**OVERVIEW: DATA WAREHOUSING & OLAP TECHNOLOGY**

Presentation: Sophia  
Discussion: Tianyu

---

**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?

<table>
<thead>
<tr>
<th>Data</th>
<th>Heterogeneous Large Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Tools</td>
<td>User Friendly Good Query Throughput</td>
</tr>
</tbody>
</table>

Data Warehouse  
OLAP tools

---

**Syllabus**

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

---

**Syllabus**

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

---

**“An Overview of Data Warehousing and OLAP Technology”**

*author: Suraljit Chaudhuri  
Umeshwar Doyal*
**Definition--data warehouse**

- A data warehouse is a

<table>
<thead>
<tr>
<th>Subject-oriented</th>
<th>Day-to-day, transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrated</td>
<td>Usually from one source</td>
</tr>
<tr>
<td>Time-variant</td>
<td>Within a short period of time</td>
</tr>
<tr>
<td>Non-volatile</td>
<td>Frequent updates</td>
</tr>
</tbody>
</table>

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!”
- Cashier: “One computer was sold and I got $1000. I need to update the database.”
  “That’s what I do everyday!”

---

**OLTP vs. OLAP**

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

---

**Why not RDBM & OLTP?**

**Decision Maker VS. Daily User**

- Decision Maker: "Alright, I’ll build my own system!
- Daily User: "Go away!"

---

**Syllabus**

<table>
<thead>
<tr>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
</tr>
<tr>
<td>Why not RDM &amp; OLTP</td>
</tr>
<tr>
<td>Typical OLAP Architecture</td>
</tr>
<tr>
<td>Logical Model</td>
</tr>
<tr>
<td>Schemas</td>
</tr>
<tr>
<td>Materialized Views</td>
</tr>
<tr>
<td>Metadata Requirements and Conclusion</td>
</tr>
</tbody>
</table>
Typical OLAP architecture

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

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 Conclusion

Database Design Methodology
- Star Schema
  - Dimension Table (attributes)
  - Dimension Table (dimensions)
  - Fact Table
  - Foreign Key
  - Dimension Table (attributes)
**Star Schema Example**

**Example Snowflake Schema**

More Problems:
- Saving of space is negligible in comparison to the typical magnitude of the fact table.
- Efficiency

**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**

**Syllabus**

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

**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:
  
  ![Diagram]
  
  - What to Materialize?
    - Workload, cost of update, storage
    - A greedy algorithm

**Use of Views**

- **Selection, Rollup**
  
  ![Diagram]
  
- **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

**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
Conclusion

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

**There’re a lot of opportunities waiting for you guys~**