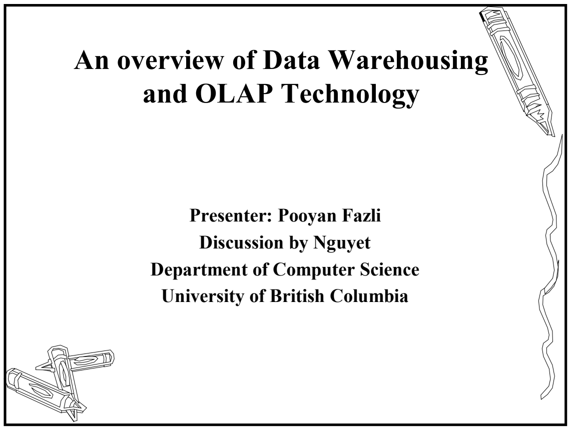
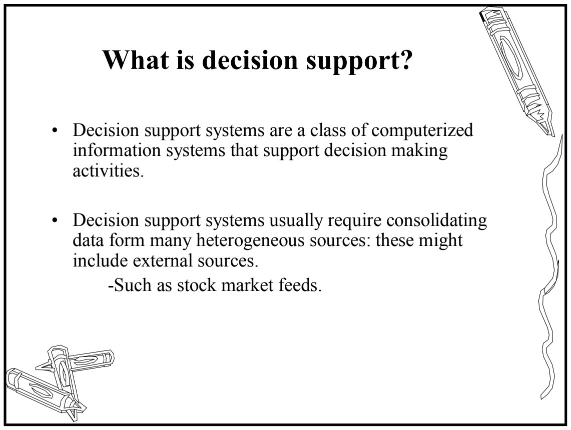


An overview of Data Warehousing and OLAP Technology

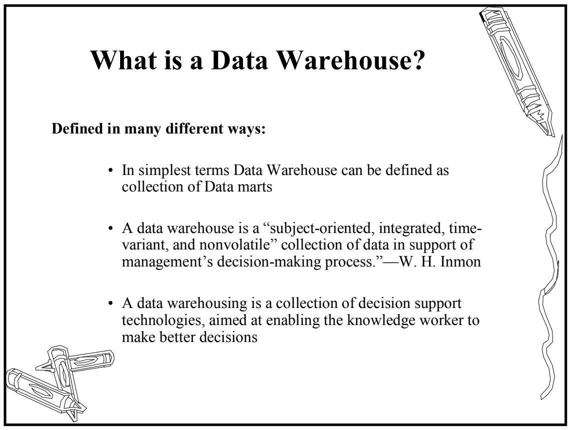
Presenter: Pooyan Fazli
Discussion by Nguyet
Department of Computer Science
University of British Columbia



What is decision support?

- Decision support systems are a class of computerized information systems that support decision making activities.
- Decision support systems usually require consolidating data form many heterogeneous sources: these might include external sources.

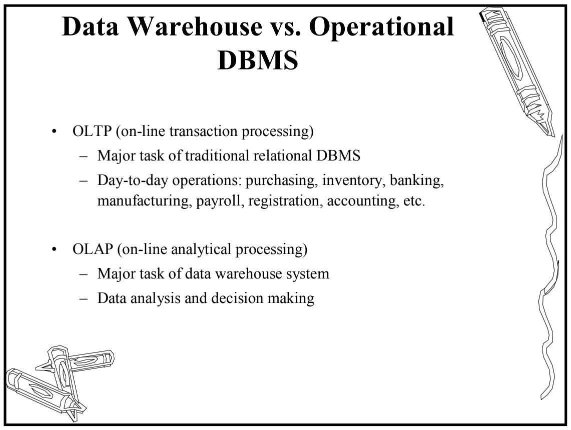
  -Such as stock market feeds.



What is a Data Warehouse?

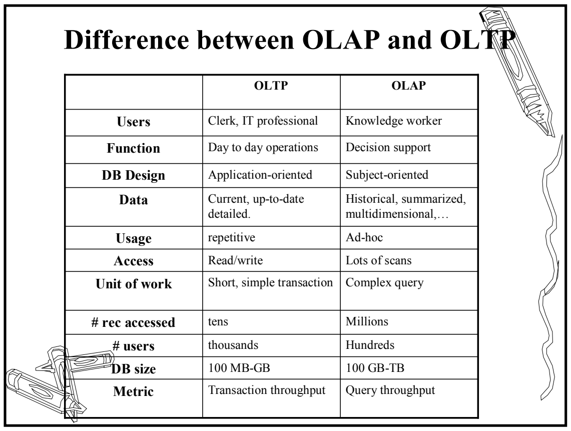
Defined in many different ways:

- In simplest terms Data Warehouse can be defined as collection of Data marts
- 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
- A data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker to make better decisions



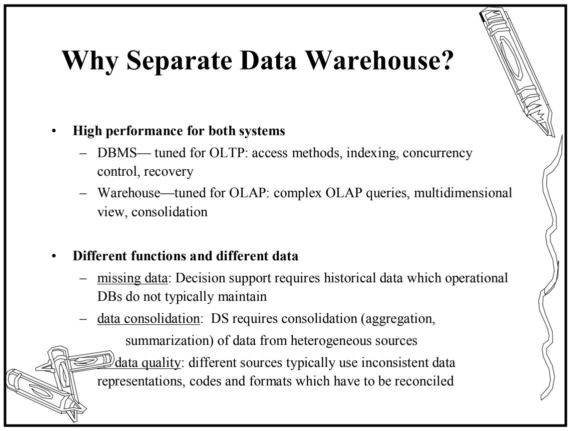
Data Warehouse vs. Operational DBMS

- 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



Difference between OLAP and OLTP

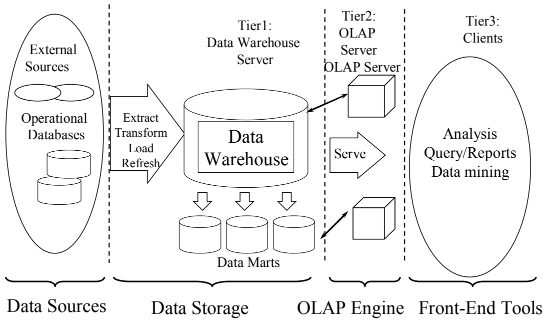
	OLTP	OLAP
Users	Clerk, IT professional	Knowledge worker
Function	Day to day operations	Decision support
DB Design	Application-oriented	Subject-oriented
Data	Current, up-to-date detailed.	Historical, summarized, multidimensional,…
Usage	repetitive	Ad-hoc
Access	Read/write	Lots of scans
Unit of work	Short, simple transaction	Complex query
# rec accessed	tens	Millions
# users	thousands	Hundreds
DB size	100 MB-GB	100 GB-TB
Metric	Transaction throughput	Query throughput



Why Separate 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

A Data Warehousing Architecture



Data Warehouse Back-End Tools and Utilities

- Data extraction
  - get data from multiple, heterogeneous, and external sources
- Data cleaning
  - detect errors in the data and rectify them when possible
- Data transformation
  - convert data from legacy or host format to warehouse format
- Load
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse

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: taking a projection of the data on a subset of dimensions for selected values of the other dimension
- Pivot (rotate): reorient the cube, visualization, 3D to series of 2D planes

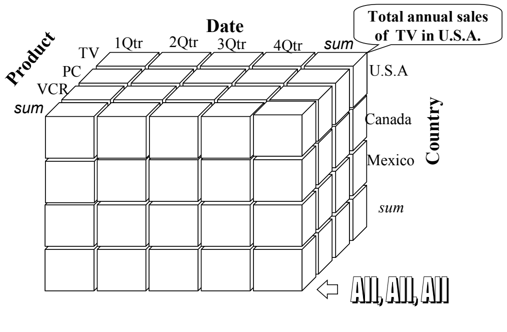
From Tables to Data Cubes

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- In a multidimensional data model, there is a set of *numeric measures* that are the objects of analysis.
- Each of the numeric measures depends on a set of *dimensions,* which provide the context for the measure.

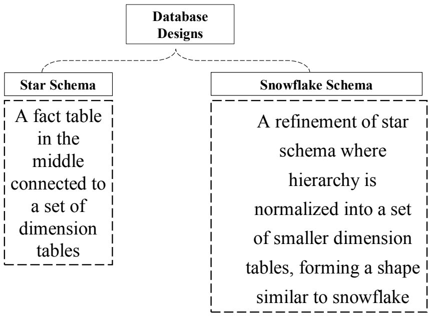
From Tables to Data Cubes

- The dimensions together are assumed to *uniquely* determine the measure.
- Each dimension is described by a set of *attributes*.
- The attributes of a dimension may be related via a hierarchy of relationships.

A Sample Data Cube



Database Design Methodology


Database Designs

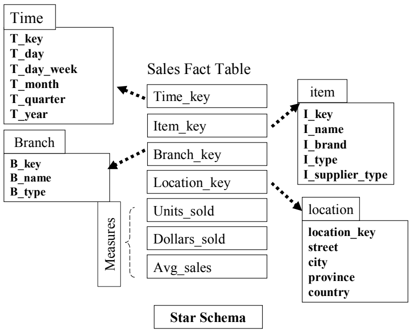
Star Schema

A fact table in the middle connected to a set of dimension tables

Snowflake Schema

A refinement of star schema where hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake

Star Schema


Time

T_key
T_day
T_day_week
T_month
T_quarter
T_year

Branch

B_key
B_name
B_type

Sales Fact Table

Time_key
Item_key
Branch_key
Location_key
Units_sold
Dollars_sold
Avg_sales

Measures

item

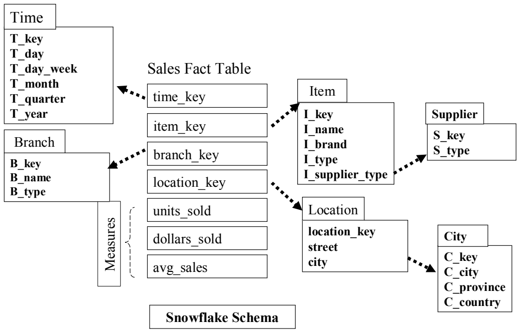
I_key
I_name
I_brand
I_type
I_supplier_type

location

location_key
street
city
province
country

Star Schema

Snowflake Schema


Time

T_key
T_day
T_day_week
T_month
T_quarter
T_year

Branch

B_key
B_name
B_type

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

Item

I_key
I_name
I_brand
I_type
I_supplier_type

Supplier

S_key
S_type

Location

location_key
street
city

City

C_key
C_city
C_province
C_country

Snowflake Schema

Materialized Views In a Warehouse

Challenges in exploiting materialized views

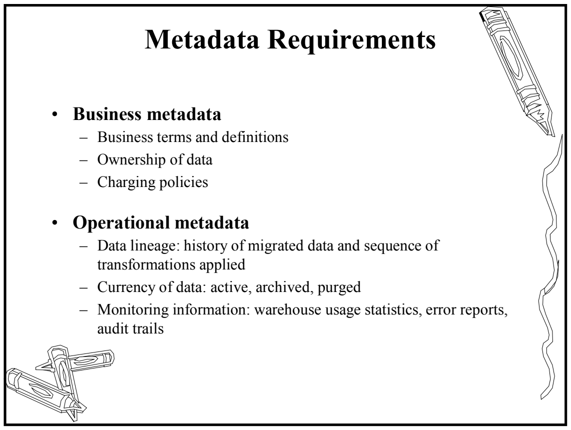
- identify the views to materialize
- exploit the materialized views to answer queries,
- efficiently update the materialized views during load and refresh.

Materialized Views In a Warehouse

- The currently adopted industrial solutions to these problems consider materializing views that have a relatively simple structure. Such views consist of joins of the fact table with a subset of dimension tables with the aggregation of one or more measures grouped by a set of attributes from the dimension tables.
- The selection of views to materialize must take into account workload characteristics, the costs for incremental update, and upper bounds on storage requirements

Metadata Requirements

- **Administrative metadata**
  - Source database and their contents
  - Source database and their contents
  - Back-end and front-end tools
  - Definitions of the warehouse schema
  - Pre-defined queries and reports
  - Data mart locations and contents
  - Data refresh and purging policies
  - User profiles and user access control policies



Metadata Requirements

- **Business metadata**
  - Business terms and definitions
  - Ownership of data
  - Charging policies
- **Operational metadata**
  - Data lineage: history of migrated data and sequence of transformations applied
  - Currency of data: active, archived, purged
  - Monitoring information: warehouse usage statistics, error reports, audit trails