Data Warehouses

Chapter 2 (Part 1)

Outline

- Introduction
- **Data Warehouses**
- Data Warehouse in **Organisation**
- OLTP vs. **OLAP**
- **Separating** Data Warehouse from Operational Database

Introduction

- **Data warehouses** can be considered as **data repositories** that are maintained separately from organization's operational database
- Functions:
  - » Allow **integration of multiple application systems**
  - » Platform of **consolidated historical data to support information processing and analysis**
- **Data warehousing** provides **architectures and tools** for business executives to systematically **organize, understand, and use data to make strategic decisions**

Data Warehouses

- According to the original definition of Bill Inmon (1996), the father of data warehouses, a data warehouse is *a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process*.

Key factors distinguishing them from other data repositories

Data Warehouse: **Subject-Oriented**

- Organized around major subjects (e.g. 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 a 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
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - 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
  - Either explicitly or implicitly
  - 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 Warehouses

- Are the foundation of the business IT infrastructures that collect data from several dispersed information sources and are designed to allow decision makers have prompt access to information for purpose of reporting

Data Warehouse in Organisation

- Data warehousing is the process of constructing and using data warehouses
- Construction includes processes of data cleaning, data integration, and data consolidation
- Utilization refers to obtaining an overview of data to make sound decisions
  - This requires various decision support technologies
- Organisations use **information from data warehouses for business decision-making activities**

Data Warehouse in Organisation

- **Aetna Life** (insurance) is an American company
  - uses IBM's data warehouse and data mining tools to have a better understanding for meeting the specific needs of its customers
  - to estimate the performance of **new products and services.**
- **Guinness Limited** (food and beverage) is a British company
  - overcome the difficulties of extracting data from transaction processing systems
  - populate a data warehouse with **valuable business information** to serve customer needs

Data Warehouse in Organisation

- Parkson Corporation Sdn Bhd (retail) is a Malaysian company
  - increased marketing program efficiency and market share through implementation of data warehouse and data mining to work for its **29** stores in Malaysia

Data Warehouse in Organisation

- Past SAS Asia Pacific Risk Management Practice Head, John Foulley has said many banks in Malaysia (back around 2000s) had the problem of integrating their data efficiently and this had led to **misplacement of information and poor quality data**
- At the time, most of Malaysian banks do not have implementation of data warehouse yet
- The local banks have to implement Basel II framework as instructed by Bank Negara (around 2008~2010).
- The framework addresses on credit and operational risks which requires ready data warehouses.

Data Warehouse in Organisation

- Alliance Banking Group allocated 36 million to build data warehouse.
- Insurance Services Malaysia (ISM) handles more than 50 insurance companies in Malaysia
  - They require insurance companies to deliver clean, structured data to them to build the data warehouse.

Data Warehouse in Organisation

- Malaysia's EON Bank had problems to access the complete view of the customer due to loan information sitting in one transactional system and credit details in another system.
- AmBank Group has invested over RM10 million involving the implementation of data integration and management solution.

Operational Database vs. Data Warehouses

- Online operational databases systems
  - perform online transaction and query processing
  - Covers most of day-to-day operations
  - **Online Transaction Processing** (OLTP)
- Data warehouse systems
  - Serves users in data analysis and decision making
  - Organize and present data in various formats to accommodate diverse needs
  - **Online Analytical Processing** (OLAP)

OLTP vs. OLAP

- There are few major distinguishing features
- User and system orientation
  - OLTP is customer-oriented for transaction and query processing
    - Used by clerks, clients, IT professionals, etc.
  - OLAP is market-oriented for data analysis
    - Used by analysts, managers, etc.
- Data contents
  - OLTP manages current data; typically too detailed and not straightforward for decision making
  - OLAP manages large amount of historical data for easier decision making
    - Has facilities for summarization, aggregation, storage, and management of information at different levels of granularity

OLTP vs. OLAP

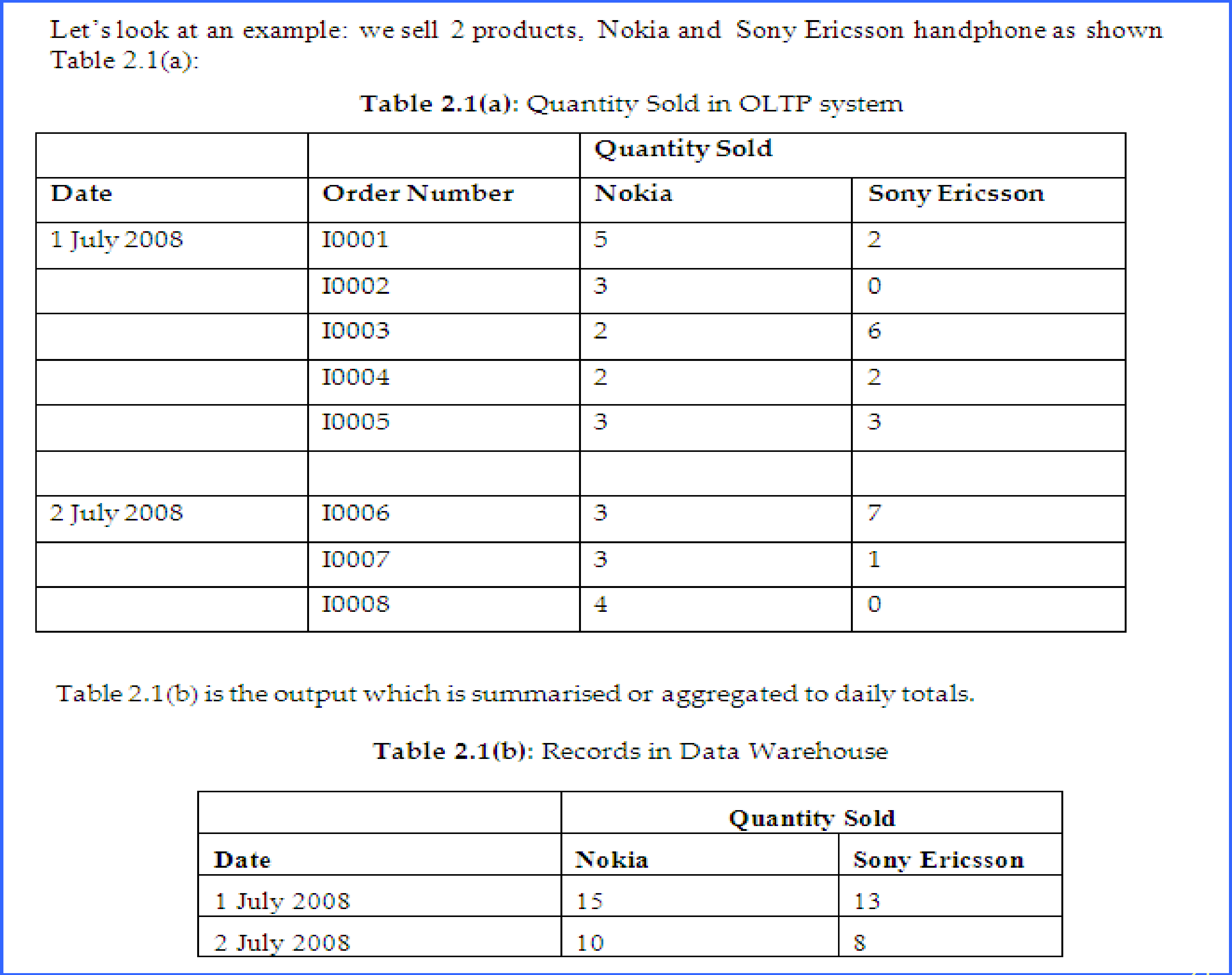
- Database design
  - OLTP usually adopts entity-relationship (ER) model
    - Application-oriented design
  - OLAP typically adopts star or snowflake model
    - Subject-oriented design
- View
  - OLTP focus mainly on current data, no reference to historical data or data from different organisations
  - OLAP may:
    - Span to multiple version of DB schema due to organisation evolution
    - Deal with data from different organisations (need to be integrated)
    - Be stored in multiple storage media due to large volume

OLTP vs. OLAP

- Access patterns
  - OLTP consist mainly of short atomic transactions
    - Require concurrency control and recovery mechanisms
  - OLAP mostly read-only operations but can have complex queries
    - Most data warehouse store historical data instead of up-to-date info.

OLTP vs. OLAP

	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, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response



Let's look at an example: we sell 2 products, Nokia and Sony Ericsson handphone as shown Table 2.1(a):

Table 2.1(a): Quantity Sold in OLTP system

		Quantity Sold	
Date	**Order Number**	**Nokia**	**Sony Ericsson**
1 July 2008	I0001	5	2
	I0002	3	0
	I0003	2	6
	I0004	2	2
	I0005	3	3
2 July 2008	I0006	3	7
	I0007	3	1
	I0008	4	0

Table 2.1(b) is the output which is summarised or aggregated to daily totals.

Table 2.1(b): Records in Data Warehouse

	Quantity Sold	
Date	**Nokia**	**Sony Ericsson**
1 July 2008	15	13
2 July 2008	10	8

Separated Data Warehouse

- Why separate Data Warehouse?
  - Common question since operational databases already store large volumes of data
  - Having a separate data warehouse allows a few key benefits at the cost of physical resources
- High performance for both systems
  - DBMS is tuned for OLTP
    - Known tasks and workload
    - Access methods, indexing, concurrency control, recovery
  - Warehouse is tuned for OLAP
    - Complex OLAP queries and computation of large data
    - Requires multidimensional view, consolidation.
  - OLAP queries on operational DBs may substantially degrade OLTP tasks

Separated Data Warehouse

- Different functions and different data:
  - missing data: decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: decision support 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
- Separation between OLTP and OLAP systems are decreasing
  - Vendors attempt to optimize DBMS to support OLAP queries

Summary

- Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process.
- Many organisations have invested in data warehouses to support and improve business decision-making
- OLTP and OLAP systems are distinguishable with key features
- Separation of Data Warehouses from Operational Databases facilitates efficient processing