University
Politehnica
of Bucharest

Faculty of
Automatic
Control and
Computers

Computer
Science and
Engineering
Department

Databases Management Systems. Data Warehouses. OLAP.

Ciprian-Octavian Truică
ciprian.truica@cs.pub.ro

Overview

- Database Management Systems
- Data Warehouses
- OLAP

Overview

- Database Management Systems
- Data Warehouses
- OLAP

Database Management Systems

- Database (DB) is an organized collection of data. The collection of data is organized in schemas, tables, reports, views and other object.
- Database Management Systems (DBMS) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyze data.
- A general-purpose DBMS is designed to allow the definition, creation, querying, update, and administration of databases.
- Source Wikipedia: https://en.wikipedia.org/wiki/Database

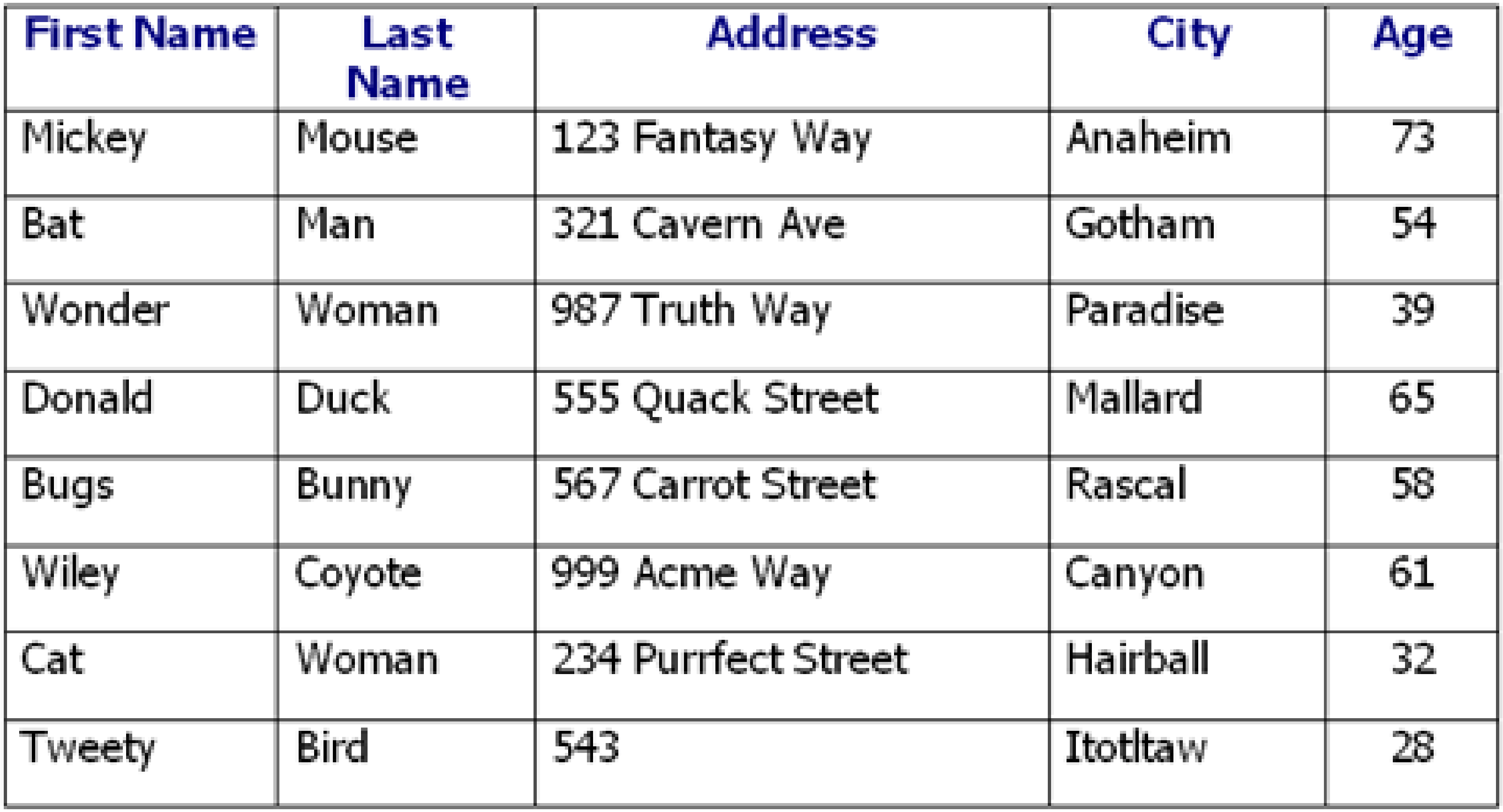
Database Management Systems

- There are multiple types of DBMS, usually classified by the way they store data:
  - Relational database management systems (RDBMSs): Oracle, Microsoft SQL Server, MySQL, PostgreSQL, IBM DB2
  - NoSQL DBMSs (not discussed):
    - Document-oriented DBMSs: MongoDB, Apache CouchDB
    - Graph DBMSs: Neo4J
    - Key-Value DBMSs: Riak, Redis
    - Object-oriented DBMSs: Caché, ObjectDB
    - Column-oriented DBMSs: Apache HBase, Cassandra

RDBMS

- Use the relational algebra model
- Store information into bi-dimensional tables
- Uses SQL (Standard Query Language) to query the data.
  - SQL is a set-oriented language

RDBMS



First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

RDBMS

- Relational algebra definitions
  - Relation r over a collection of sets (domain values) $D_1, D_2, \ldots, D_3$ is a subset of the Cartesian Product $D_1 \times D_2 \times \cdots \times D_n$
  - Thus, a relation r is a set of n-tuples $(d_1, d_2, \ldots, d_n)$ where $d_i \in D_i$
  - E.g.

StudId = {412, 307, 540}

StudName = {Smith, Jones}

Major = {CS, CSE, BIO }

Then r = {(412, Smith, CS), (307, Jones, CSE), (412, Smith, CSE)} is a relation over StudId x StudName x Major

RDBMS

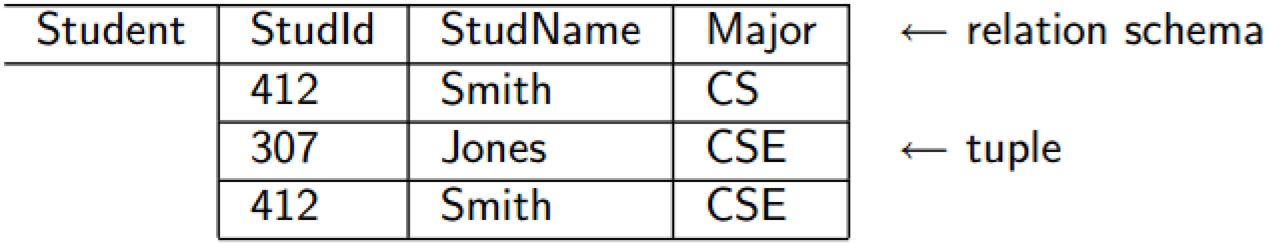
- Let $A_1, A_2, \ldots, A_n$ be attribute names with associated domain $D_1, D_2, \ldots, D_n$, then $R(A_1{:}\, D_1, A_2{:}\, D_2, \ldots, A_n{:}\, D_n)$ is a relation schema
- E.g.

Student(StudId : integer, StudName : string, Major: string)

- A relation schema specifies the name and the structure of the relation.
- A collection of relation schemas is called a relational database schema.

RDBMS

- A relation instance $r(R)$ of a relation schema can be thought of as a table with n columns and a number of rows.
- Note. Instead of a relation instance the term relation is used.
- An element $t \in r(R)$ is called a $tuple$ (row)
- A relation has the following properties:
  - The order of rows is irrelevant
  - There are no duplicate rows in the relationship



Student	StudId	StudName	Major
	412	Smith	CS
	307	Jones	CSE
	412	Smith	CSE

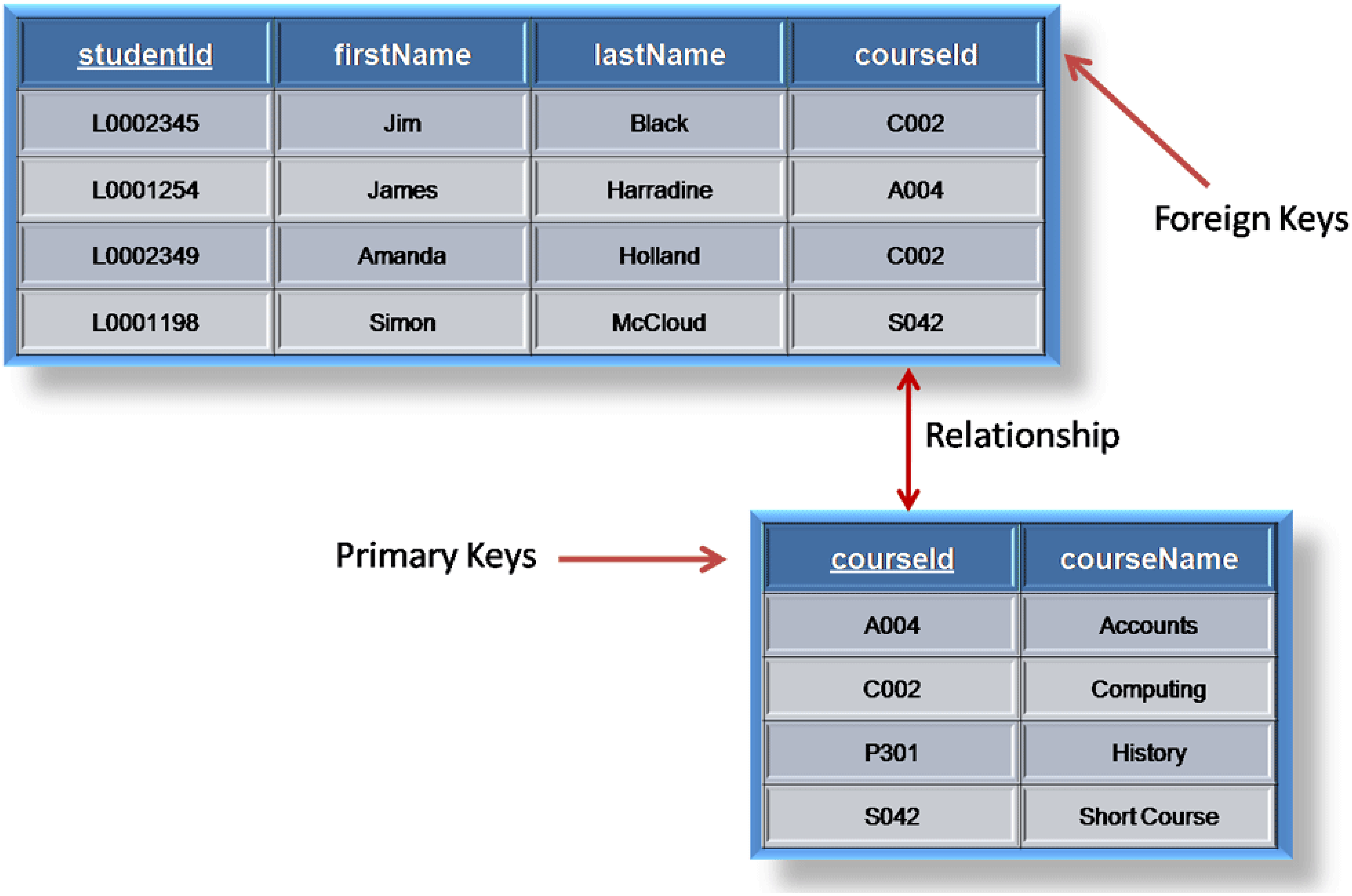
← relation schema

← tuple

RDBMS

- Integrity constraints:
  - Primary Key
    - The unique identifier of a tuple
    - Can be composed of one or more attributes.
  - Foreign Key
    - Set of attributes in one relation (child relation) that is used to “refer” to a tuple in another relation (parent relation).
    - Foreign key must refer to the primary key of the referenced relation.

RDBMS



studentId	firstName	lastName	courseId
L0002345	Jim	Black	C002
L0001254	James	Harradine	A004
L0002349	Amanda	Holland	C002
L0001198	Simon	McCloud	S042

Foreign Keys

Relationship

Primary Keys

courseId	courseName
A004	Accounts
C002	Computing
P301	History
S042	Short Course

RDBMS

- Relational algebra is used to formalize the query language
- Queries in relational algebra are applied to relation instances, result of a query is again a relation instance
- Operations
  - Set operations
  - Unary operations
  - Binary operations

RDBMS

- Given tow relations R and S
- Set operations:
  - Union is the operation that includes all the tuples that are either in R or in S, but duplicate tuples are removes, the end result being also a relation $T = R \cup S = \{ t_i \mid t_i \in \mathrm{R} \vee t_i \in S \}$
  - Union all is the union operation between two relations that keep the duplicates, the end result being also a relation, $T = R \uplus S = \{t_i \mid t_i \in R \vee t_i \in S\}$.
  - Intersection is the operation that includes only tuples that are both in R and S, the end result being also a relation, $T = R \cap S = \{t_i \mid t_i \in R \wedge t_i \in S\}$

RDBMS

- Set operations:
  - Difference is the operation that includes the tuples that are only in R and not in S, the end result being also a relation, $T = R - S = \{t_i \mid t_i \in R \wedge t_i \notin S\}$
  - Cartesian product is the operation that matches each tuple from relation R with each tuple from relations S, the end result being also a relation, $T = R \times S = \{t = (r, s) \mid r \in R, s \in S\}$. This operations is also known in literature as Cross Join or Cross Product.

RDBMS

r

A	B
a	1
a	2
b	1

s

A	B
a	2
b	3

$r \cup s$

A	B
a	1
a	2
b	1
b	3

$r \uplus s$

A	B
a	1
a	2
b	1
a	2
b	3

$r \cap s$

A	B
a	2

$r - s$

A	B
a	1
b	1

$r \times s$

r.A	r.B	s.A	s.B
a	1	a	2
a	2	a	2
b	1	a	2
a	1	b	3
a	2	b	3
b	1	b	3

RDBMS

- Unary operations are operations with only one operand applied on only one relation instance ($R = (r_1, r_2, \ldots, r_n)$)
  - Rename ($\rho_{\frac{a}{r_i}}(R), i = \overline{1, n}$) is an operation that, when applied to a relation R, returns the same result except that the attribute r_i is changed to a for all the tuples.
  - Projection ($\pi_{a_1, a_2, \ldots, a_p}(R)$) is an operation that restricts the set of attributes selected.

RDBMS

- Unary operations:
  - Selection ($\sigma_{\varphi}(R)$) is an operation that selects the number of tuples for which the propositional formula (condition) φ holds. The formal definition is $\sigma_{\varphi}(R) = \{r \mid r \in R, \varphi(r)\}$.
  - Grouping operator and aggregation functions ($\gamma_L(R)$) are used to return a single result using a group. L is a list that contains the group by attributes and the aggregation functions

RDBMS

r

A	B
a	1
a	2
b	1

s

A	B
a	2
b	3

$\rho_{\frac{C}{A}}(r)$

C	B
a	1
a	2
b	1

$\pi_A(r)$

A
a
a
b

$\sigma_{B=1}(r)$

A	B
a	1
b	1

$\gamma_{A,sum(B)}(r)$

A	B
a	3
b	1

RDBMS

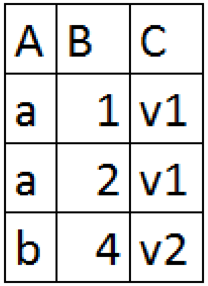
- Binary operations
  - Joins are binary operations that correlate, usually using a propositional formula (condition) θ, two relations.
    - Inner Joins
      - Cross Join
      - Natural Join
      - θ-Join
    - Outer Joins
      - Left Join
      - Right Join
      - Full Outer Join

RDBMS

- Inner join:
  - Natural join ($R \bowtie S = \{\ r \cup s \mid r \in\ R \wedge\ s \in\ S \wedge \phi(r \cup\ s)\ \}$) is a binary operation which result is the set of all combinations of tuples in R and S that are equal on their common attribute names. The predicative function ϕ verifies if the following conditions are true: $R.t_k\ =\ S.t_k, k\ =\ \overline{1, p}$.
  - θ-Join is a binary operation which result is the set of all combinations of tuples in R and S that respect a correlation condition between their attributes. Formally, $R\ \bowtie_\theta S = \{r \cup s \mid\ r \in\ R \wedge\ s \in\ S \wedge \phi_\theta(r \times\ s)\}$. In this case, $\theta\ = \{ <, \leq, =\ , \geq, > \}$ is a binary relational operator and the predicative function ϕ_θ verifies if following conditions are true $R.t_i'\theta S.t_j''$, with $i\ = j$. If $\theta\ = \{=\}$ then the join becomes an equi-join.

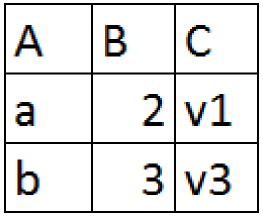
RDBMS

r



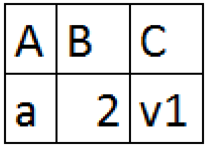
A	B	C
a	1	v1
a	2	v1
b	4	v2

s



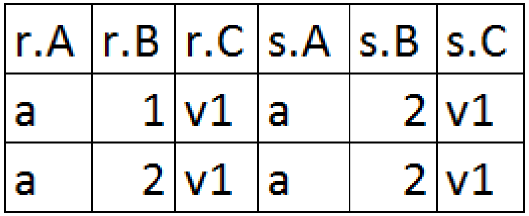
A	B	C
a	2	v1
b	3	v3

$r \bowtie s$



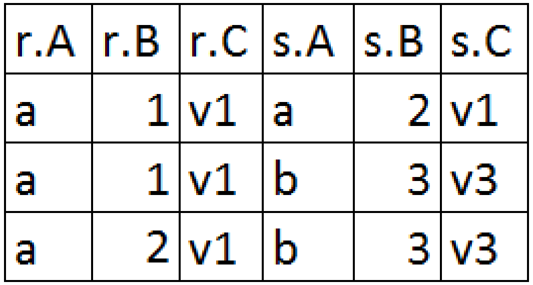
A	B	C
a	2	v1

$r \bowtie_{r.C=s.C} s$



r.A	r.B	r.C	s.A	s.B	s.C
a	1	v1	a	2	v1
a	2	v1	a	2	v1

$r \bowtie_{r.B<s.B} s$



r.A	r.B	r.C	s.A	s.B	s.C
a	1	v1	a	2	v1
a	1	v1	b	3	v3
a	2	v1	b	3	v3

RDBMS

- Outer joins are binary operations that extend the Inner Join operations by adding new values when not matching tuples between the relations are not found.
  - Left Outer Join $R⟕S = (R \bowtie S) \cup \left(\left(R - \pi_{r_1,\dots,r_n}(R \bowtie S)\right) \times \{(\omega,\ldots,\omega)\}\right)$
  - Right Outer Join $R⟖S = (R \bowtie S) \cup \left(\{(\omega,\ldots,\omega)\} \times \left(S - \pi_{s_1,\dots,s_n}(R \bowtie S)\right)\right)$
  - Full Outer Join R$⟗S = R⟕S \cup R⟖S$

RDBMS

r

A	B	C
a	1	v1
a	2	v1
b	4	v2

s

A	B	C
a	2	v1
b	3	v3

r⟕$_{r.C=s.C}s$

r.A	r.B	r.C	s.A	s.B	s.C
a	1	v1	a	2	v1
a	2	v1	a	2	v1
b	4	v2	ω	ω	ω

r⟖$_{r.C=s.C}s$

r.A	r.B	r.C	s.A	s.B	s.C
a	1	v1	a	2	v1
a	2	v1	a	2	v1
ω	ω	ω	b	3	v3

r⟗$_{r.C=s.C}s$

r.A	r.B	r.C	s.A	s.B	s.C
a	1	v1	a	2	v1
a	2	v1	a	2	v1
b	4	v2	ω	ω	ω
ω	ω	ω	b	3	v3

RDBMS

- Normalization is the process of dividing the data into multiple tables, so that data redundancy and data integrities are achieved.
- De-Normalization is the opposite process of normalization where the data from multiple tables are combined into one table, so that data retrieval will be faster.

Overview

- Database Management Systems
- Data Warehouses
- OLAP

Data warehousing

- Data warehouse is a repository of an organization's electronically stored data.
- Data warehouses are designed to facilitate reporting and analysis.
- A data warehouse stores a standardized, consistent, clean and integrated form of data sourced from various operational systems in use in the organization, structured in a way to specifically address the reporting and analytic requirements.

Data warehousing

R. Kimball (see [Kimball 2002]):

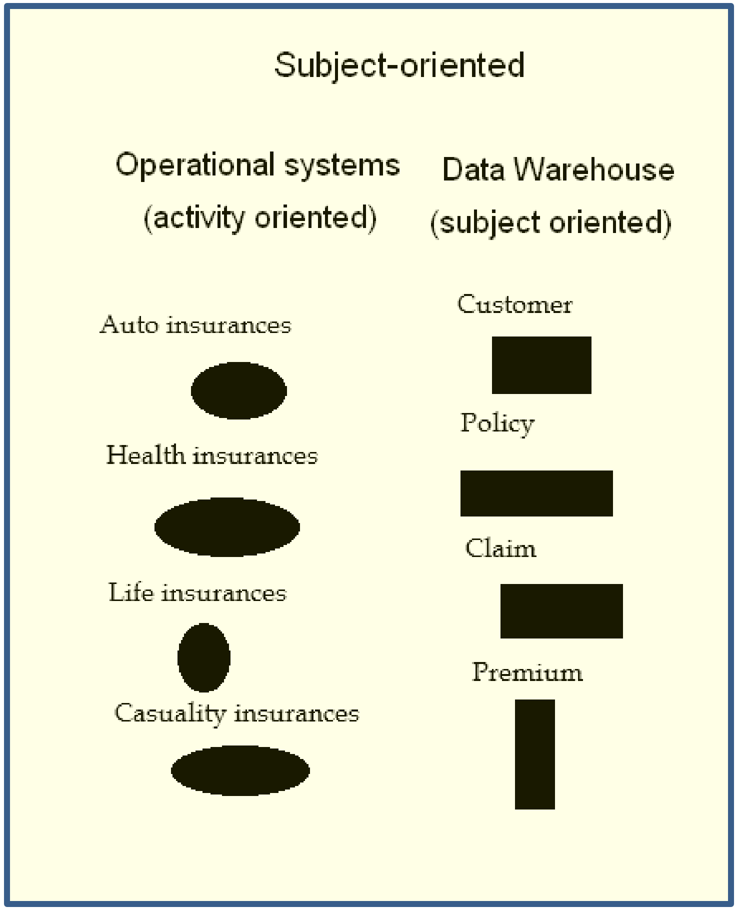
- A data warehouse is a copy of transactional data specifically structured for querying and analysis.
- According to this definition:
  - The form of the stored data (RDBMS, flat file) is not linked with the definition of a data warehouse.
  - Data warehousing is not linked exclusively with "decision makers" or used in the process of decision making.

Data warehousing

- **W.H. Inmon ([Inmon 2002]):**
- A data warehouse is a: subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management’s decisions.
- The data warehouse contains granular corporate data.

Data warehousing

- Operational data systems of a company are organized considering the main activities, so they are activity-oriented and not subject oriented.
- A classical example in the literature is an insurance company where the ***main activities*** are auto insurances, health insurances, life insurances and casualty insurances.



Data warehousing

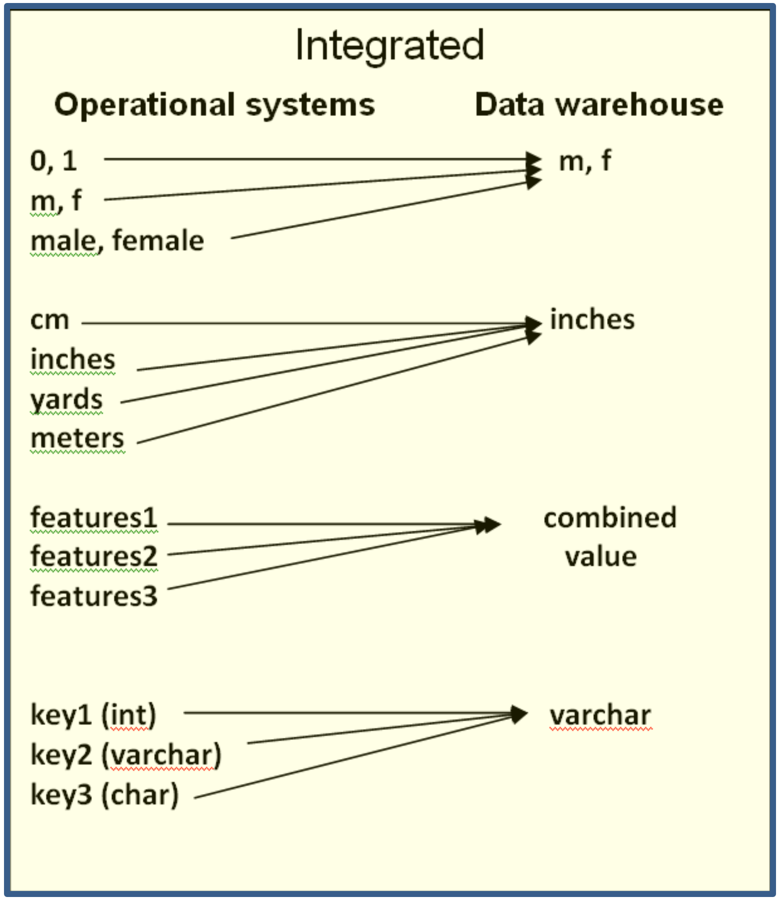
- Subject-oriented
  - For each activity there is possibly another software system managing data on the main ***subject areas***: policies, customers, claims and premiums in the area, so there are possible four separate databases, one for each activity, with similar but not identical structures.
  - When uploading data in the company data warehouse, the data must first be ***restructured*** on these major subject areas, integrating data on customers, policies, claims and premiums from each activity.

Data warehousing

- Integrated
  - When preparing data for uploading in the data warehouse, one of the most important activities is the integration.
  - Data is loaded from operational sources and must be converted, summarized, re-keyed, etc., before loading it in the data warehouse.

Data warehousing

- Integrated
  - Combine multiple encodings in a single one.
  - For example, the gender may be encoded as (0, 1), (m, f), (male, female) in separate operational systems. If (m, f) is chosen as the data warehouse encoding, all data encoded using other convention must be converted.



Integrated

Operational systems

Data warehouse

0, 1

m, f

male, female

m, f

cm

inches

yards

meters

inches

features1

features2

features3

combined value

key1 (int)

key2 (varchar)

key3 (char)

varchar

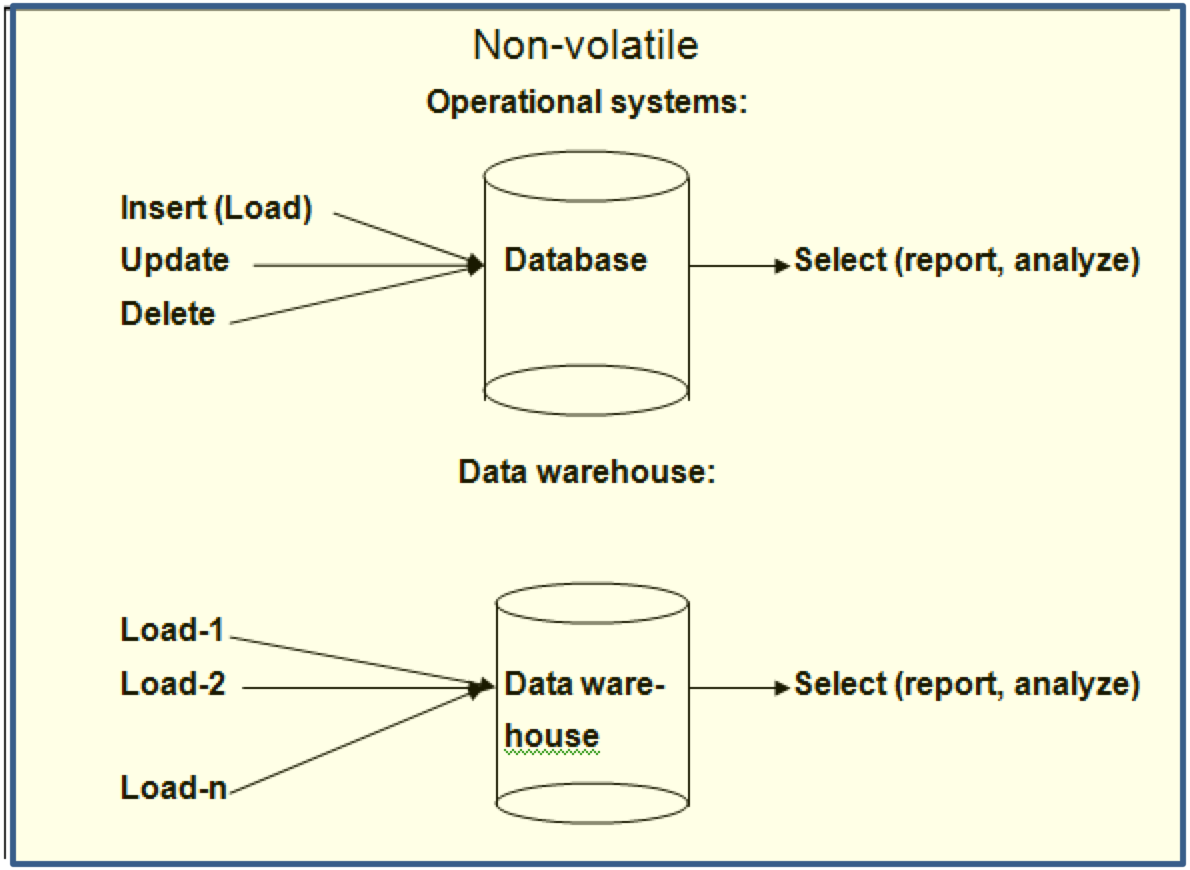
Data warehousing

- Actions performed for data integration:
  - Chose a unique measure unit for each piece of information. For example, if length is measured in cm, inches, yards and meters in different operational systems, one unit must be chosen for the data warehouse and all other values must be converted.
  - If the same object has in some data sources different values for the same attribute (e.g. description, name, features, etc.), these must be combined in a single one.
  - If the same object has different keys in the source systems it must be re-keyed to have a single key in the data warehouse.

Data warehousing

- Non-volatile:
  - In usual operational systems data is updated or deleted to reflect the actual values.
  - In a data warehouse data is never updated and deleted: after data is loaded, it stays there for future reporting, like a snapshot reflecting the situation in a certain moment.
  - The next load operations, instead of changing the old snapshots, are added as new snapshots and so the data warehouse is a sequence of such snapshots that coexist.

Data warehousing



Non-volatile

Operational systems:

Data warehouse:

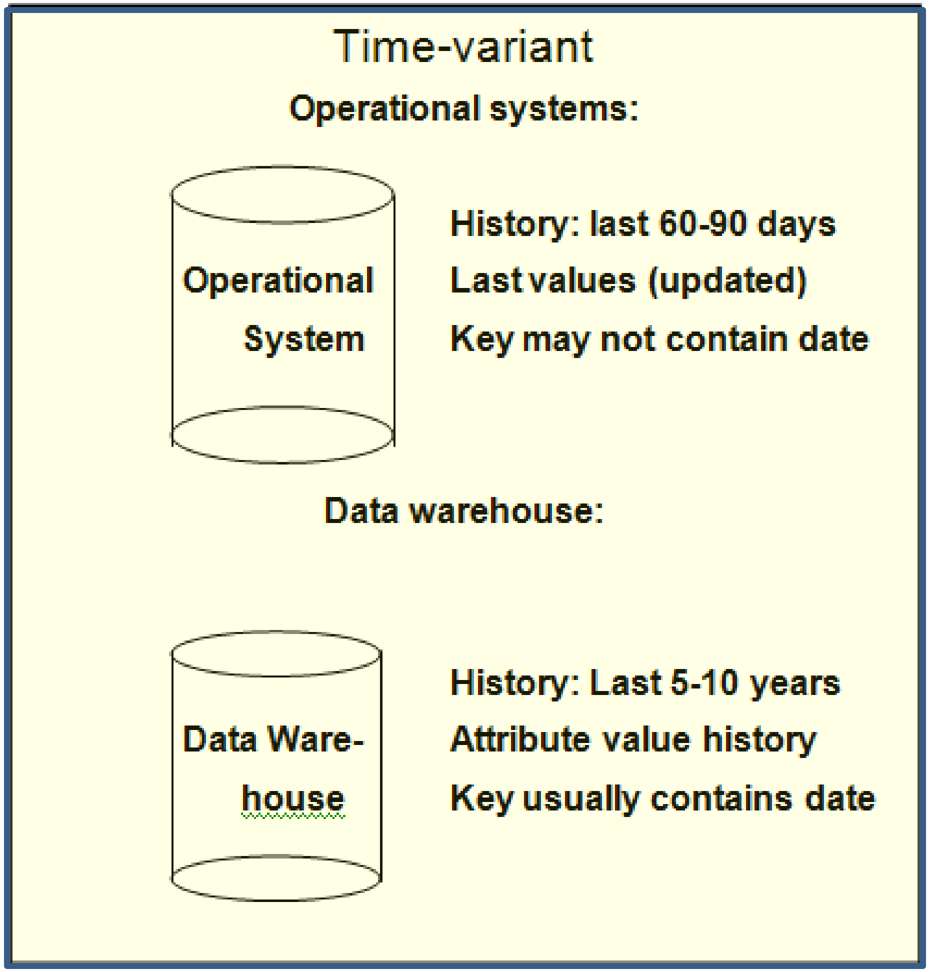
Data warehousing

- Non-volatile:
  - In this way the data warehouse contains not the operational data at a given moment but all the history of operational data.
  - Because of this lack of change, once loaded, data in a data warehouse may be considered as read-only.

Data warehousing

- Time variant
  - As described above, a data warehouse contains a sequence of snapshots, each snapshot being actual at a given moment of time.
  - Because a DW contains the whole history of a company, it is possible to retrieve information from a time window (slice).
  - Each unit of information is stamped or linked with the moment during which that information was accurate.

Data warehousing



Time-variant

Operational systems:

Operational System

History: last 60-90 days

Last values (updated)

Key may not contain date

Data warehouse:

Data Ware-house

History: Last 5-10 years

Attribute value history

Key usually contains date

Data warehousing

- Time variant:
  - In an operational system only the current data is kept. For example, if a customer changes address, in the operational system old address is replaced (update) with the new one.
  - In the data warehouse all successive addresses of a customer are kept.
  - Because date and time are very important in analyzing data and reporting, the key structure contains usually the date and sometimes the time.

Data warehousing

- Requirements for a DW [Kimball 2002]:
  - Information must be easy accessible
  - Information must be consistent
  - Flexibility
  - Security
  - Decision Support
  - User Acceptance

Data warehousing

- Why build a DW? [Kimball 2002]
  - “We have mountains of data in this company, but we can’t access it.”
  - “We need to slice and dice the data every which way.”
  - “You’ve got to make it easy for business people to get at the data directly.”
  - “Just show me what is important.”
  - “It drives me crazy to have two people present the same business metrics at a meeting, but with different numbers.”
  - “We want people to use information to support more fact-based decision making.”

Data warehousing

- Information must be easy accessible:
  - DW content must be understandable.
  - DW content must be intuitive or obvious to the non-database specialists, because they are the key users of the system.
  - Names must be meaningful (for data categories, features, attributes and so on, so the structure of the DW must be understandable for a non-specialist user).
  - The DW must provide options for combining data in the DW, the process being known and referred to as *slicing and dicing.*
  - The methods and tools for accessing data in the data warehouse must be simple, easy to use, and the answer must be returned in a short time.

Data warehousing

- Information must be consistent:
  - The process of fueling a data warehousing with data contains a step of preprocessing, where data is assembled from many sources, cleansed, quality assured. Data is released (published) to the users only when it is fit for usage.
  - As described earlier, an integration step is performed when data is load from operational sources, unifying encodings, units of measure, keys, names and common values/features, etc.
  - Common definitions for the contents of the data warehouse must be available for DW users.

Data warehousing

- Flexibility:
  - A data warehouse must be designed to be flexible considering the inevitable changes in computer science and engineering.
  - Its content must be structured in such a way that changes in the software and hardware platform must be possible.
  - Adding new data, reports, queries must be possible and must not interfere with existing ones.

Data warehousing

- Security
  - Because of its confidential content, the data warehouse must have the means for rejecting unauthorized access.
  - Potential leaks of content may be harmful for the company if competitors have access to the data in the DW.

Data warehousing

- Decision support:
  - The primary goal of implementing a data warehouse in an organization is the decision support
  - The ultimate output from a DW is the set of decisions based on its content, analyzed and presented in different ways to the decision makers.
  - The original label for a data warehouse and the tools around it was ‘decision support system’.

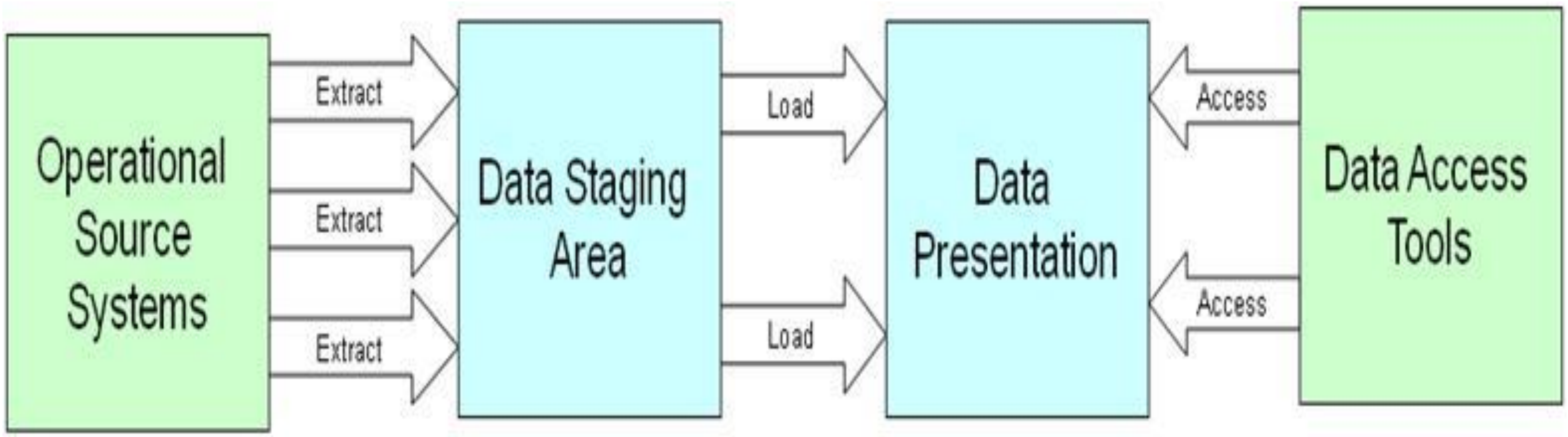
Data warehousing

- Acceptance
  - The ultimate test for the success in implementing a data warehouse is the acceptance test.
  - If the business community does not continue to use it in the first six months after training, then the system has failed the acceptance test, no mater how bright is the technical solution.
  - It is possible to ignore using it because decisions may be adopted also without a decision support system.
  - Key point in user acceptance is simplicity and user friendliness.

DW architecture

- ❑ Operational Source Systems. These are the source of the data in the DW, and are placed outside of the data warehouse
- ❑ Data Staging Area. Here data is prepared (transformed) for loading in the presentation area. This area is not accessible to the regular user.
- ❑ Data Presentation. This part is what regular users see and consider to be a DW.
- ❑ Data Access Tools. These tools are used for analyzing and reporting. They provide the interface between the user and the DW.

DW architecture



DW architecture

- The **data staging area** (DSA) of a data warehouse is [Kimball 2002] :
  - A storage area
  - A set of processes performing the so-called Extract-Transform-Load (ETL) operation:
    - Extract – Extracting data from Operational Source Systems
    - Transform – Integrating data from all sources
    - Load – Publishing data for users, meaning loading data in the Data presentation area

DW architecture

- DSA contains everything between the operational source systems and the data presentation area.
- This area is not accessible to the regular users of the data warehouse.
- Storing data in a DW (so also in DSA) may be done following two main approaches:
    1. The normalized approach ([Inmon 2002])
    2. The dimensional approach ([Kimball 2002])
- These approaches are not mutually exclusive, and there are other approaches.
- Dimensional approaches can involve normalizing data to a degree.

DW architecture

- ***Normalized approach***
  - In the ***normalized approach***, data are stored following database normalization rules.
  - Tables are grouped by subject areas (data on customers, policies, claims and premiums for example).
  - The main ***advantage*** of this approach is that loading data is straightforward because the philosophy of structuring data is the same for operational source systems and the data warehouse.

DW architecture

- ***Normalized approach***
  - The main ***disadvantage*** of this approach is the number of joins needed to obtain meaningful information.
  - A regular user needs also to have a good knowledge about the data in the DW and also a training period in obtaining de-normalized tables from normalized ones.
  - Missing a join condition when performing a query may lead to Cartesian products instead of joins. In other words, regular user may need assistance from a database specialist to perform usual operations.

DW architecture

- ***Dimensional approach***
  - Data is partitioned in two main categories:
    - Facts (numeric transaction data). In a retail example, the fact table contains quantity sold, total price, total cost, total gross profit.
    - Dimensions (standardized contexts for facts). In a retail example, dimensions may be: product, date, time, location, customer, salesperson, etc.
    - Main approach OLAP

DW architecture

- ***Advantages*** of the dimensional approach are:
  - Data is easy to understand, easy to use, no need for assistance from a database specialist, speed in solving queries.
  - Data being de-normalized (or partially de-normalized) the number of joins needed for performing a query is lower than in the normalized approach.
  - Joins between the fact table and its dimensions is easy to perform because the fact table contains surrogate keys for all involved dimension tables.

DW architecture

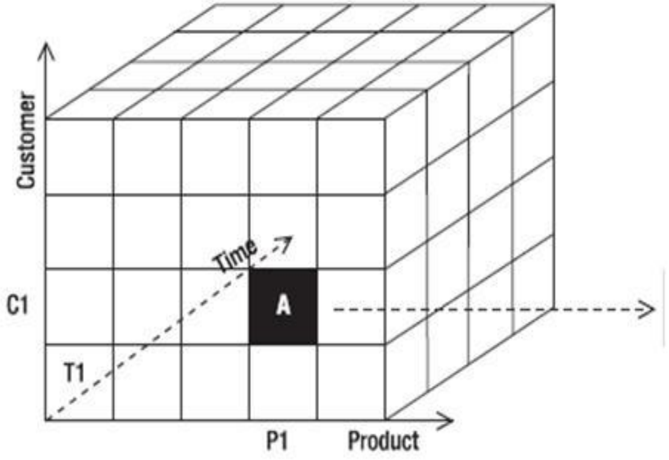
- ***Disadvantages*** of dimensional approach:
  - The ETL process is harder to be performed because of the different philosophy in structuring data in the operational systems and the data warehouse: transform and load steps are more complicated than in the normalized approach.
  - A second disadvantage is that is more difficult to modify the data warehouse scheme when the company changes its way to do business.

DW architecture

- Data presentation area:
  - At the end of the ETL process prepared data is loaded in the Data Presentation Area (DPA).
  - After that moment, data is available for users for querying, reporting and other analytical applications.
  - Because regular users have access only to that area, they may consider the presentation area as being the data warehouse.
  - This area is structured as a series of integrated **data marts**, each presenting the data from a single business process.

DW architecture

- Data presentation area:
  - In the DPA data is stored, presented, and accessed in dimensional schemas.
  - We can imagine a hypercube with edges labeled with the dimensions, e.g. customer, product and time.



DW architecture

- Data access tools:
  - Almost all DW regular users (80% to 90%) will access the data via some prebuilt parameter-driven analytic applications.
  - Generally a user has four channels to interact with a DW:
    - Ah-hoc query tools.
    - Report writers
    - Analytic applications.
    - Modeling tools.

DW architecture

- Ah-hoc query tools:
  - By this channel the user obtains raw data verifying the conditions specified in the ad-hoc query.
  - For using this channel the user must have a good knowledge on the DW structure and on query language used.
  - This channel is for specialists and experienced users.

DW architecture

- Report writers:
  - This channel is at the same level as the first one.
  - Raw data is presented as a report.
  - Usually there are several pre-built reports that user may run without knowing the DW structure and query language.
  - Building new reports may need extra abilities.

DW architecture

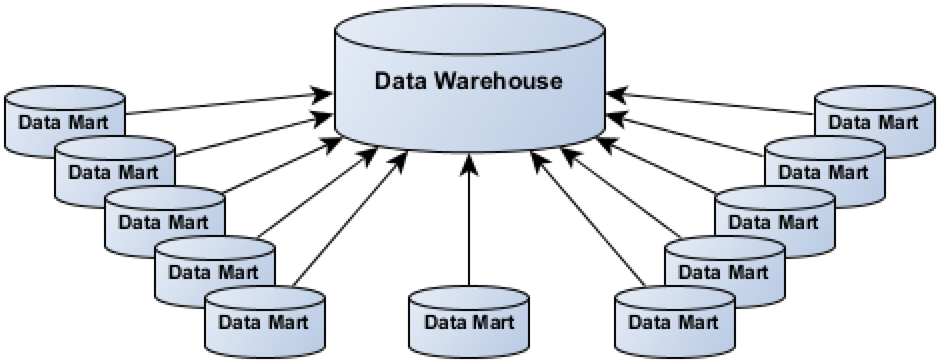
- Analytic applications
  - Interactive reports
  - Dashboards
  - Scorecards
  - Other reporting tools allowing users to access and analyze on data in the DW.

DW architecture

- Modeling tools:
  - In this category can be mentioned data mining products, forecasting and scoring tools.
  - At this level the result is not only a sophisticated report on existing data but also extracted new knowledge, models for forecasting and other outputs providing new knowledge to the user.

DW architecture

- A ***data mart*** is defined as a repository of data gathered from operational data and other sources that is designed to serve a particular community of knowledge workers.
- Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization.
- ***The data warehouse is created from the union of organizational data marts***.



Overview

- Database Management Systems
- Data Warehouses
- OLAP

OLAP

- OLAP (On-Line Analytical Processing) is a technology used to organize information in Data Warehouses which offers support for knowledge discovery useful in the decision making process

OLAP

- Concepts:
  - Hierarchy is a logical tree-type structure used to organize the members of a dimension
  - Level – the information from a hierarchy can be organized using different inferior or superior levels
  - Member is an element in a hierarchy
  - Calculated Member is a member in a hierarchy whose value is computed during execution
  - Dimension is a set that contain one or more hierarchies
  - Cube is a data structure that aggregates the measures by the levels and hierarchies of each analyzed dimension
  - Measure is a set of values in a cube based on a column in the Facts table.

OLAP

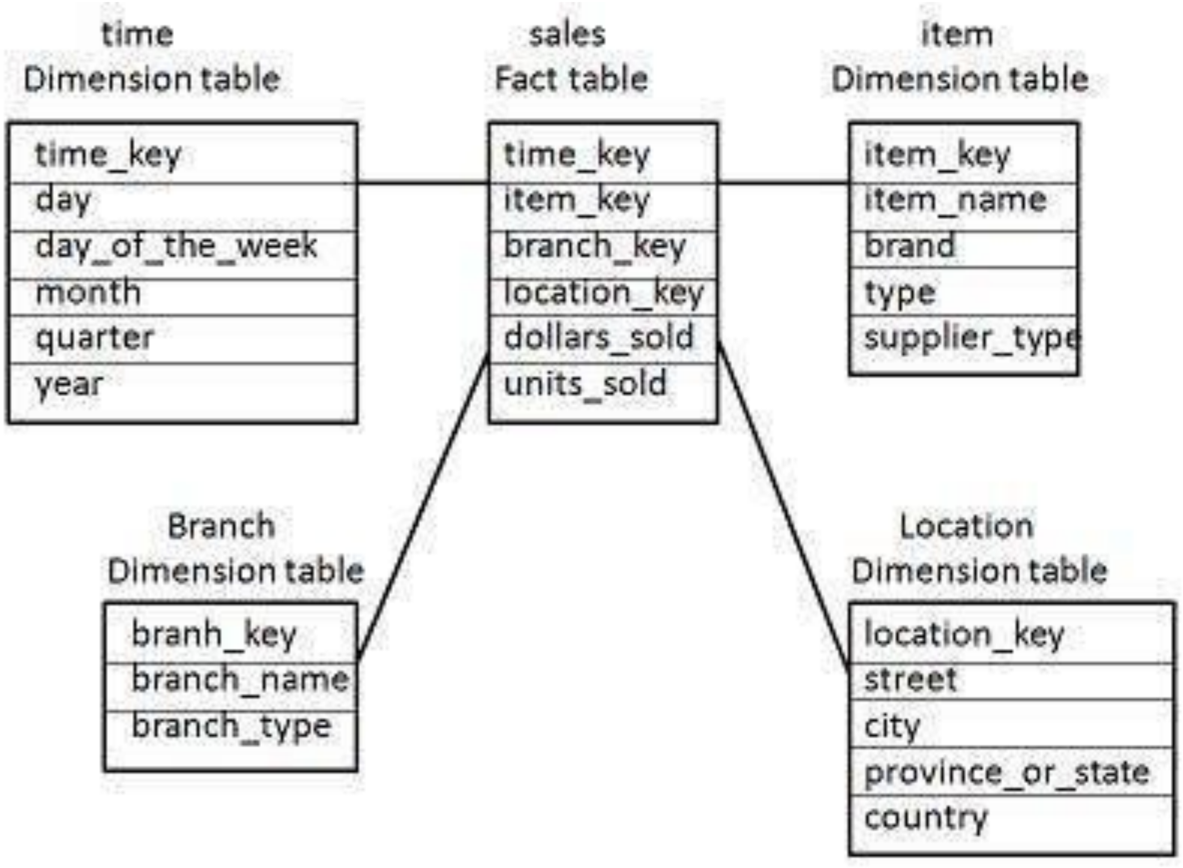
- In practice these concepts are used to create different schemas:
  - Star schema
  - Snowflake schema
  - Constellation schema

OLAP

- Star schema:
  - Each dimension is modeled as a table
  - The dimension tables contain the set of attributes
  - The Facts table constrains:
    - Foreign keys that reference the primary keys in the dimension tables
    - Calculated members based on the dimension hierarchies

OLAP

- Star Schema



time
Dimension table

time_key
day
day_of_the_week
month
quarter
year

sales
Fact table

time_key
item_key
branch_key
location_key
dollars_sold
units_sold

item
Dimension table

item_key
item_name
brand
type
supplier_type

Branch
Dimension table

branh_key
branch_name
branch_type

Location
Dimension table

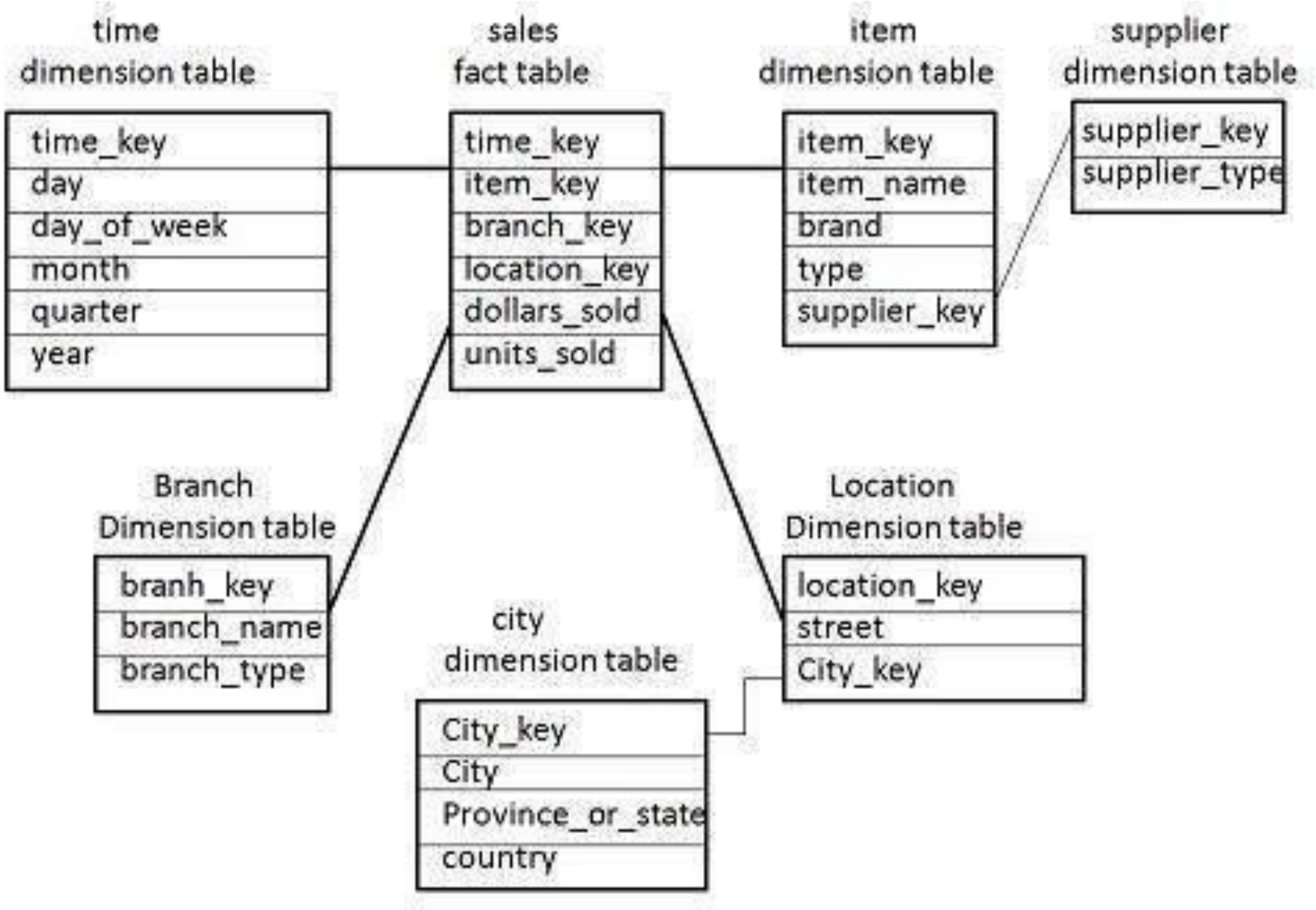
location_key
street
city
province_or_state
country

OLAP

- Snowflake Schema
  - The dimension are normalized
  - The data from dimensions and hierarchies are split intro multiple tables

OLAP

- Snowflake Schema



time
dimension table

time_key
day
day_of_week
month
quarter
year

sales
fact table

time_key
item_key
branch_key
location_key
dollars_sold
units_sold

item
dimension table

item_key
item_name
brand
type
supplier_key

supplier
dimension table

supplier_key
supplier_type

Branch
Dimension table

branh_key
branch_name
branch_type

Location
Dimension table

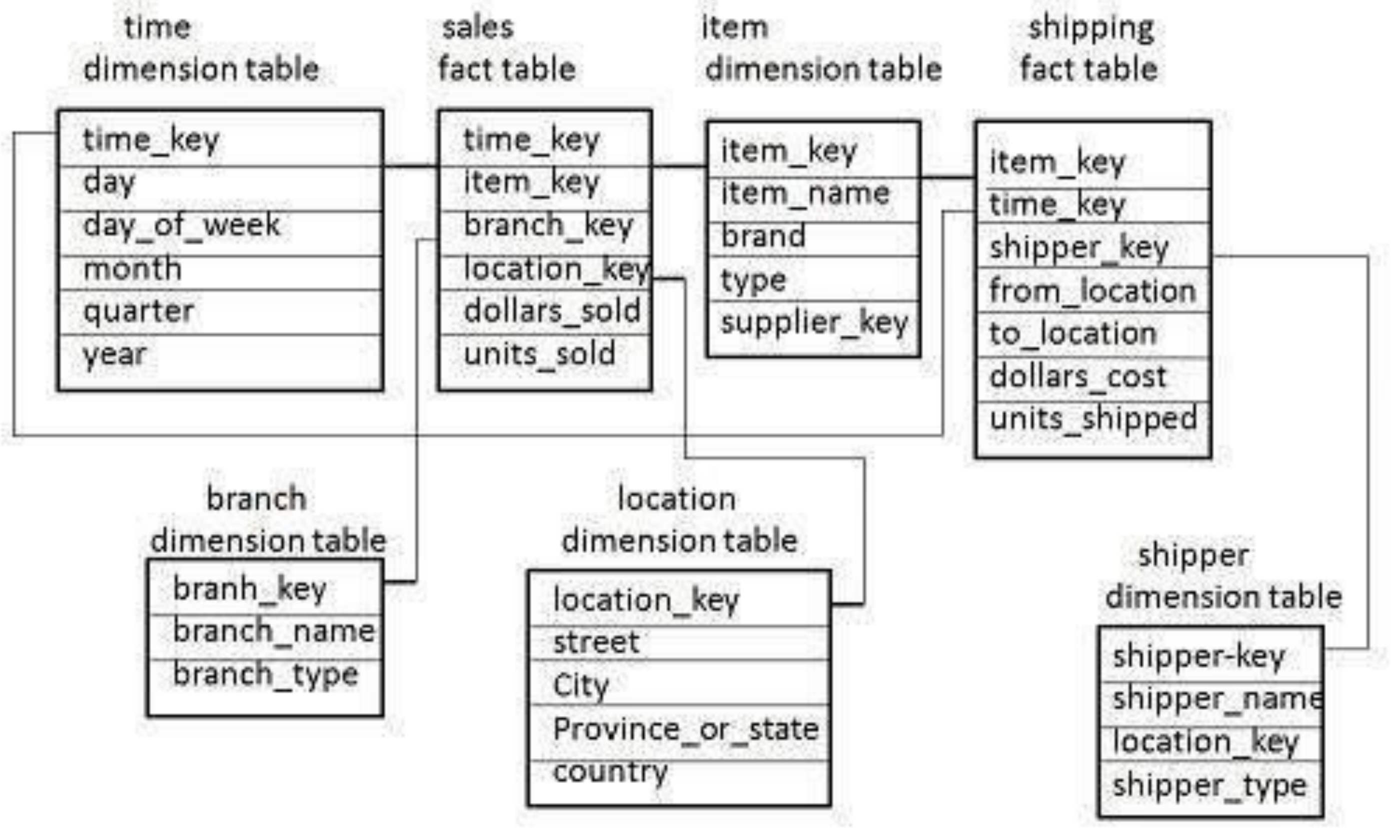
location_key
street
City_key

city
dimension table

City_key
City
Province_or_state
country

OLAP

- Constellation Schema: contains multiple Fact Tabeles



time
dimension table

time_key
day
day_of_week
month
quarter
year

sales
fact table

time_key
item_key
branch_key
location_key
dollars_sold
units_sold

item
dimension table

item_key
item_name
brand
type
supplier_key

shipping
fact table

item_key
time_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

branch
dimension table

branh_key
branch_name
branch_type

location
dimension table

location_key
street
City
Province_or_state
country

shipper
dimension table

shipper-key
shipper_name
location_key
shipper_type

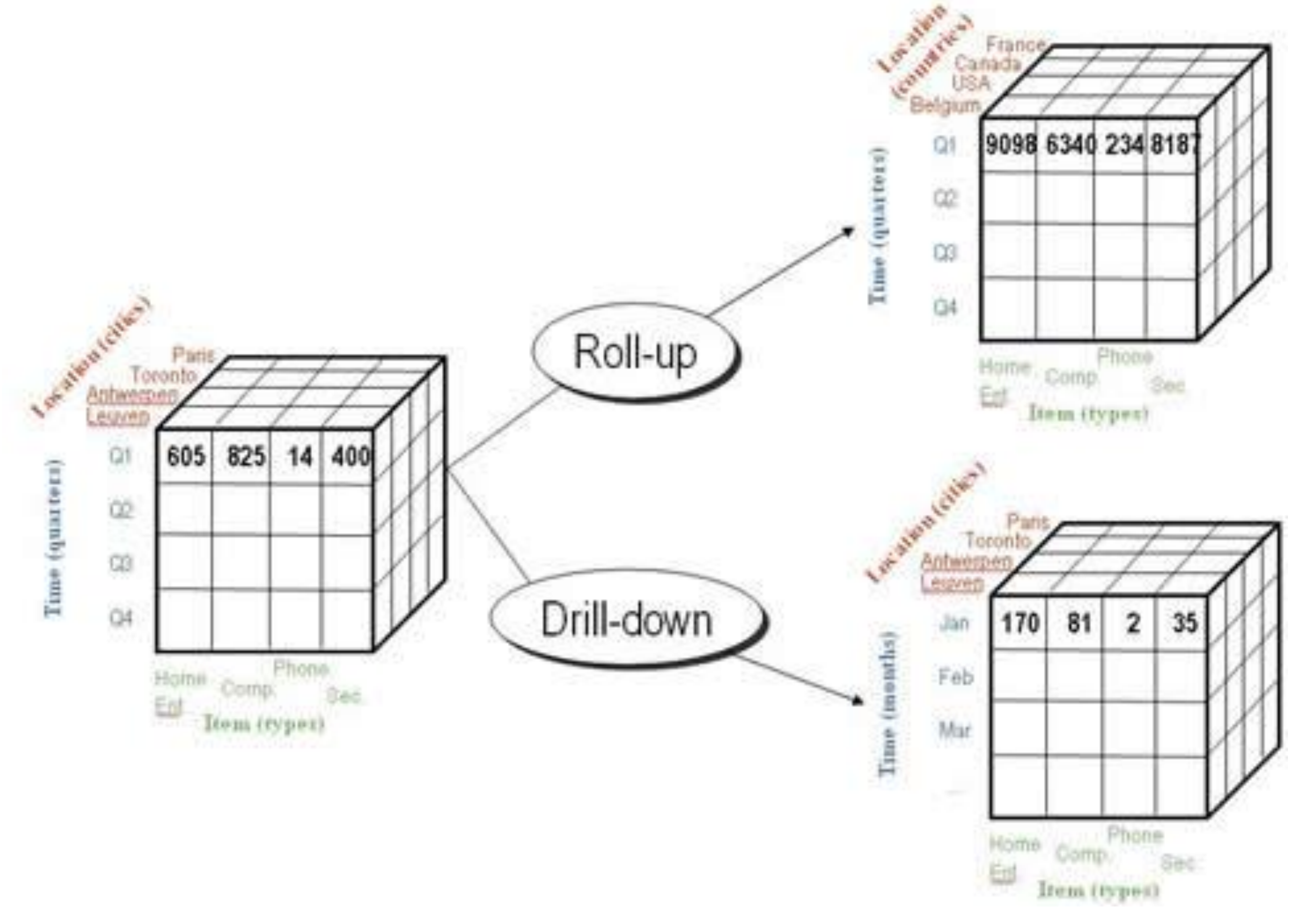
OLAP Operations

1. Roll-up is used to climb up a concept hierarchy for a dimension or by reducing dimensions
2. Drill-down is used o step down a concept hierarchy or by introducing a new dimension

- Roll-up is the reverse operation of Drill-down

OLAP Operations

- Roll-up and Drill-down

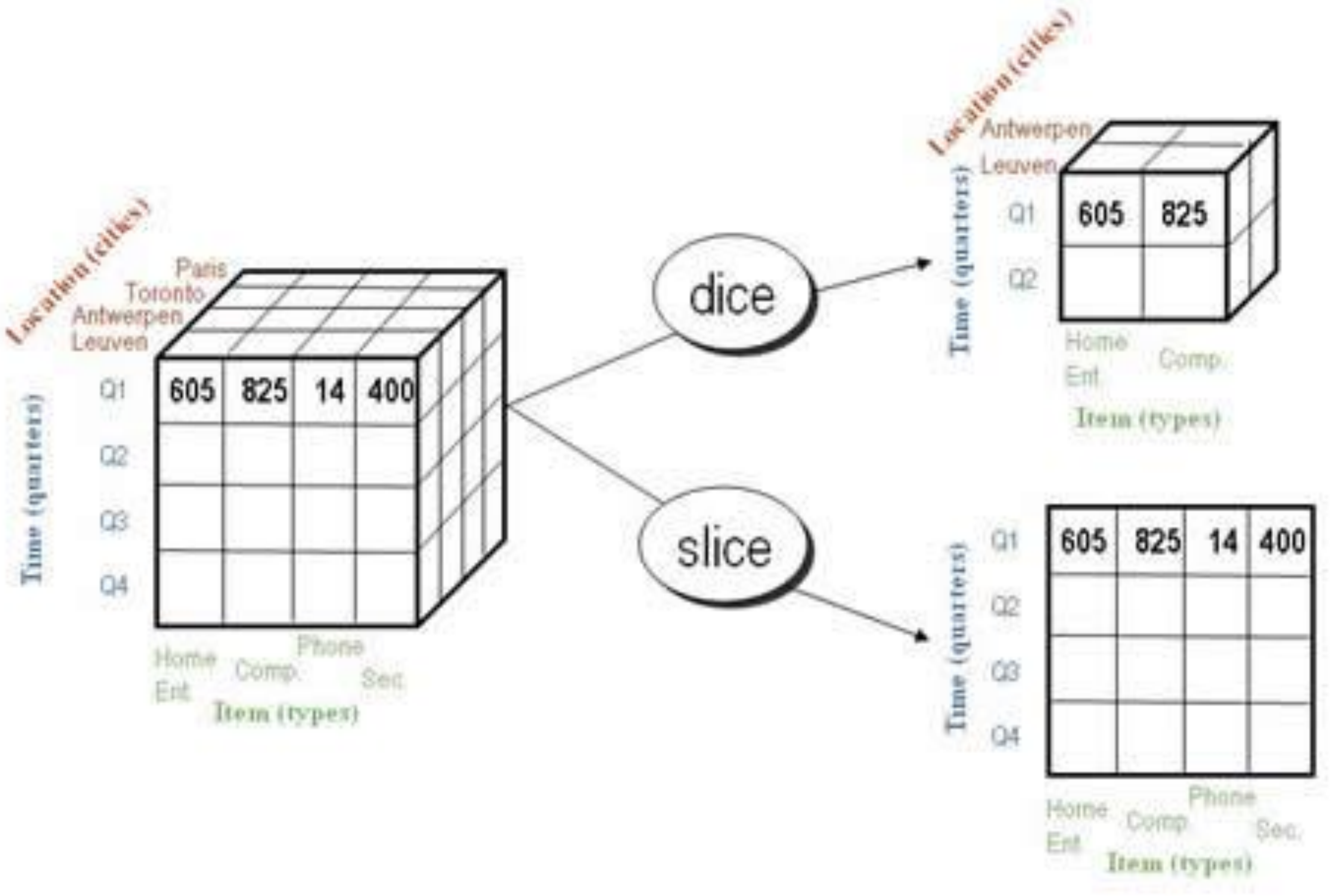


OLAP Operations

3. Slice operation selects one particular dimension from a given cube and provides a new sub-cube
4. Dice selects two or more dimensions from a given cube and provides a new sub-cube

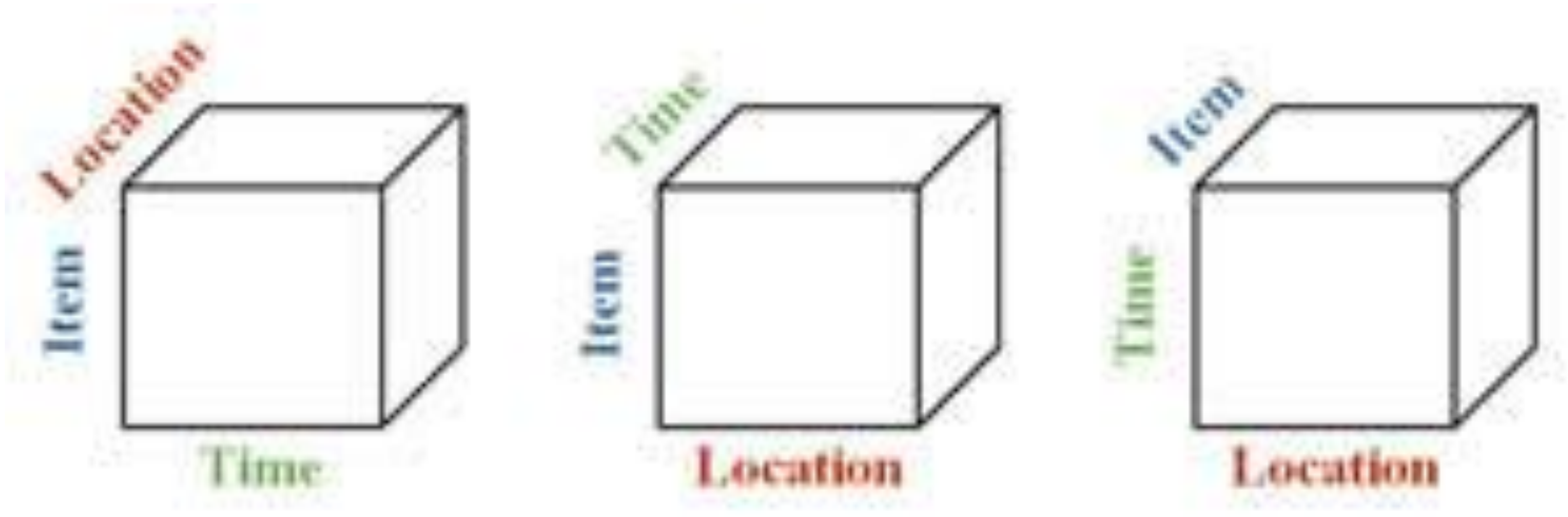
OLAP Operations

- Slice and Dice



OLAP Operations

5. Pivot (or rotate) is used to rotate the cube on one of the axes to visualize the date from different perspective



Summary

- This course presented:
  - A brief introduction to relational algebra and relational databases
  - A brief example of data warehouses
  - OLAP

References

- [Inmon 2002] W.H. Inmon - Building The Data Warehouse. Third Edition, Wiley & Sons, 2002
- [Kimball 2002] Ralph Kimball, Margy Ross - The Data Warehouse Toolkit, Second Edition, Wiley & Sons, 2002