Data Warehousing
and
Data Mining

S. K. Ghosh

Department of Computer Science and Engineering

Indian Institute of Technology, Kharagpur 721302

skg@cse.iitkgp.ac.in

Data, Data everywhere yet ...

- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences
- I can't get the data I need
  - need an expert to get the data
- I can't understand the data I found
  - available data poorly documented
- I can't use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

What is Data Warehousing?

A process of transforming data into information and making it available to users in a timely enough manner to make a difference

[Forrester Research, April 1996]

Data Warehouse?

- Different definitions -
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses

Data Warehouse—Subject-Oriented

- Organized around major subjects.
  [For example - customer, product, sales]
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - “Interoperability”
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element".

Data Warehouse—Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Data Warehouse vs. Heterogeneous DBMS

- Traditional heterogeneous DB integration:
  - Build wrappers/mediators on top of heterogeneous databases
  - Query driven approach
    - When a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set
    - Complex information filtering, compete for resources
- Data warehouse: update-driven, high performance
  - Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

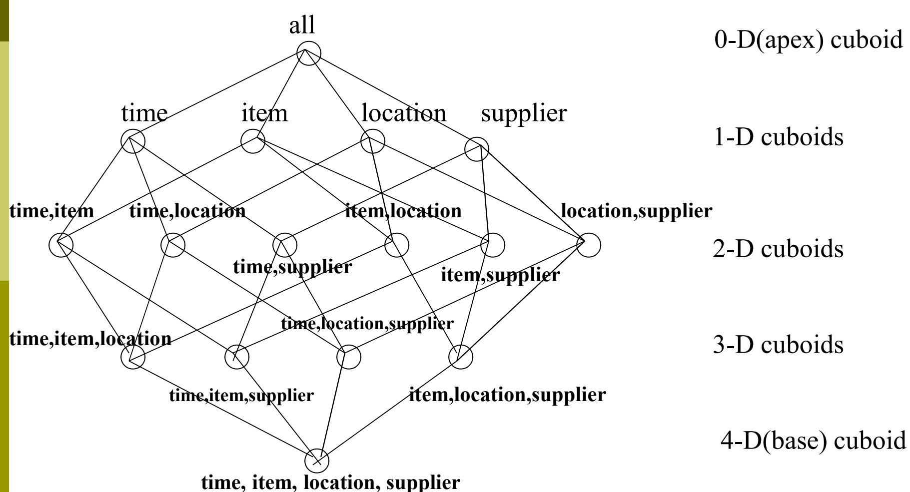
Why Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation.
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled

Multi-dimensional Data Model – From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

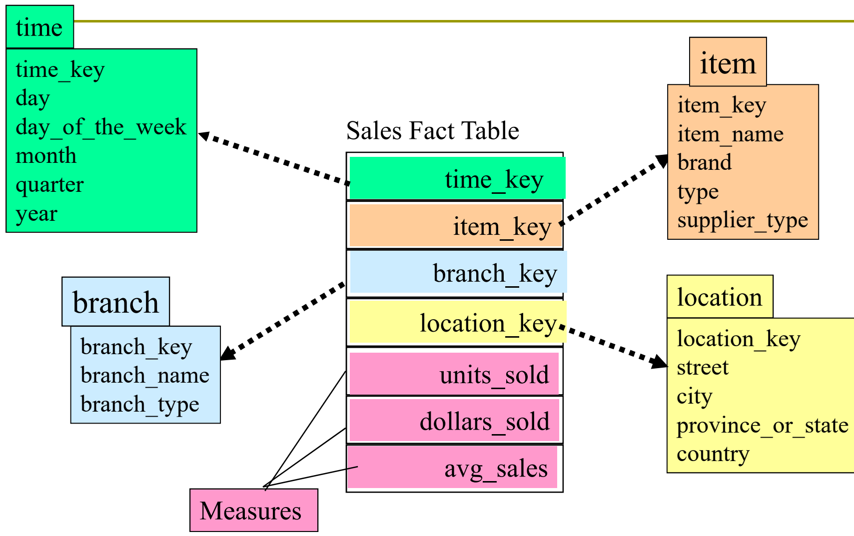
Cube: A Lattice of Cuboids



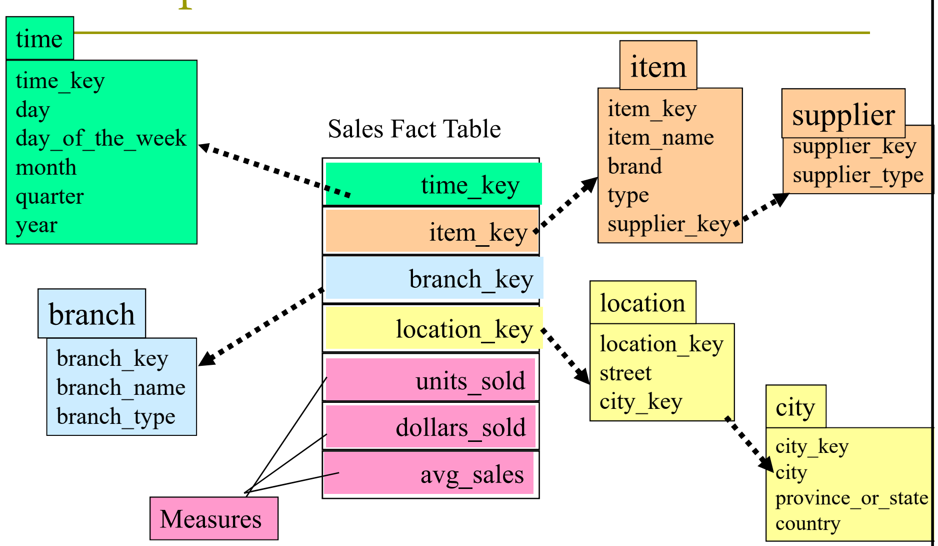
Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

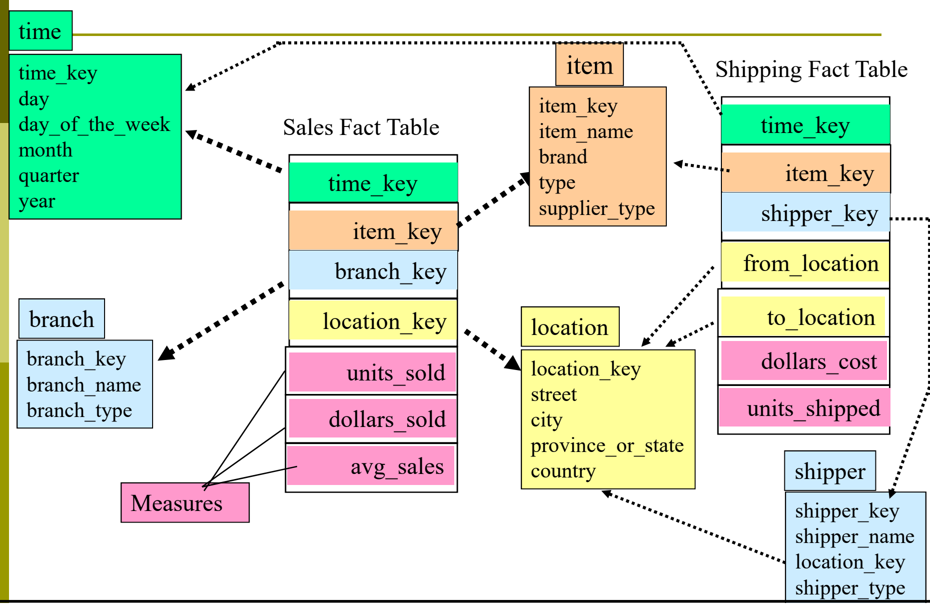
Example of Star Schema



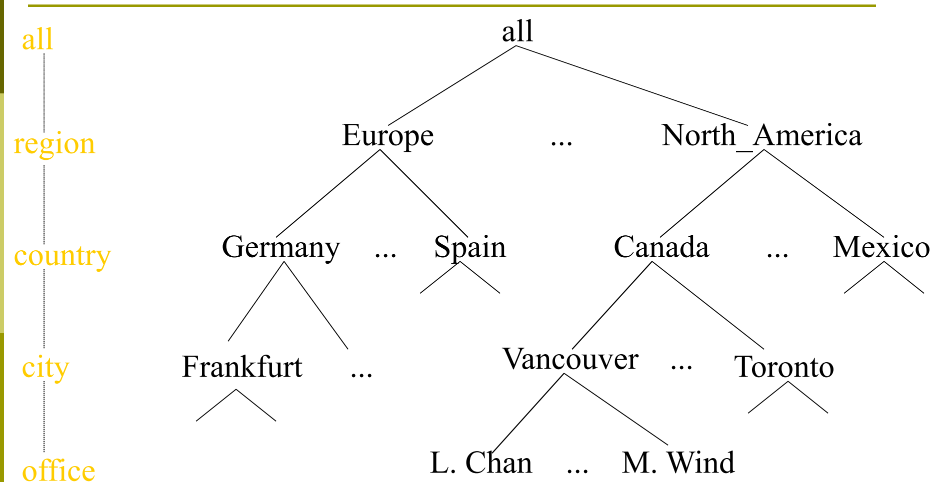
Example of Snowflake Schema



Example of Fact Constellation



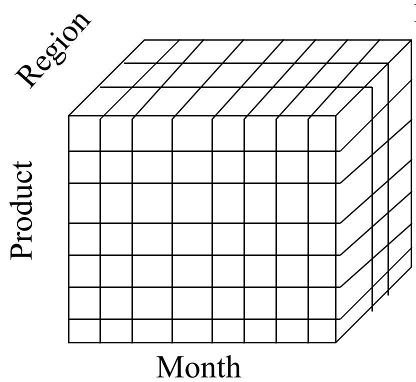
A Concept Hierarchy: Dimension (location)

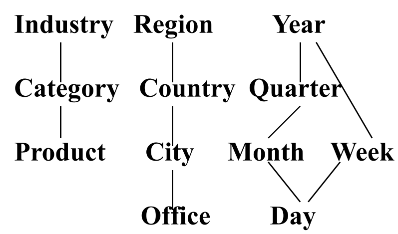


Multidimensional Data

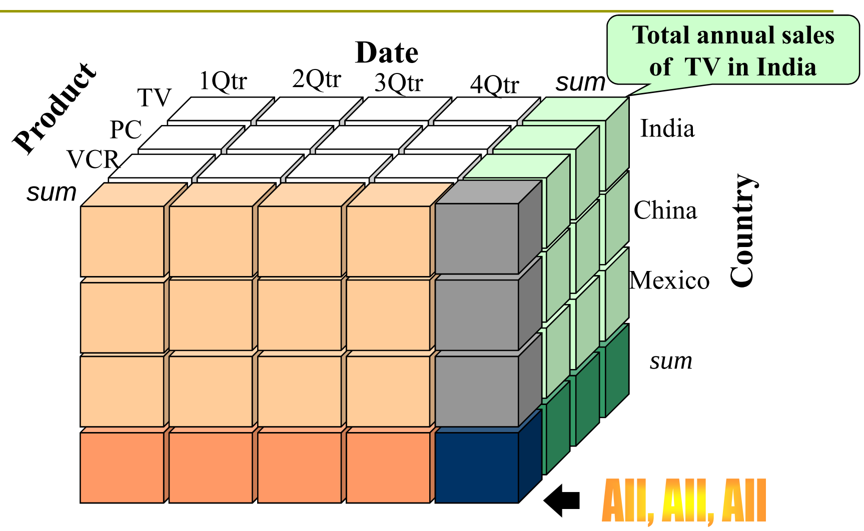
- Sales volume as a function of product, month, and region

Dimensions: Product, Location, Time
Hierarchical summarization paths

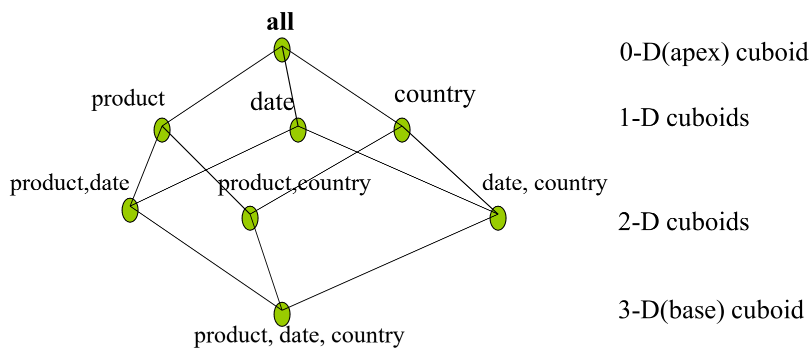




A Sample Data Cube



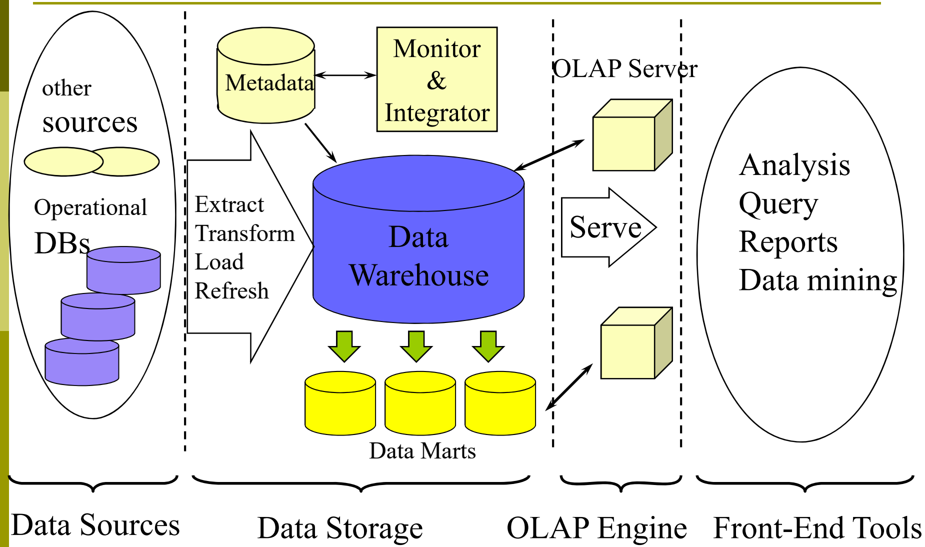
Cuboids Corresponding to the Cube



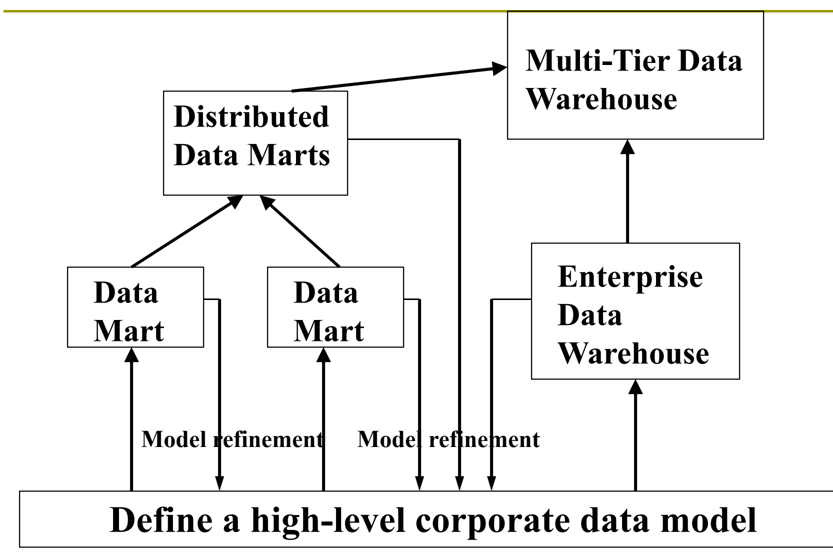
Typical OLAP Operations

- **Roll up (drill-up):** summarize data
  - *by climbing up hierarchy or by dimension reduction*
- **Drill down (roll down):** reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- **Slice and dice:**
  - *project and select*
- **Pivot (rotate):**
  - *reorient the cube, visualization, 3D to series of 2D planes.*
- Other operations
  - ***drill across:*** *involving (across) more than one fact table*
  - ***drill through:*** *through the bottom level of the cube to its back-end relational tables (using SQL)*

Multi-Tiered Architecture



Data Warehouse Development

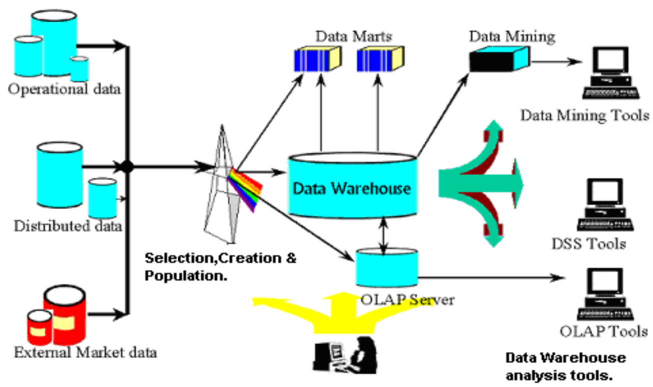


DATA MINING

Data Mining

- Data mining refers to the discovery of new information in terms of patterns or rules from vast amounts of data
- **Data warehousing and Data mining**
  - The goal of data warehouse is to support decision making process
  - Data mining can be used in conjunction with a data warehouse to help with certain decisions
  - Data mining can be applied to operational databases but to make it more efficient and meaningful it is applied to data warehouses
- Data mining applications should be considered early during the design of a data warehouse

"DW-DM" Architecture



Data Mining and Knowledge Discovery

- Knowledge discovery in databases (KDD) --- more general than data mining
- KDD process consists of six phases
  1. Data selection   2. Data cleaning
  3. Enrichment   4. Data transformation
  5. Data mining   6. Display and reporting
- Example
  Consumer goods retailer
  - Association rule: whenever a customer buys product X he also buys product Y
  - Sequential pattern: whenever a customer buys a camera then within six months he buys photographic supplies
  - Classification trees: credit-card customers, cash customers, etc.

Goals of Data Mining

- **Prediction** --- data mining can show how certain attributes within the data will behave in the future
- **Identification** --- data patterns can be used to identify the existence of an item, event, or an activity
- **Classification** --- data mining can partition the data so that different classes or categories can be identified based on combinations of parameters
- **Optimization** --- one eventual goal of data mining may be to optimize the use of limited resources such as time, space, money, or materials

Knowledge Discovery during Data Mining

- Raw data $\Rightarrow$ Information $\Rightarrow$ knowledge
- Deductive knowledge
  - Deduce new information based on applying pre-specified logical rules of deduction on the given data
- Inductive knowledge
  - Discover new rules and patterns from the available data
- Data mining addresses inductive knowledge
  - Discovered knowledge can be
    - Unstructured like rules or propositional logic
    - Structured like decision trees, semantic network, neural networks, etc

Types of Knowledge Discovered

- The knowledge discovered during data mining can be described as
  - **Association rules** --- correlate the presence of a set of items with another range of values for another set of variables
  - **Classification hierarchies** --- create hierarchies of classes
  - **Sequential patterns** --- sequence of actions or events
  - **Pattern with time series** --- similarities detected within positions of the time series
  - **Categorization and segmentation** --- partition a given population of events or items into sets of “similar” elements.

Association Rules

- An **association rule** is of the form $X \Rightarrow Y$
  where $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_m\}$ are sets of distinct items
  The rule states that if a customer buys X, he is also likely to buy Y
- The set LHS $\cup$ RHS is called an **itemset**
- **Interest measures**
  1. **Support** (*prevalence*) for the rule LHS $\Rightarrow$ RHS is the percentage of transactions that hold all the items in the itemset.
  2. **Confidence** (*strength*) for the rule LHS $\Rightarrow$ RHS is the percentage (fraction) of all transactions that include items in LHS and out of these the ones that include items of RHS.
     - Confidence is computed as support (LHS $\cup$ RHS) / support (LHS)

Example

Tid	time	Items
101	6:35	milk, bread, cookies, juice
102	7:38	milk, juice
103	8:05	milk, eggs
104	8:40	bread, cookies, coffee

- Consider two rules *milk* $\Rightarrow$ *juice* and *bread* $\Rightarrow$ *juice*

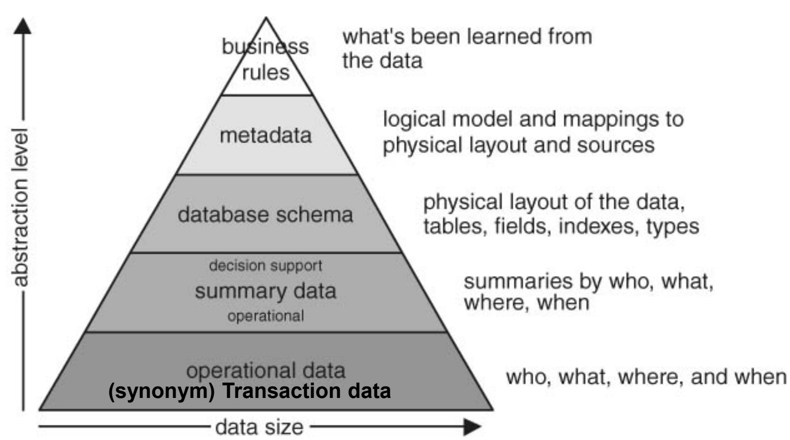
  Support {milk, juice} is 50%

  Support {bread, juice} is 25%

  Confidence of *milk* $\Rightarrow$ *juice* is 66.7%

  Confidence of *Bread* $\Rightarrow$ *juice* is 50%

Generic Architecture of Data



Data Mining Objectives:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events