Data Warehousing and Data Mining

Lecture # 5

Sajid Majeed

Data Warehouse Vs. OLTP

OLTP (On Line Transaction Processing) specific query

Select tx_date, balance from tx_table
Where account_ID = 23876;

Data Warehouse Vs. OLTP

DWH specific query

Select balance, age, sal, gender from customer_table, tx_table
Where age between (30 and 40) and
Education = ‘graduate’ and
CustID.customer_table =
Customer_ID.tx_table;

Data Warehouse Vs. OLTP

OLTP	DWH
Primary key used	Primary key NOT used
No concept of Primary Index	Primary index used
Few rows returned	Many rows returned
May use a single table	Uses multiple tables
High selectivity of query	Low selectivity of query
Indexing on primary key (unique)	Indexing on primary index (non-unique)

Data Warehouse Vs. OLTP

OLTP: OnLine Transaction Processing (MIS or Database System)

	Data Warehouse	OLTP
Scope	* Application –Neutral * Single source of "truth" * Evolves over time * How to improve business	* Application specific * Multiple databases with repetition * Off the shelf application * Runs the business
Data Perspective	* Historical, detailed data * Some summary * Lightly denormalized	* Operational data * No summary * Fully normalized
Queries	* Hardly uses PK * Number of results returned in thousands	* Based on PK * Number of results returned in hundreds
Time factor	* Minutes to hours * Typical availability 6x12	* Sub seconds to seconds * Typical availability 24x7

Table-4.2: Detailed comparison of OLTP and DWH

Comparison of Response Times

- On-line analytical processing (OLAP) queries must be executed in a small number of seconds.
  - Often requires denormalization and/or sampling.
- Complex query scripts and large list selections can generally be executed in a small number of minutes.
- Sophisticated clustering algorithms (e.g., data mining) can generally be executed in a small number of hours (even for hundreds of thousands of customers).

Typical Applications

Impact on organization's core business is to streamline and maximize profitability.

- Fraud detection.
- Profitability analysis.
- Direct mail marketing.
- Yield management.

ROI on any one of these applications can justify HW/SW & consultancy costs in most organizations.

Fraud detection

- By observing data usage patterns.
- People have typical purchase patterns.
- Deviation from patterns.
- Certain cities notorious for fraud.
- Certain items bought by stolen cards.

Profitability Analysis

- Banks know if they are profitable or not.
- Don’t know which customers are profitable.
- Typically more than 50% are NOT profitable.
- Don’t know which one?
- Balance is not enough, transactional behavior is the key.
- Restructure products and pricing strategies.

Direct mail marketing

- Targeted marketing.
- Offering high bandwidth package NOT to all users.
- Know from call detail records of web surfing.
- Saves marketing expense, saving pennies.
- Knowing your customers better.

Yield Management

- Works for fixed inventory businesses.
- Item prices vary for varying customers.
- Example: Air Lines, Hotels etc.
- Price of (say) Air Ticket depends on:
  - How much in advance ticket was bought?
  - How many vacant seats were present?
  - How profitable is the customer?
  - Ticket is one-way or return?

1. What do we mean by strategic information. For a commercial bank, name some types of strategic objectives?

2. Why are operational systems not suitable for providing strategic information? Give specific reasons and explain?

3. A data warehouse is an environment, not a product. Discuss?

4. For an airline company, how can strategic information increase the number of frequent flyers? Discuss giving specific details.

5. Name at least six characteristics or features of a data warehouse.

6. Why is data integration required in a data warehouse, more so than in an operational application?

7. Why do you need a separate data staging component?

8. Name any six different methods for information delivery.