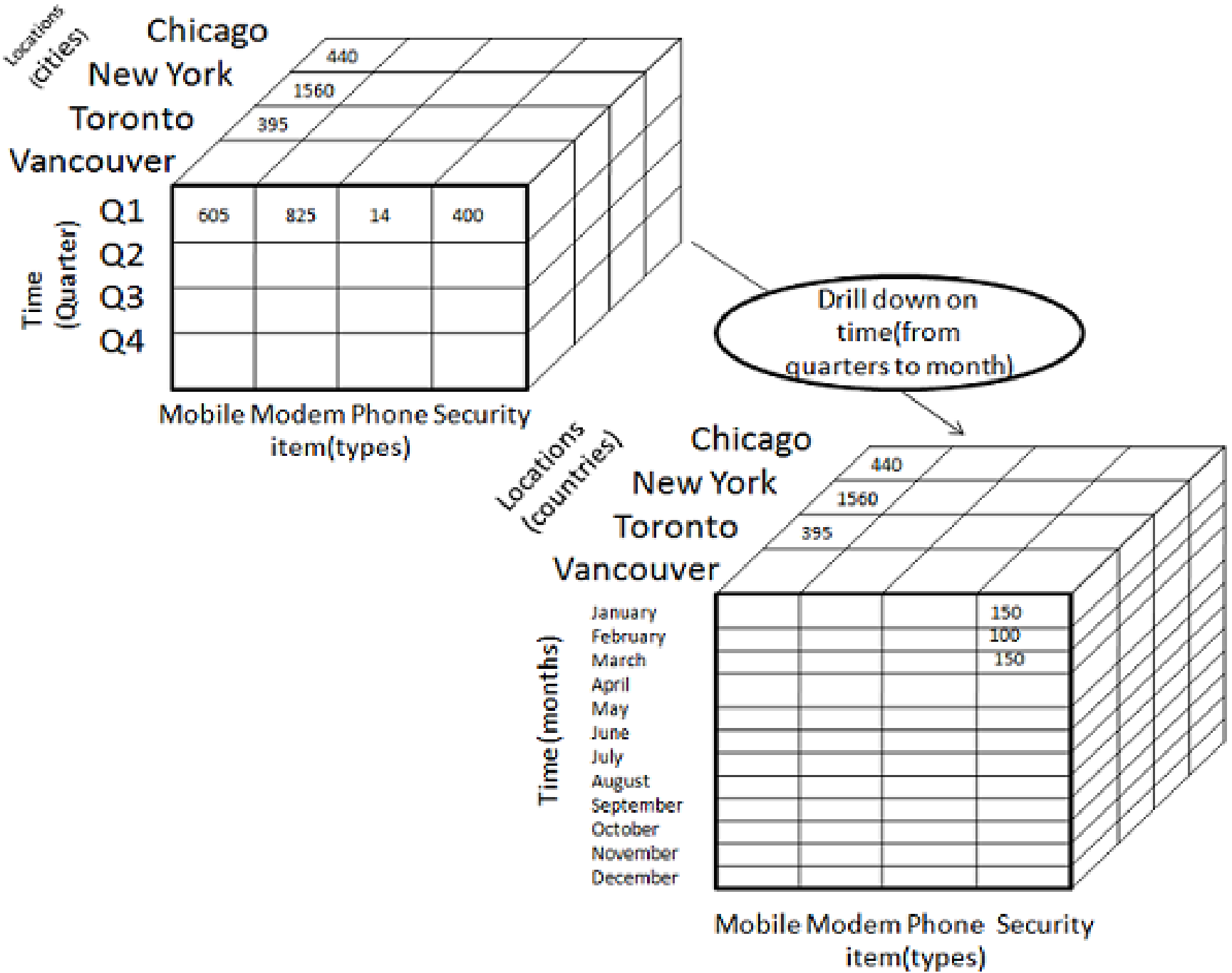
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

# OLAP Operations

## 2. Drill-down

- Drill-down is the reverse operation of roll-up.
- In drill-down operation, the less detailed data is converted into highly detailed data.
- It can be done by:
  - By stepping (moving) down in the concept hierarchyfor a dimension
  - By introducing (adding) a new dimension
- The following diagram illustrates how drill-down works

# Drill-down


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
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)

# OLAP Operations

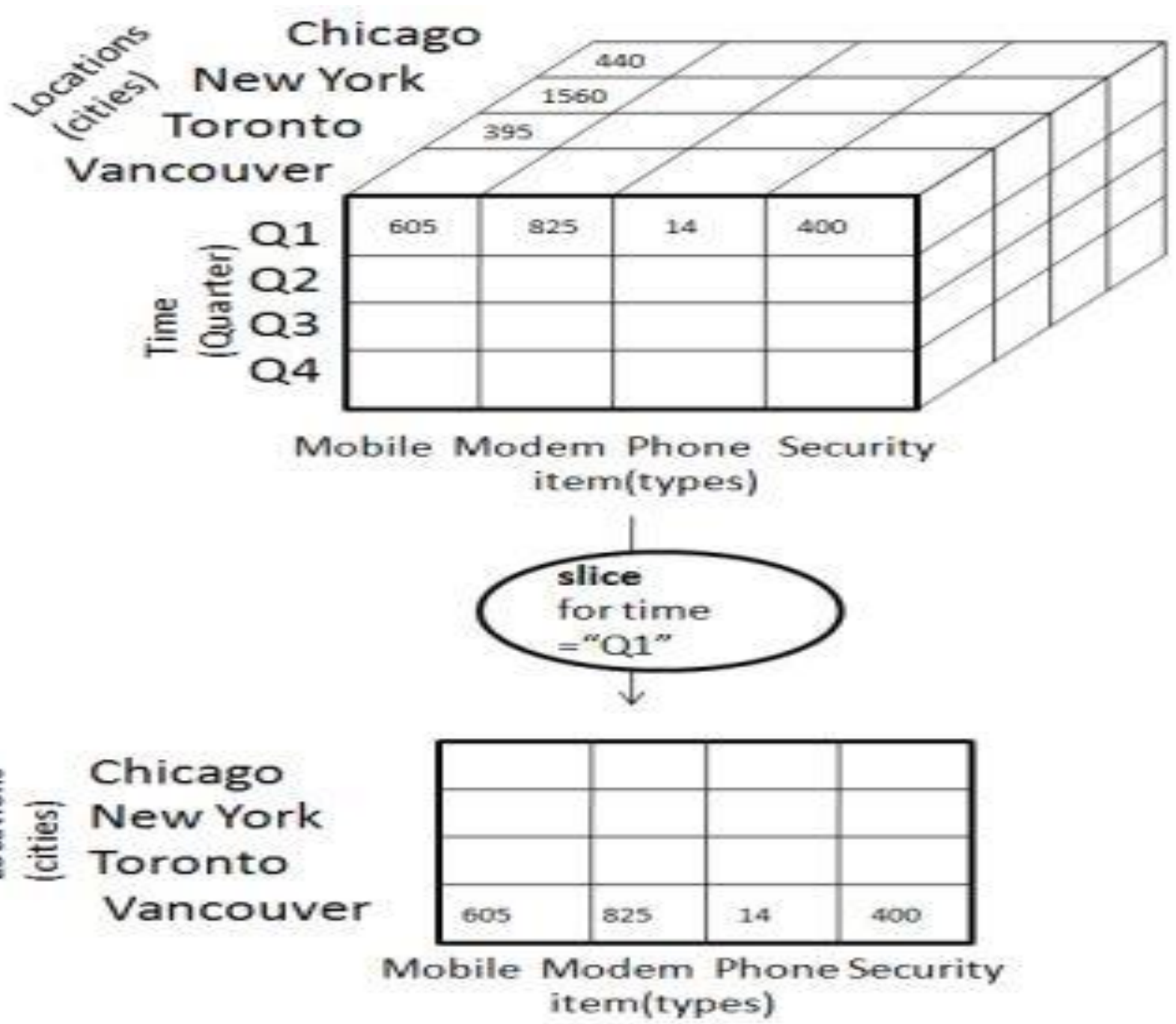
- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

# OLAP Operations

3. **Slice**

- The slice operation selects one particular dimension from a given cube and provides a new sub-cube.
- Consider the diagram in the next slide that shows how slice works.

# Slice


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

# OLAP Operations

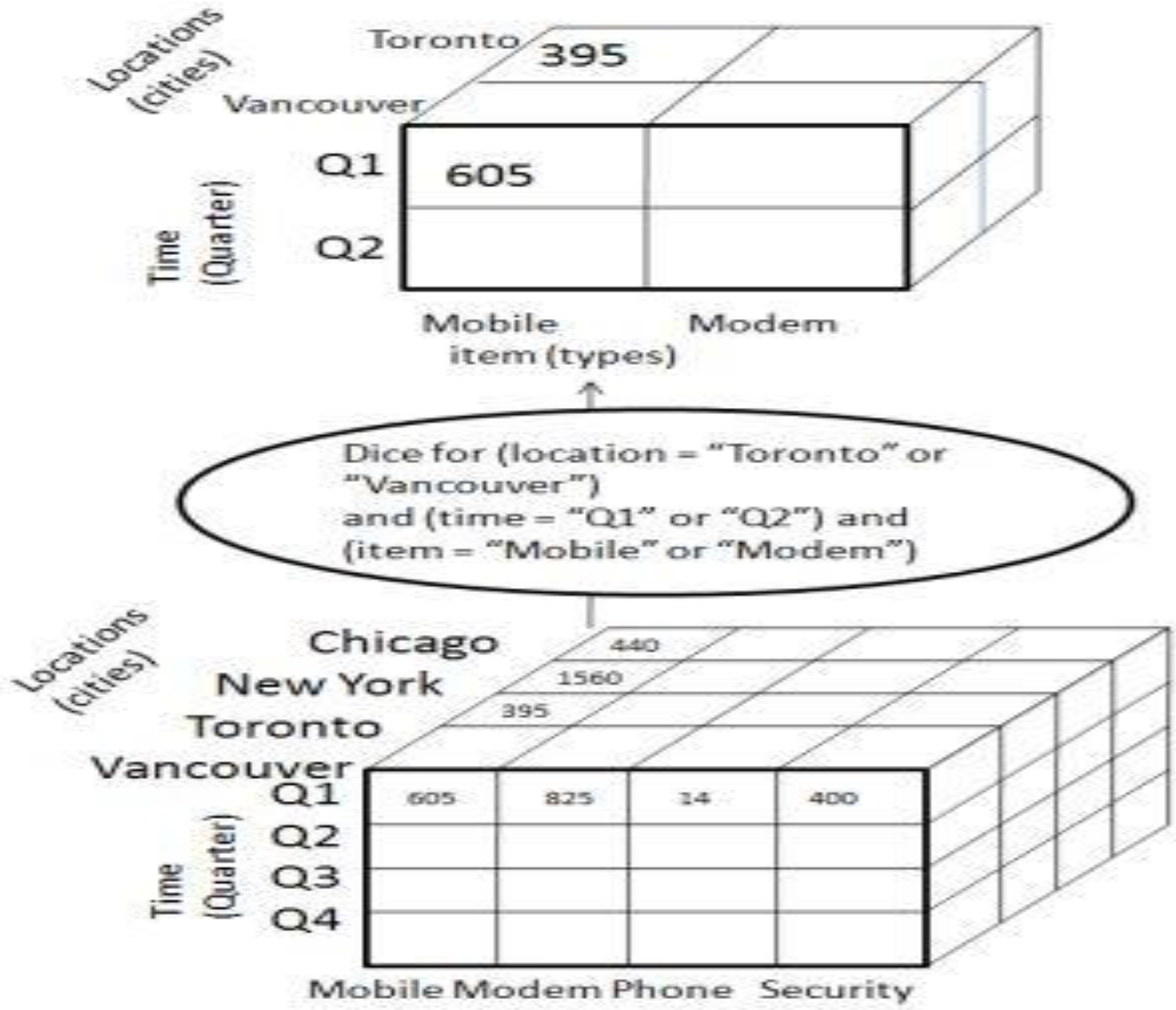
- The Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

# OLAP Operations

4. **Dice**

- Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- Consider the diagram in the next slide that shows the dice operation.

# Dice


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

# OLAP Operations

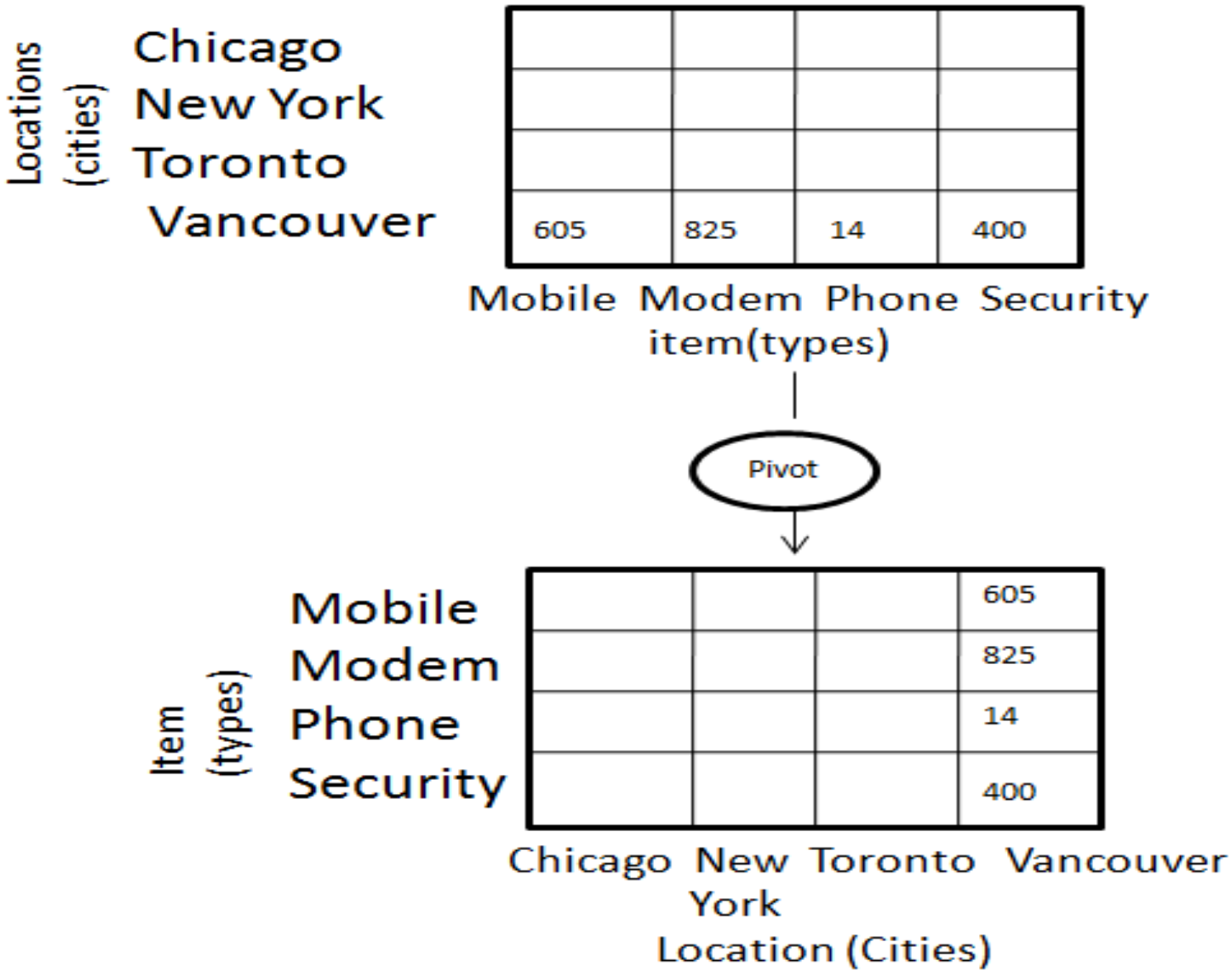
- The dice operation on the cube is based on the following selection criteria which involves three dimensions:
  - (location = "Toronto" or "Vancouver")
  - (time = "Q1" or "Q2")
  - (item =" Mobile" or "Modem")

# OLAP Operations

5. **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 diagram in the next slide that shows the pivot operation.

# Pivot


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)

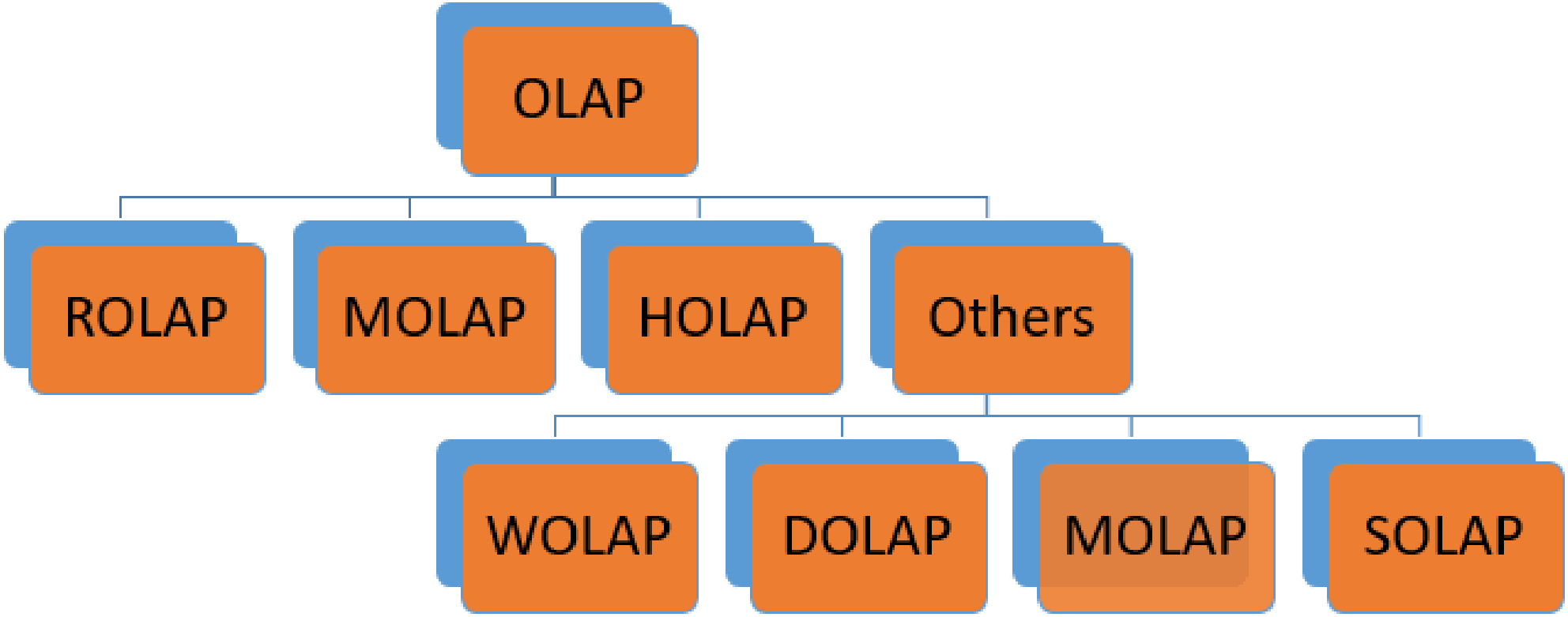
# OLAP Operations

- The 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.

# Types of OLAP Servers

- We have four types of OLAP servers −
  1. Relational OLAP (ROLAP)
  2. Multidimensional OLAP (MOLAP)
  3. Hybrid OLAP (HOLAP)
  4. Specialized SQL Servers

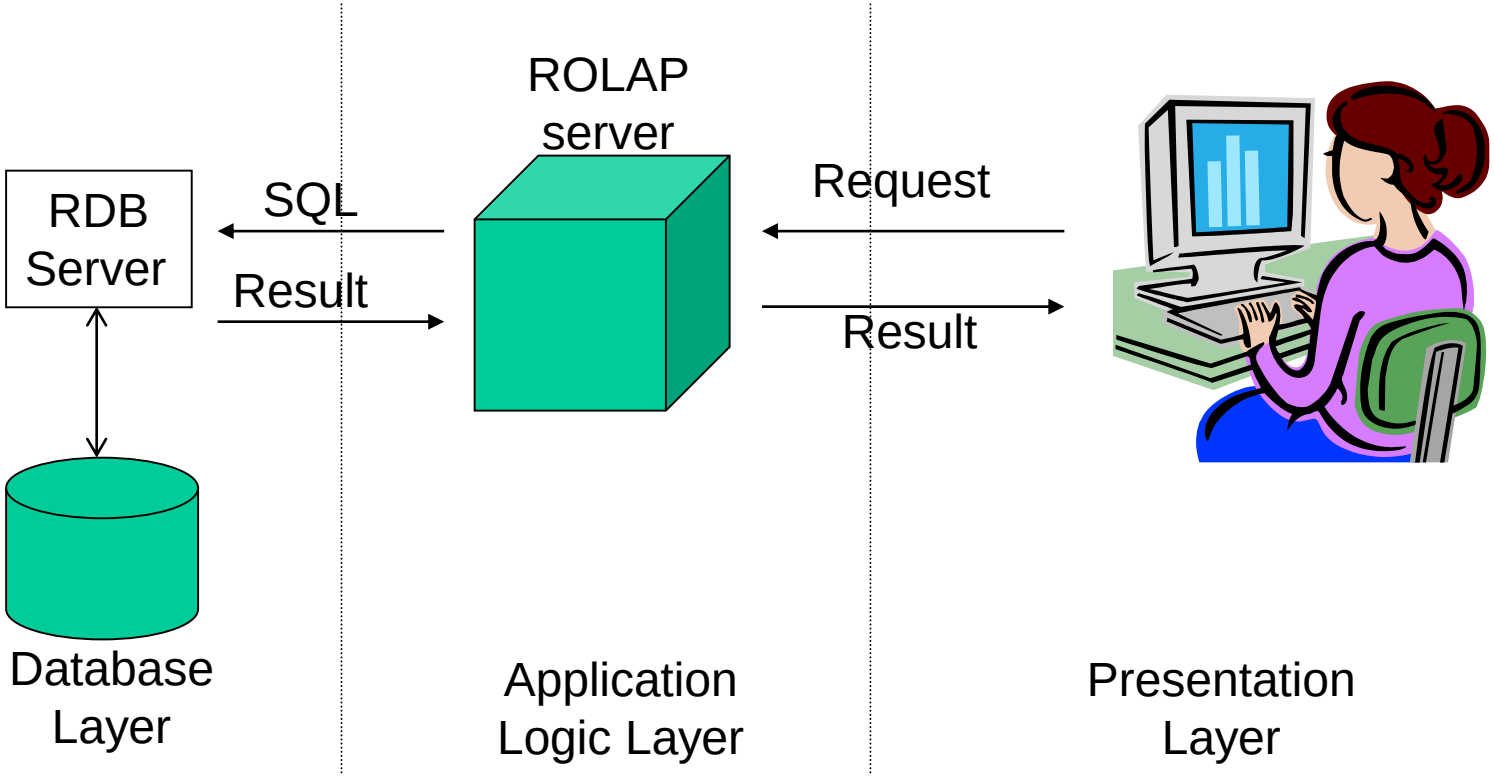
# Types of OLAP Servers



# Relational OLAP

- ROLAP is the fastest growing technology.
  - Works by providing multi-dimensional views of 2D data.
  - SQL is enhanced to increase performance and support complex operations on multi dimensions
- ROLAP servers are placed between relational back-end server and client front-end tools.
  - To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.
- ROLAP includes the following –
  - Implementation of aggregation navigation logic.
  - Optimization for each DBMS back end.
  - Additional tools and services.

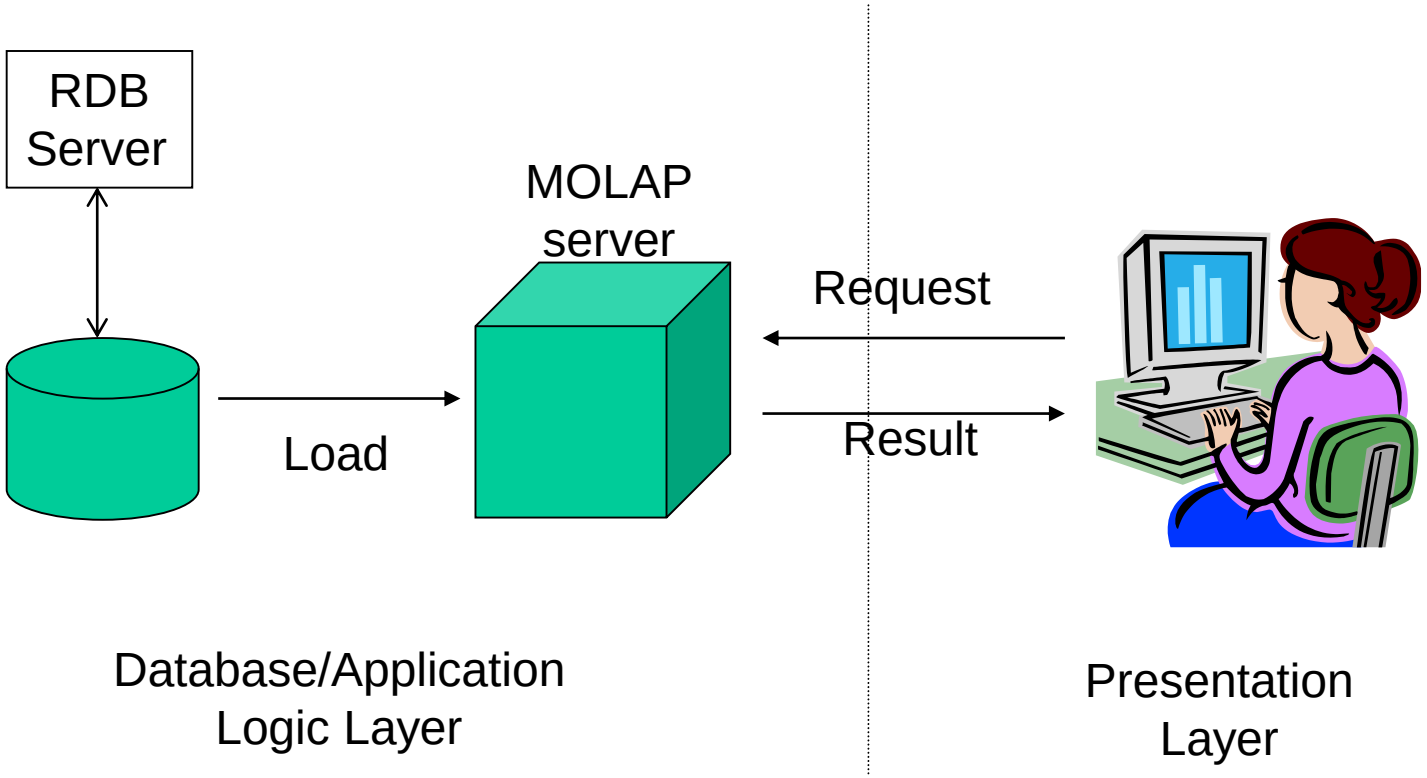
# ROLAP


RDB
Server
SQL
Result
ROLAP
server
Request
Result
Database
Layer
Application
Logic Layer
Presentation
Layer

# Multidimensional OLAP

- MOLAP uses array-based multidimensional storage engines for multidimensional views of data.
  - With multidimensional data stores, the storage utilization may be low if the data set is sparse.
  - Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

# MOLAP


RDB
Server
MOLAP
server
Request
Result
Load
Database/Application
Logic Layer
Presentation
Layer

# Hybrid OLAP

- Hybrid OLAP is a combination of both ROLAP and MOLAP.
  - It offers higher scalability of ROLAP and faster computation of MOLAP.
- HOLAP servers allows to store the large data volumes of detailed information.
- The aggregations are stored separately in MOLAP store.

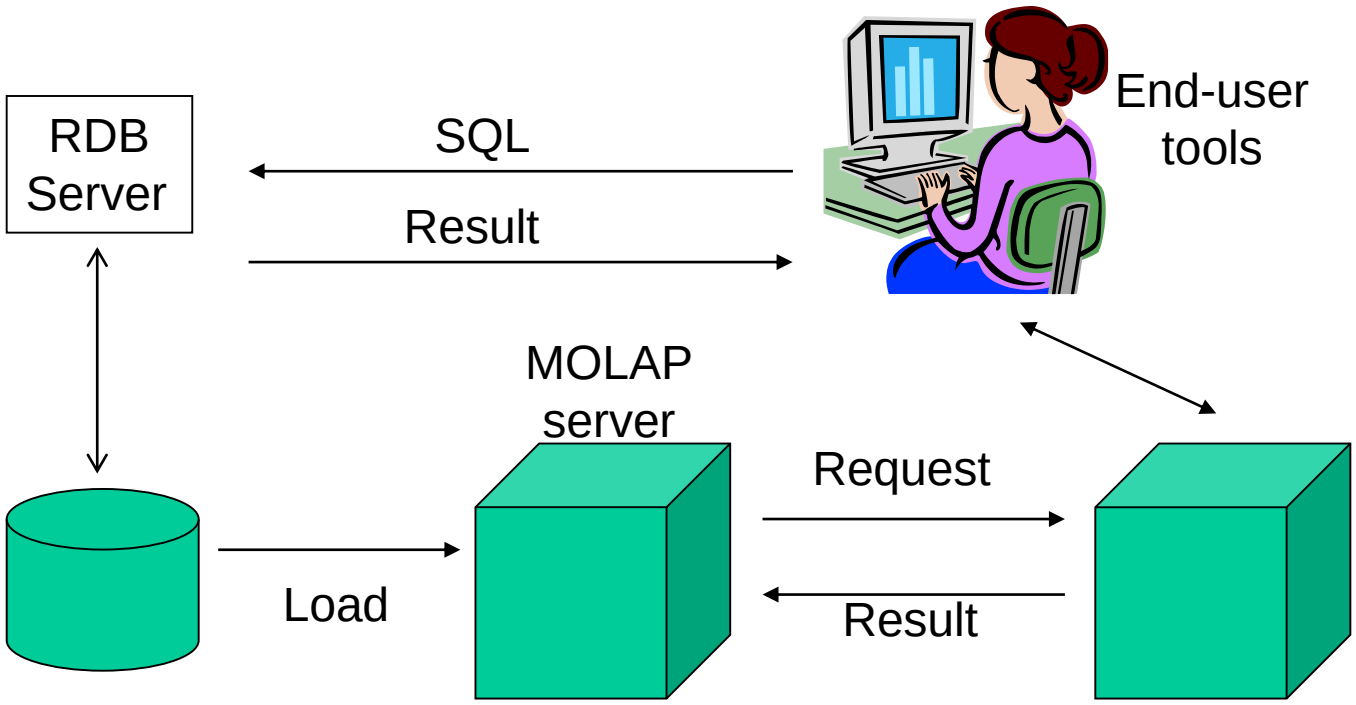
# Specialized SQL Servers

- Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

# Managed Query Environment (MQE)

- MQE is a newer technology.
  - Data can be delivered either directly from the RDB or from a MOLAP/ROLAP server in the form of a data cube.
  - The data cube is stored and analysed locally – therefore they are simple to install, and each user can build a custom data cube.

# MQE


RDB
Server
SQL
Result
End-user
tools
MOLAP
server
Request
Load
Result

# Real World Scenarion

## Casino

# Understand the desired objectives

Step 1: Determine business objectives

- Improve customer experience

Step 2: Collect appropriate data to help obtain your business objective

- Target the right customer

Step 3: Identify what success looks like

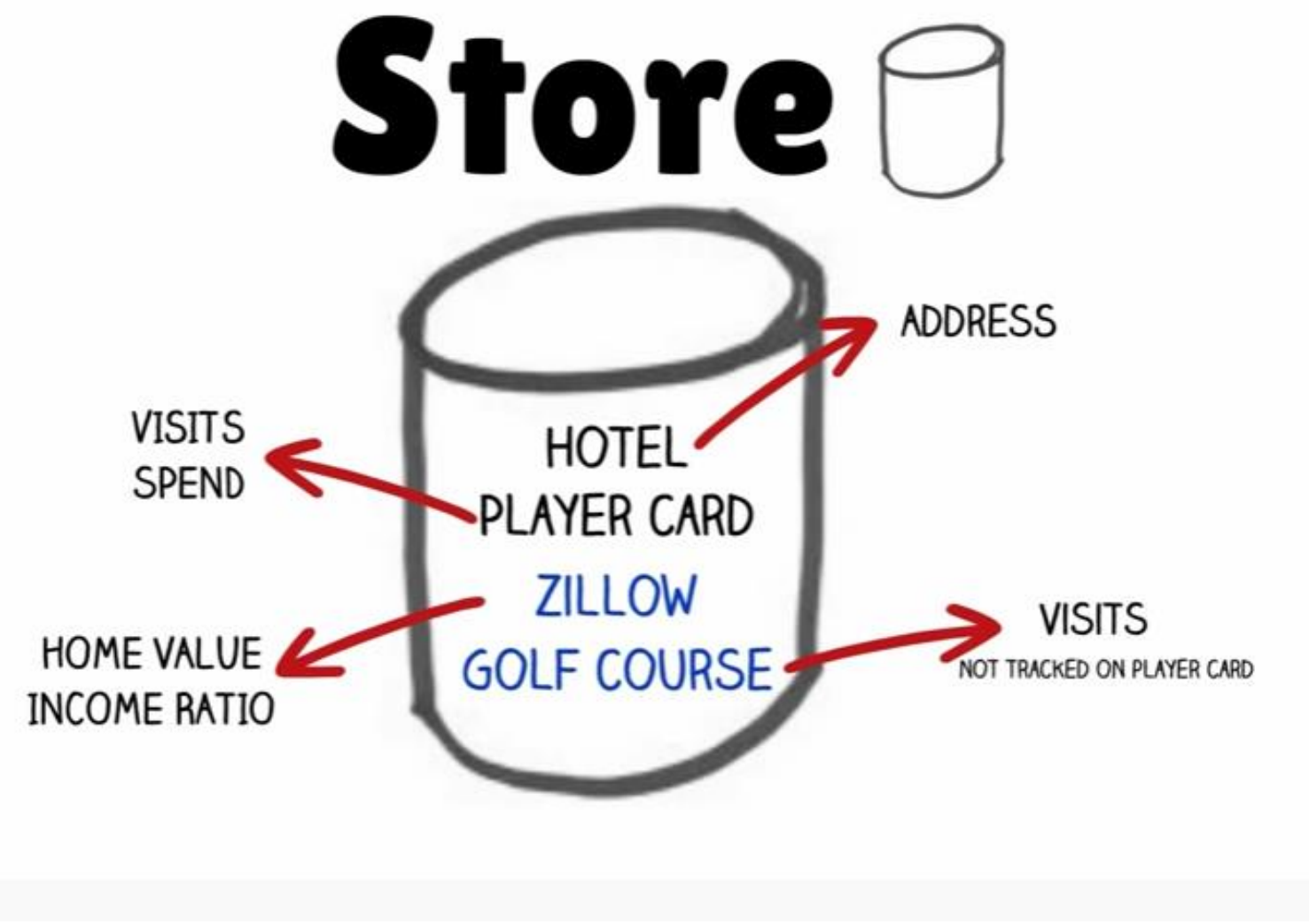
- Increase customer visits

# Collect the Right data about your customer

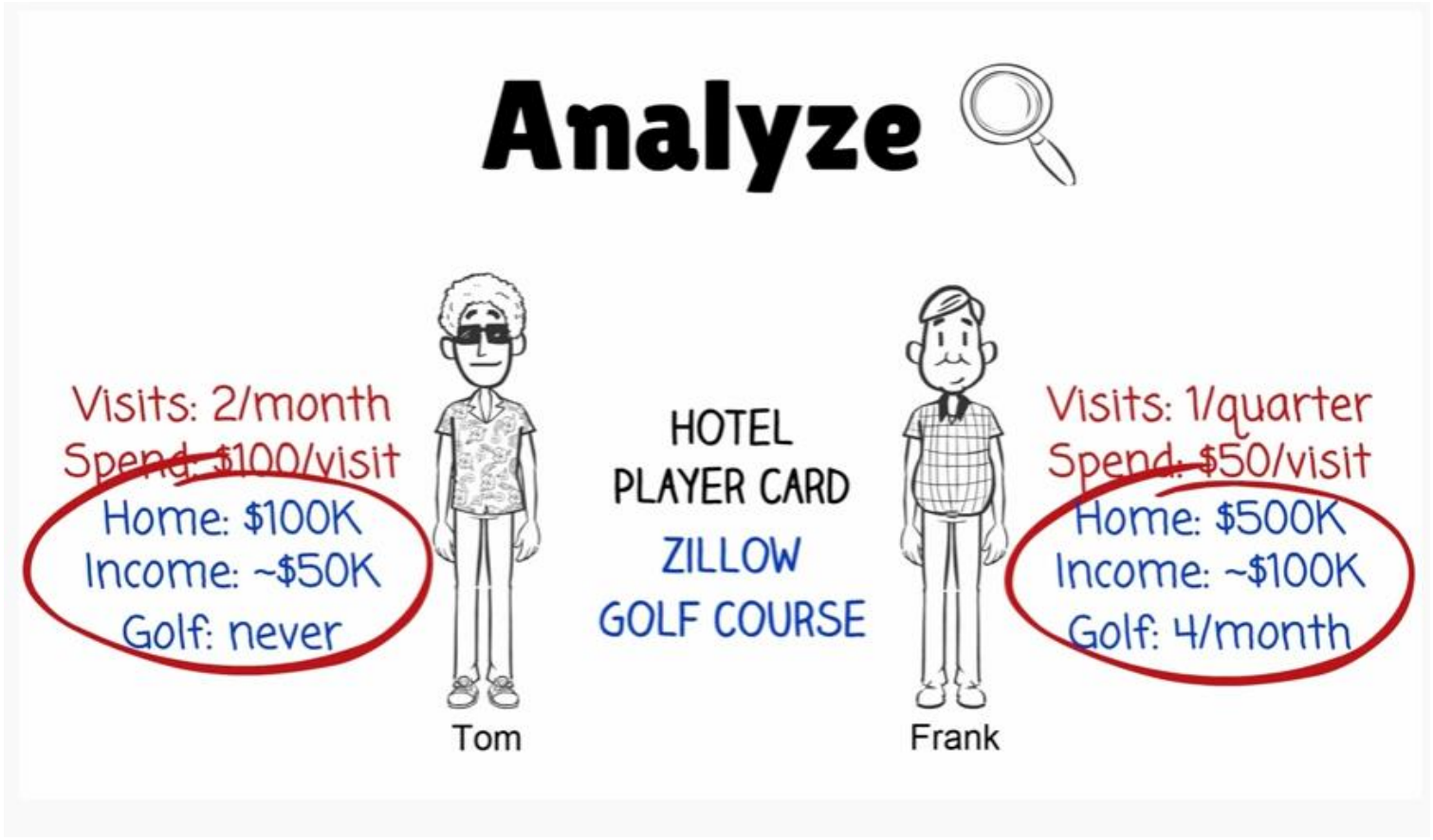
FACEBOOK
TWITTER
HOTEL
PLAYER CARD

ZILLOW
GOLF COURSE

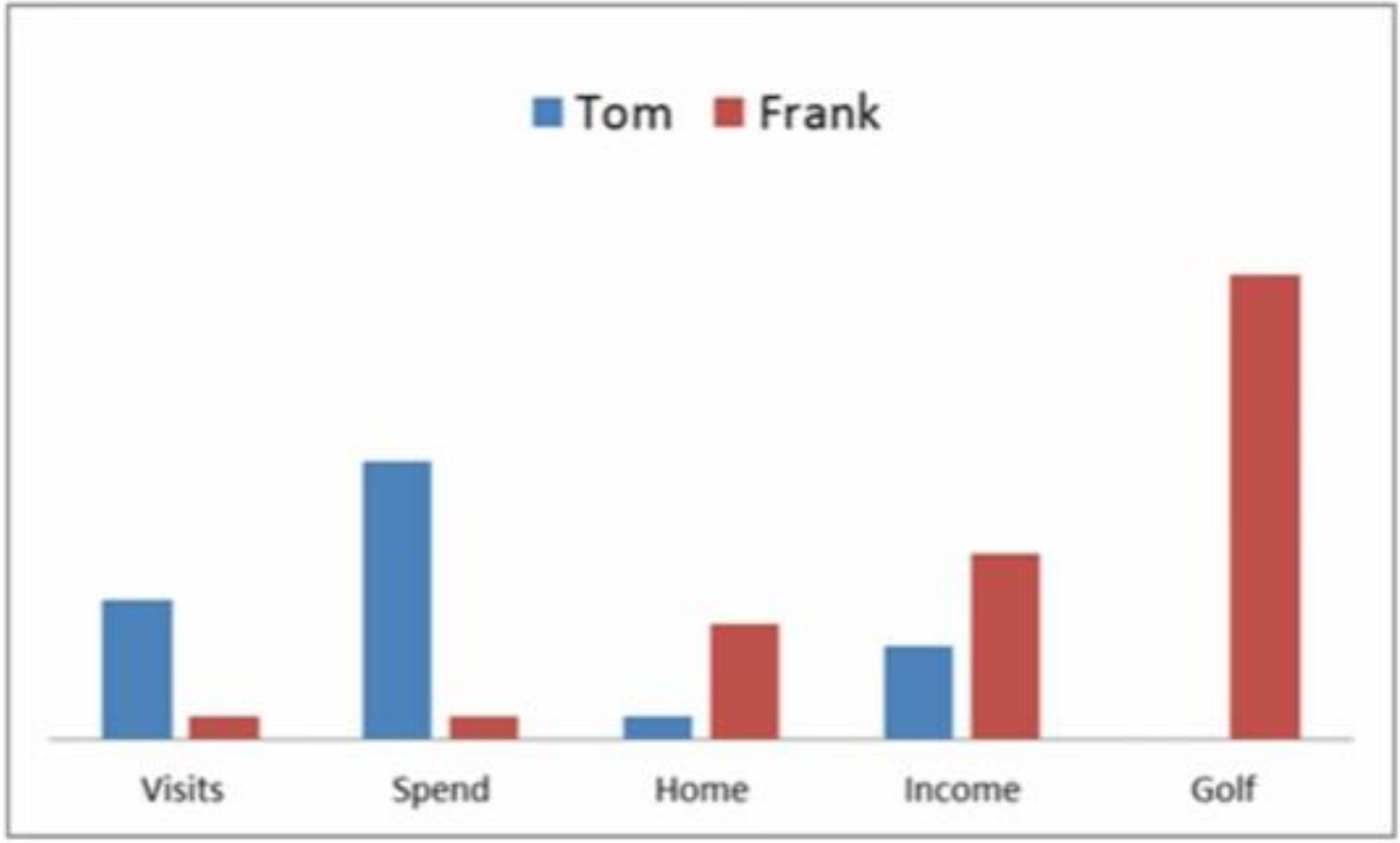
# Store the data e.g. in a data warehouse



# Analyse data to better understand their customer



# Visualize the Data to know the target customer



# Assignment

Enumerate the major differences between the following:

- Data lake and Big Data
- Data mining and Business Intelligence

# Further Reading

- Connolly and Begg, chapters 31 to 34.
- W H Inmon, *Building the Data Warehouse*, New York, Wiley and Sons, 1993.
- Benyon-Davies P, Database Systems (2nd ed), Macmillan Press, 2000, ch 34, 35 & 36.