OVERVIEW: DATA WAREHOUSING & OLAP TECHNOLOGY

Presentation: Sophia

Discussion: Tianyu

"An Overview of Data Warehousing and OLAP Technology"

author: Surajit Chaudhuri
Umeshwar Dayal

Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclusion

Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclution

Motivation– Decision Support

- Decision Making: Everyday, Everywhere
- Decision Support System:
  - a class of computerized information systems that support decision making activities.
- What are needed to support decision making?

Data	Heteroguous Large Scale	Data Warehouse
Analysis Tools	User Friendly Good Query Throughput	OLAP tools

Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclution

Definition--data warehouse

- A data warehouse is a

Subject-oriented	Day-to-day, transaction
Integrated	Usually from one source
Time-variant	Within a short period of time
Non-volatile	Frequent updates

RDMS

OLAP OLTP

collection of data in support of management's decision making process.

OLAP VS. OLTP

- Manager: "I need to analyze the curve of sales of our company within the past 10 years."
  - "Let's try to find something interesting!"
- Casher: "One computer was sold and I got $1000. I need to update the database."
  - "That's what I do everyday!"

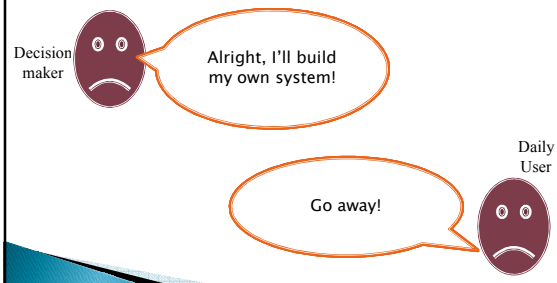
	OLTP	OLAP
Users	Clerk, IT professional	Knowledge worker
Function	Day to day operations	Decision support
DB Design	Application-oriented	Subject-oriented
Data	Current, up-to-date detailed.	Historical, summarized, multidimensional,...
Usage	repetitive	Ad-hoc
Access	Read/write	Lots of scans
Unit of work	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

Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclution

Why not RDBM & OLTP?

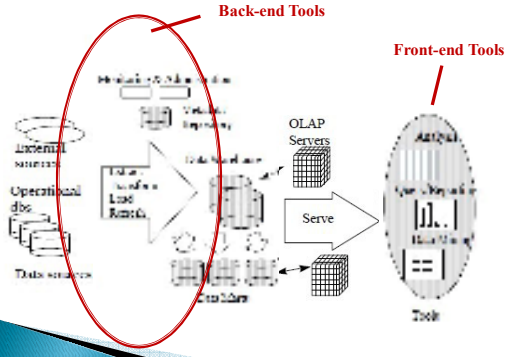
Decision Maker VS. Daily User



Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclution

Typical OLAP architecture



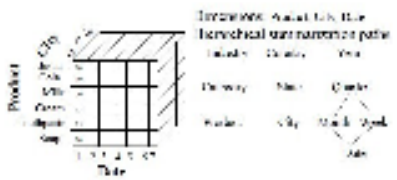
Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclution

Logical Model

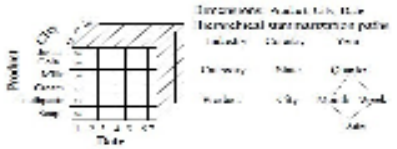
- Multidimensional

  Numerical Measure – Dimension -- Attribute



Logical Model

- Front-end operations
  - Pivot
  - Rollup
  - Drilldown
  - Slice and dice

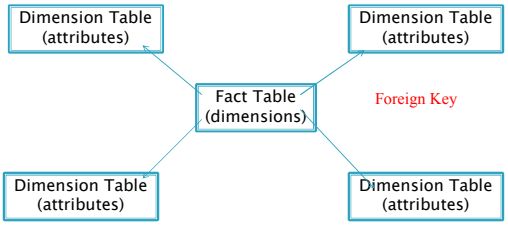


Syllabus

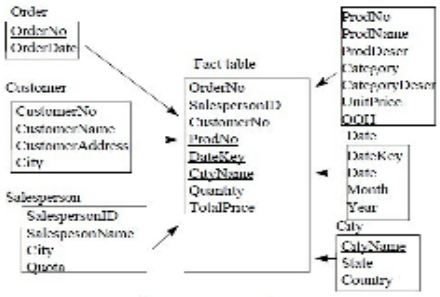
- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclution

Database Design Methodology

- Star Schema



Star Schema Example



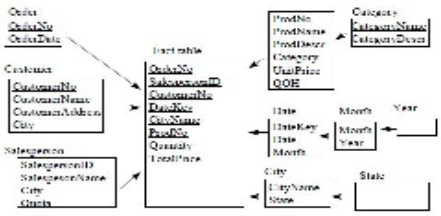
Database Design Methodology

- Problem of Star Schema
  - Un-normalized (redundancy)
  - No attribute hierarchy

Ex: (Vancouver, BC, Canada, North America)
(Victoria, BC, Canada, North America)

Normalize dimension tables→
Snowflake Schema

Example Snowflake Schema



More Problems:

- saving of space is negligible in comparison to the typical magnitude of the fact table.
- Efficiency

Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclution

Warehouse Servers

- Indices
- Materialized Views
- Answer queries using indices and views
- Optimization of Complex queries
- Parallel Processing
- Server Architecture:
  - Specialized SQL Servers
  - ROLAP
  - MOLAP
- SQL Extensions: next paper

Indices

- Bit map – One bit for each record
- Bitmap Compression
- Joint Indices

Materialized Views

- **Joins of the fact table with a subset of dimension tables, with the aggregation of one or more measures.**

  Example:

  Total Sale of each product in each State

- What to Materialize?
  - Workload, cost of update, storage
  - A greedy algorithm

Use of Views

- Selection, Rollup
- Generator(V: view Q: query):
  - Same dimensions
  - Selection clause of Q implies that of V
  - Group by of V is a subset of Q
- Minimal Generator M':
  - A set of generators (could be more than one)
  - Other generators generate some member of it

- Query: Total Sales of clothing in Washington
- Minimal Generators:
  - Total sales by each state for each product
  - Total sales by each city for each category
- Other Generators:
  - Total sales by each city for each product

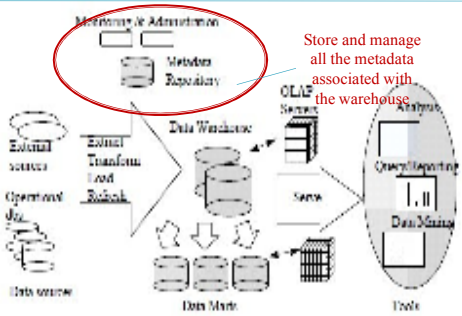
Syllabus

- Motivation
- Definition
- Why not RDMS & OLTP
- Typical OLAP Architecture
- Logical Model
- Schemas
- Materialized Views
- Metadata Requirements and Conclusion

Metadata requirements

- **Administrative metadata**
  - **Information necessary for setting up and using a warehouse**
- **Business metadata**
  - **Business terms; ownership; charging policy**
- **Operational metadata**
  - **Information collected during the operation of the warehouse**

Metadata requirements



Conclusion

- Data Warehousing is really hard
  - Data Cleansing
  - Physical Design
  - Management
  - Visualization

THERE'RE A LOT OF OPPORTUNITIES WAITING FOR YOU GUYS~

31