

Skill Based Subject – III: DATA MINING AND WAREHOUSING - 18BIT55S

UNIT V: Data Warehousing: Introduction – Operational Data Stores – Data Warehouses – Data warehouse Design – Guidelines for Data Warehouse Implementation – Data Warehouse Metadata. Online Analytical Processing (OLAP): Introduction – OLAP – Characteristics of OLAP Systems – Multidimensional View and Data Cube – Data Cube Implementation – Data Cube Operations.

TEXT BOOK

G.K Gupta, “Introduction to Data Mining with Case Studies”, Prentice Hall of India(Pvt) Ltd, India, 2008.

Prepared by : Mrs. G. Shashikala, Assistant Professor, PG Department of Information Technology

Introduction

- A large company might have the following systems :
 - Human resources
 - Financials
 - Billing
 - Sales leads
 - Web sales
 - Customer support

Such systems are called online transaction processing (OLTP) systems.

These are mostly relational database systems designed for transaction processing

- The performance of OLTP systems is very important because these are used to support the users (staff) that provide service to the customers
- These systems must be able to deal with insert, update operations and answering simple queries quickly
- But these systems are not designed to handle management queries efficiently
- In addition to OLTP systems (using relational database), an enterprise may use file systems using complex data structures.

- In such cases, there is no unified view of the enterprise and generating reports becomes difficult.
- The focus on operational managers is to improve business management and processes like customer support, inventory, marketing etc.
- To achieve this, they require:
 - A single sign-on path to the enterprise information
 - A single version of the enterprise information
 - A high level of data accuracy
 - A user-friendly interface to the information
 - Easy sharing of data across enterprise business units
- Some managers are not involved in operational management, but they look for overall trends and aggregations of the operational data
- Solutions to the variety of needs of management staff are :
 - Pose queries to a mediator system that decomposes each query into appropriate subqueries, obtain results, and then combines and presents the result to the user – this is called lazy or on-demand query processing. (the user is provided up-to-date information, but generates a heavy load on the OLTP systems)


- Another approach is that, one can collect the most common queries that the managers ask and have the results available – this is called eager approach (this is quick but the information is not up-to-date, and queries that are not pre-computed has to be run in the lazy mode)
- The third approach is to create a separate database and involves the following two steps :
 - (i) The information needs of the management staff are analysed, a suitable data model is developed, the information is then extracted and loaded in the new database
 - (ii) The new database is used to answer management queries and the OLTP systems are not accessed for such queries.

The two approaches of this type are the operational data store(ODS) and the data warehouse(DW). An ODS is a special type of DW

Operational data stores

- An ODS is designed to provide a consolidated view of the enterprises current operational information
- An ODS is defines as a subject-oriented, integrated, volatile, current valued data store, containing only corporate detailed data
- In contrast, a data warehouse doesn't contain operational data in it. A data warehouse is a reporting database that contains recent and historical data and also aggregate data
- An ODS may be of different types depending on the needs of the enterprise.
- An ODS could be a
 - Reporting tool – updated daily
 - Designed to track more complex information like product and location codes – database updated hourly
 - Designed to support customer relationship management
- Benefits of ODS
 - ODS is the unified operational view of the enterprise and provides improved access to important operational data
 - Assists in better understanding of business and customer

- ODS is effective in generating current reports without accessing the OLTP
- Shortens the time required to implement and populate a data warehouse system
- ODS becomes a source of reliable information for some other applications
- ODS design and implementation
 - To implement an ODS, a data model should be developed.
 - The database model should match the structure of the enterprise
 - The attributes that are needed by the operational management staff should be identified and included
 - The extraction of information from source databases has to be efficient and quality of data has to be maintained after each refresh

- 
- An ODS has to deal with integrity constraints like, existential integrity, referential integrity and null data
 - An ODS is a read-only database, refreshed by OLTP systems
 - Populating an ODS involves an acquisition process of extracting, transforming and loading data from OLTP source systems. This process is called ETL
 - Completing populating the database, checking for anomalies, and testing for performance are necessary before an ODS system can go online

Architecture of an ODS

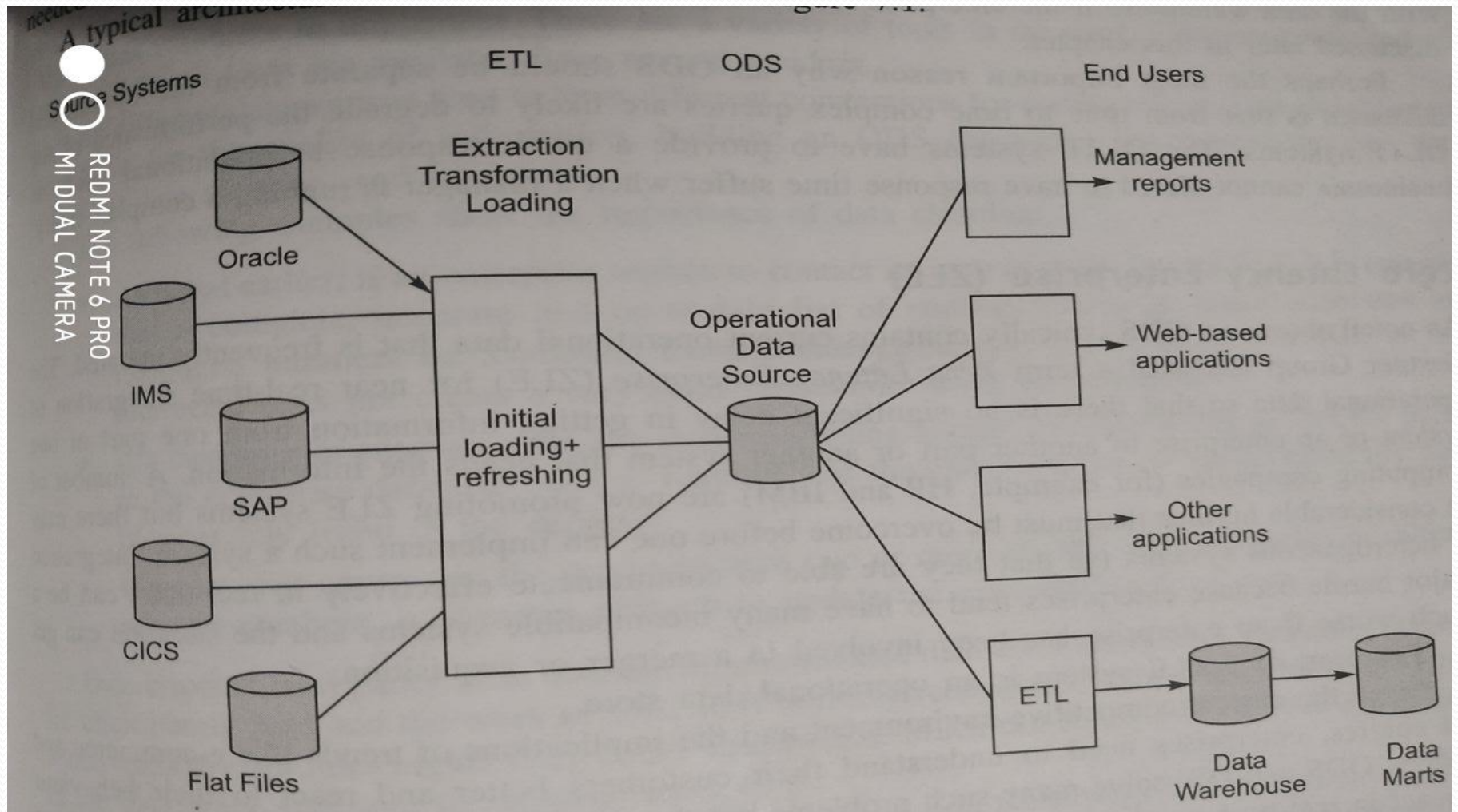


Figure 7.1 A possible Operational Data Store structure.

needs to be efficient and the quality of data checks

- Why a separate database ?
 - An ODS makes query processing more efficient for operational managers, without loading the OLTP system
 - ODS should be separate from the operational database since complex queries are likely to degrade the performance of the OLTP systems.
 - The OLTP systems has to provide a quick response to operational users
- Zero latency enterprise(ZLE)
 - Near real-time integration of operational data so that there is no delay in getting information from one part of an enterprise to another , that needs the information
 - The heart of a ZLE system is an operational data store
 - A ZLE is an ODS that is integrated and up-to-date
 - The aim of a ZLE data store is to allow management a single view of enterprise information by bringing together relevant data in real-time and providing management a 360-degree view of the customer
 - The characteristics of ZLE are
 - It has a unified view of the enterprise operational data
 - It has a high level of availability
 - Involves online refreshing of information
 - Supports large number of concurrent users (like call centre users)

Data warehouses

Data warehousing is a process for assembling and managing data from various sources for the purpose of gaining a single detailed view of an enterprise. It is an integrated subject oriented and time variant repository of information in support of management's decision making process

- The benefits of implementing a data warehouse are
- 1 To provide a single version of truth about enterprise information
- 2 To speed up ad hoc reports and queries that involve aggregations across many attributes
- 3 To provide a system in which managers who do not have a strong technical background are able to run Complex queries
- 4 To provide a database that stores relatively clean data
- 5 to provide a database that stores historical data that may have been deleted from the OLTP systems
-

ODS and DW Architecture

A typical ODS structure was shown in Figure 7.1. It involved extracting information from source systems by using ETL processes and then storing the information in the ODS. The ODS could then be used for producing a variety of reports for management.

The architecture of a system that includes an ODS and a data warehouse shown in Figure 7.4 is more complex. It involves extracting information from source systems by using an ETL process area. Another ETL process is used to transform information from the staging area to populate the warehouse which in turn feeds a number of data marts that generate the reports required by management. It should be noted that not all ETL processes in this architecture involve data cleaning, some may only involve data extraction and transformation to suit the target systems.

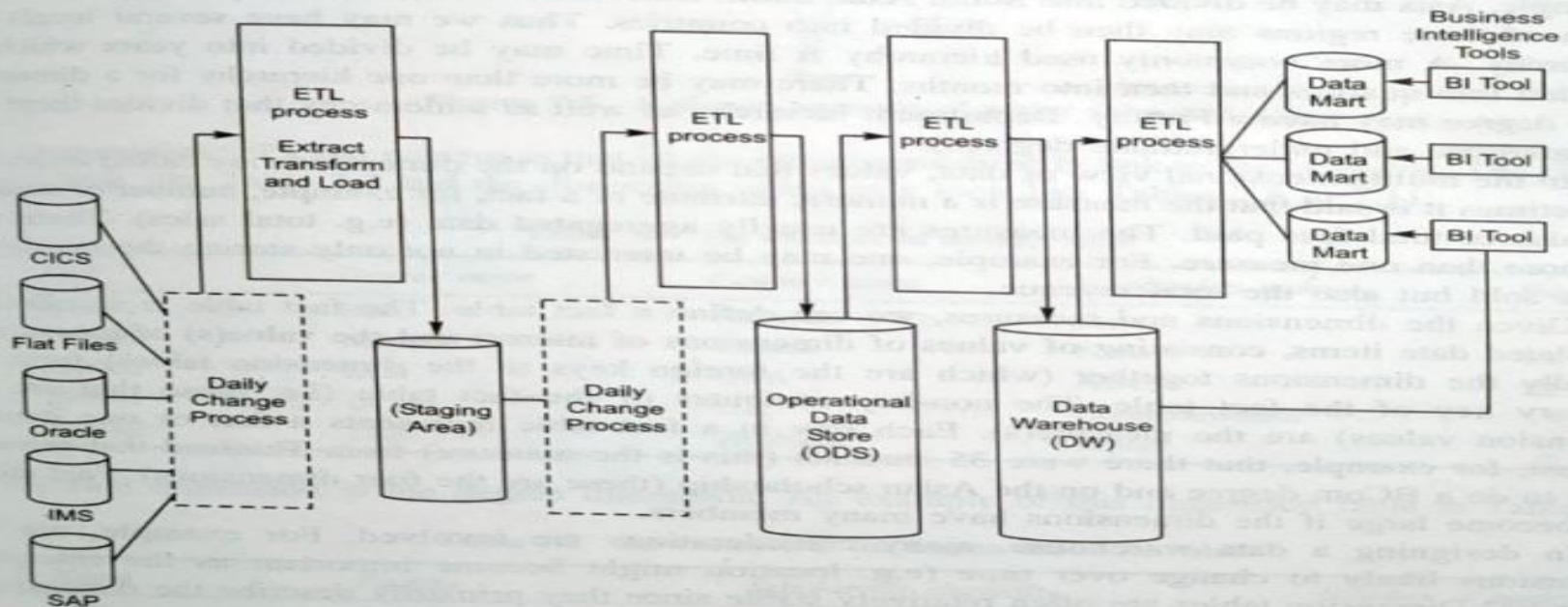


Figure 7.4 Another structure for ODS and DW.

7.5 DATA WAREHOUSE DESIGN

There are a number of ways of conceptualizing a data warehouse. One approach is to view it as a three-level structure. The lowest level consists of the OLTP and legacy systems, the middle level consists of the global or central data warehouse while the top level consists of local data warehouses

Datawarehouse metadata

- Meta data is data about data or documentation about the data that is needed by the users It is not the actual data warehouse but answers the “Who, what, where, when, why and how” questions about the data warehouse. Meta data is a structured data which describes the characteristics of a resource.
- some examples of meta data are are
- 1 a library catalogue
- 2 the table of contents and the index in a book
- 3 the tables and figures in a document like book

Online analytical processing(OLAP)

- A dimension is an attribute within a multidimensional structure consisting of a list of values.
- Dimensions are used for selecting and aggregating data at the desired level of detail. A dimension does not include ordering of values. But a dimension may have one or more hierarchies that show parent/child relationships between the members of a dimension.
- OLAP
- It is a software technology concerned with fast analysis of enterprise information. Often OLAP systems are data warehouse front-end software tools to make aggregate data available efficiently, for advanced analysis, to an enterprise's managers. It is implemented as a special database (eg a data warehouse) to improve response time

- Business intelligence is used to mean both data warehousing and OLAP
- Characteristics of OLAP systems:
 - The differences between OLTP and OLAP are :
 - 1 Users : OLTP systems are designed for office workers while the OLAP systems are designed for decision makers
 - 2 functions : OLTP systems support an enterprise's day to day operations and are performance and availability driven. OLAP systems support an enterprise's decision support functions using analytical investigations. They are more functionality driven.
 - 3 Nature : although SQL queries return a set of records, OLTP systems are designed to process one record at a time. OLAP systems are not designed to deal with individual customer records. They involve queries that deal with many records at a time and provide summary data to the manager. OLAP systems involve data stored in a data warehouse that has been extracted from many tables

4 Design : OLTP database systems are designed to be application oriented while OLAP systems are designed to be subject oriented . OLTP systems view the enterprise data as a collection of tables based on an Entity relationship model . OLAP systems view enterprise information as multidimensional

5 Data : OLTP systems normally deal with the current status of information. On the other hand OLAP systems require historical data over several years since trends are often important in decision making

6 Kind of use : : OLTP systems are used for read and write operations while OLAP systems normally do not update the data

- **Multidimensional view and the data cube**
- The multidimensional view of data is a natural view of any enterprise for managers. The triangle diagram shows that as we go higher in the triangle hierarchy the managers need for detailed information declines

Data cube implementations

- When millions of aggregates are likely in a large enterprise some possible solutions are :
 - 1 precompute and store all
 - 2 precompute (and store) none
 - 3 precompute and store some

Data cube operations

- The common operations applied to data cube are :
Roll-up, Drill down , slice and dice and pivot

- Roll up

- Roll up is like zooming out on the data cube. It is required when the user needs further abstraction or less detail . This operation performs further aggregations on the data
- Drill down
- Drill down is like zooming in on the data and is therefore the reverse of roll-up . It is an appropriate operation when the user needs further details or when the user wants the partition more finely or wants to focus on some particular values of certain dimensions. Drill down adds more details to the data.
- slice and dice
- slice and dice are operations for browsing the data in the cube . The terms refer to the ability to look at information from different viewpoints
- pivot or rotate
- The pivot operation is used when the user wishes to re-orient the view of the data cube. It may involve swapping the rows and columns, or moving one of the row dimensions into the column dimension

