Data warehousing

Data Warehouse

What is a Data Warehouse?

- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process
- **Subject Oriented:**

  Data that gives information about a particular subject instead of about a company's ongoing operations.
- **Integrated:**

  Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- **Time-variant** : A time dimension is explicitly included in the data so that trends & changes over time can be studied.
- **Non-volatile :** The data in the DW is not as volatile as data in an operational database.

In the data warehouse, data is not stored by operational applications, but by business subjects.

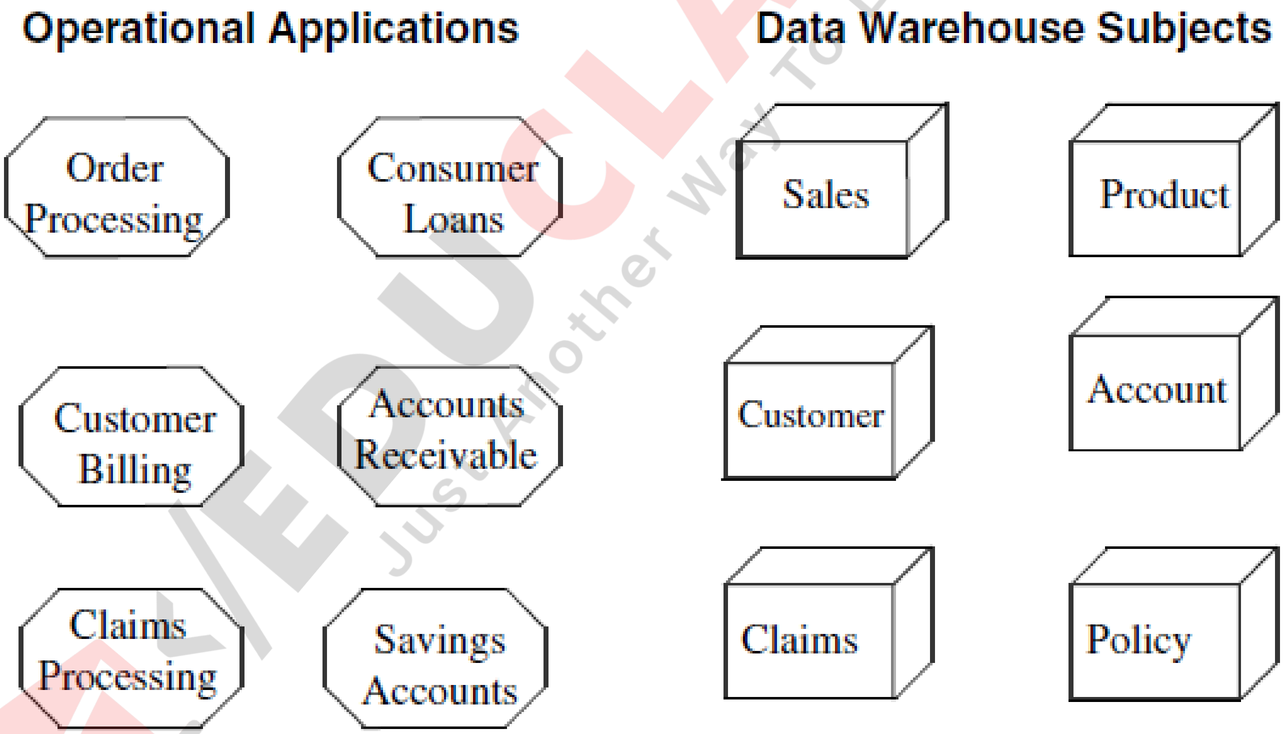


Figure 2-1 The data warehouse is subject oriented.

Data inconsistencies are removed; data from diverse operational applications is integrated.

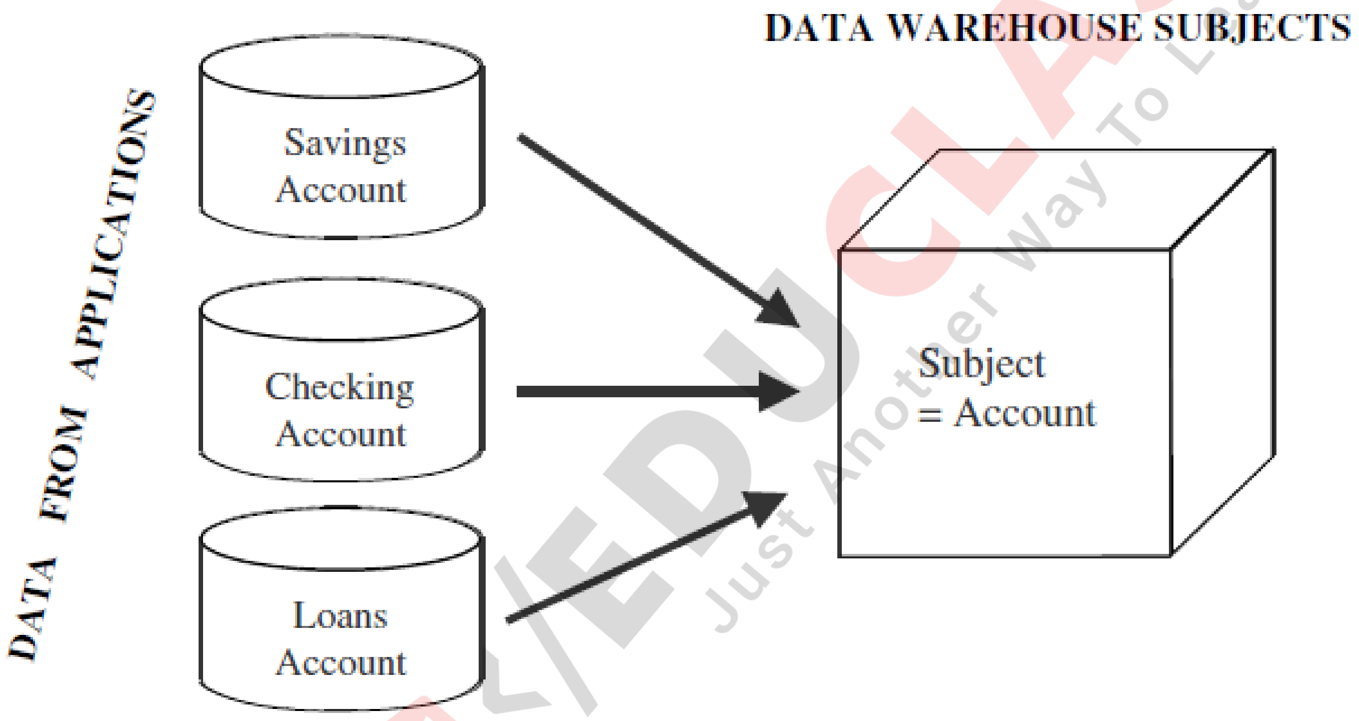


Figure 2-2 The data warehouse is integrated.

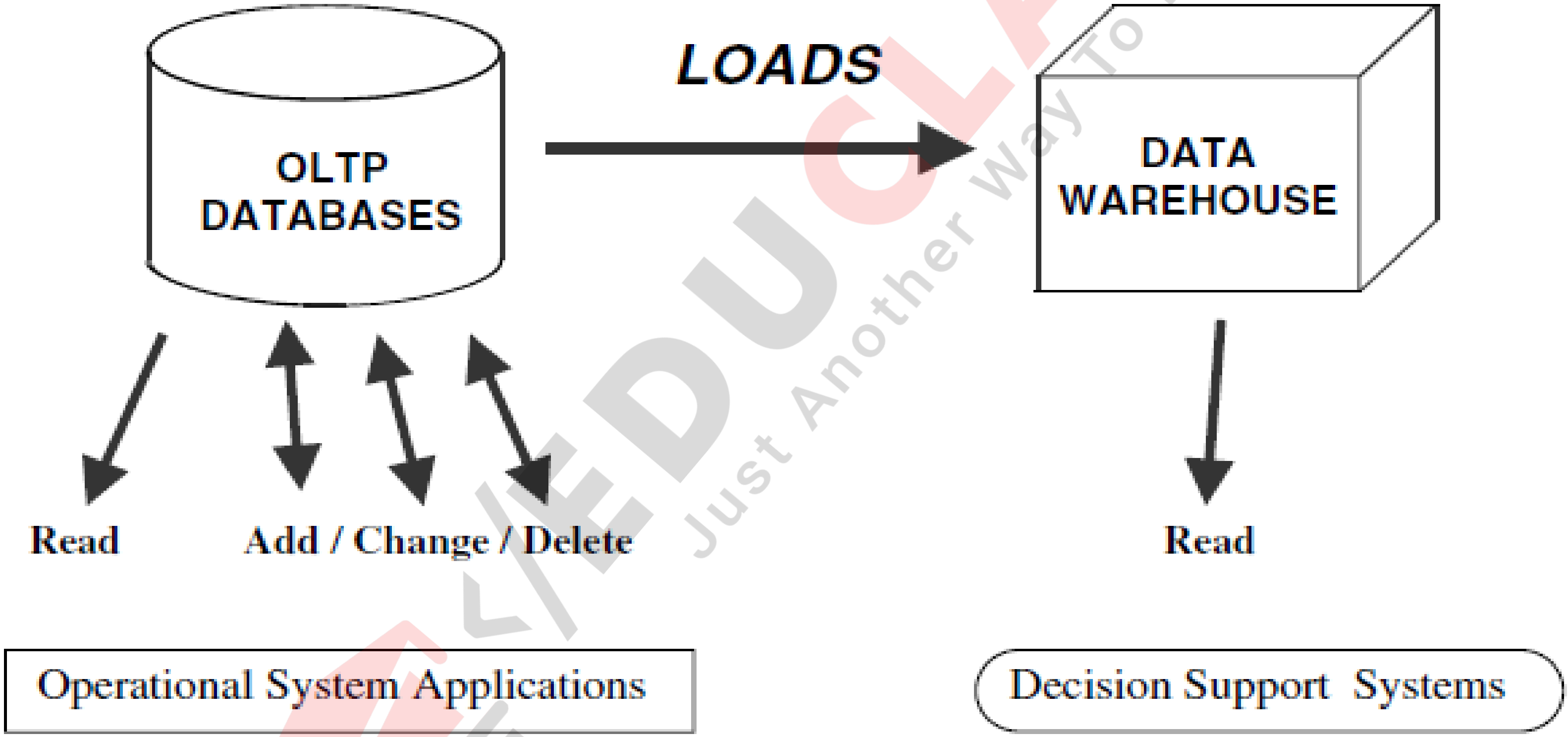
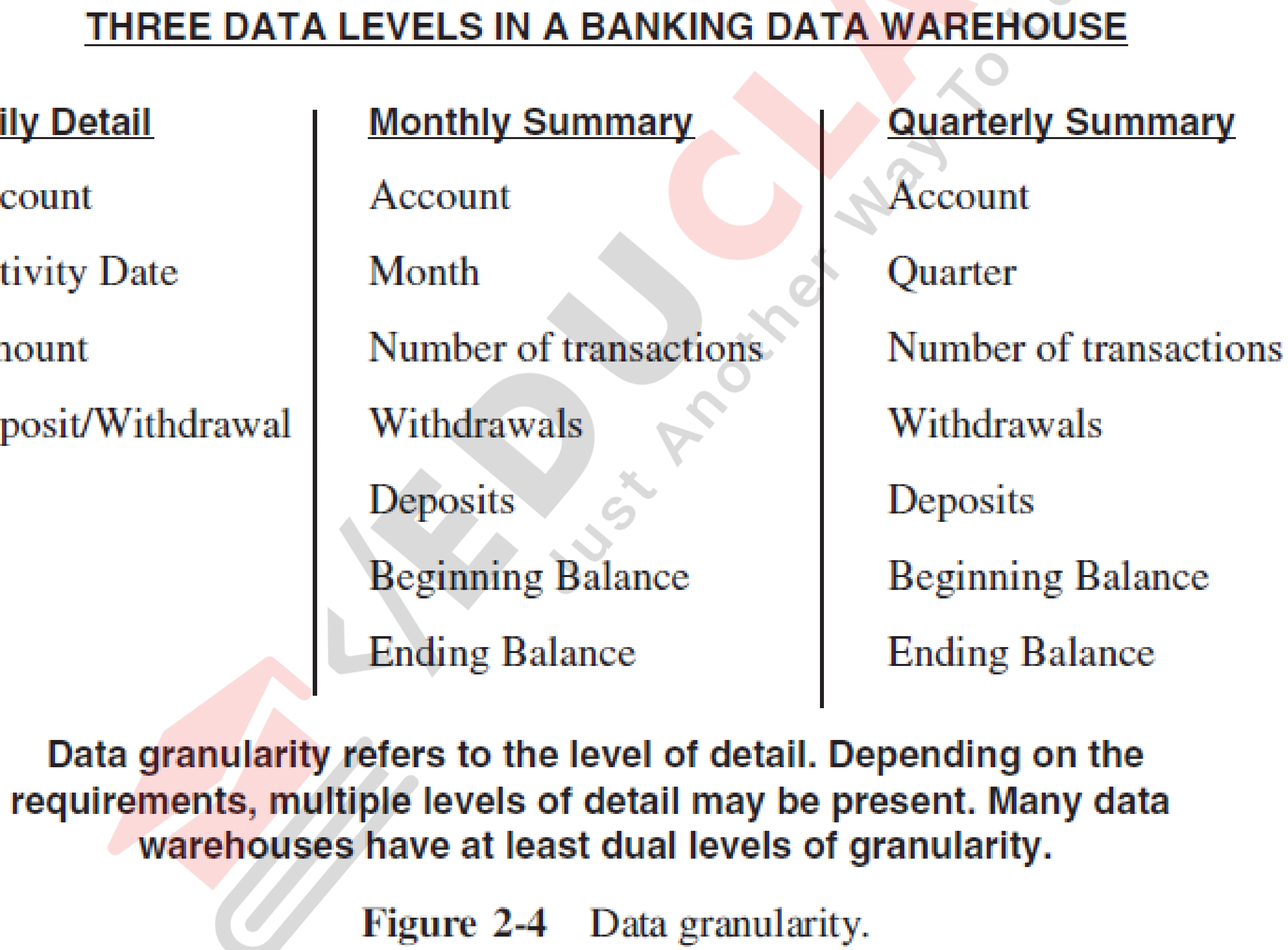


Figure 2-3 The data warehouse is nonvolatile.

Data Granularity

- Refers to the level of detail.
- Keeping data summarized at diff levels.



THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

Daily Detail	Monthly Summary	Quarterly Summary
Account	Account	Account
Activity Date	Month	Quarter
Amount	Number of transactions	Number of transactions
Deposit/Withdrawal	Withdrawals	Withdrawals
	Deposits	Deposits
	Beginning Balance	Beginning Balance
	Ending Balance	Ending Balance

Data granularity refers to the level of detail. Depending on the requirements, multiple levels of detail may be present. Many data warehouses have at least dual levels of granularity.

Figure 2-4 Data granularity.

Data warehousing

- Data Warehouse is a central managed and integrated database containing data from the operational sources in an organization (such as SAP, CRM, ERP system).

- DW is a repository of information gathered from multiple sources, stored under a unified schema, at a single site.

- DW provide the user a single consolidated interface to data, making decision support queries easier to write.

- It may gather manual inputs from users determining criteria and parameters for grouping or classifying records.
- DW holds business intelligence for the enterprise.

Why do we need DW?

- Operational databases are for On Line Transaction Processing (OLTP)
  - automate day-to-day operations (purchasing, banking etc).
  - transactions access (and modify) a few records at a time.
  - database design is application oriented.
  - metric: transactions/sec .
- Data Warehouse is for On Line Analytical Processing (OLAP)
  - complex queries that access millions of records.
  - need historical data for trend analysis .
  - metric: query response time.

Data warehousing

- A source for the data warehouse is a data extract from operational databases.
- The data is validated, cleansed, transformed and finally aggregated and it becomes ready to be loaded into the data warehouse.
- Sometimes, where only a portion of detailed data is required, it may be worth considering using a **data mart**.
- A smaller collection, usually relating to one specific aspect of an organization is called data mart.
- A data mart is generated from the data warehouse and contains data focused on a given subject and data that is frequently accessed or summarized.

ARCHITECTURAL FRAMEWORK

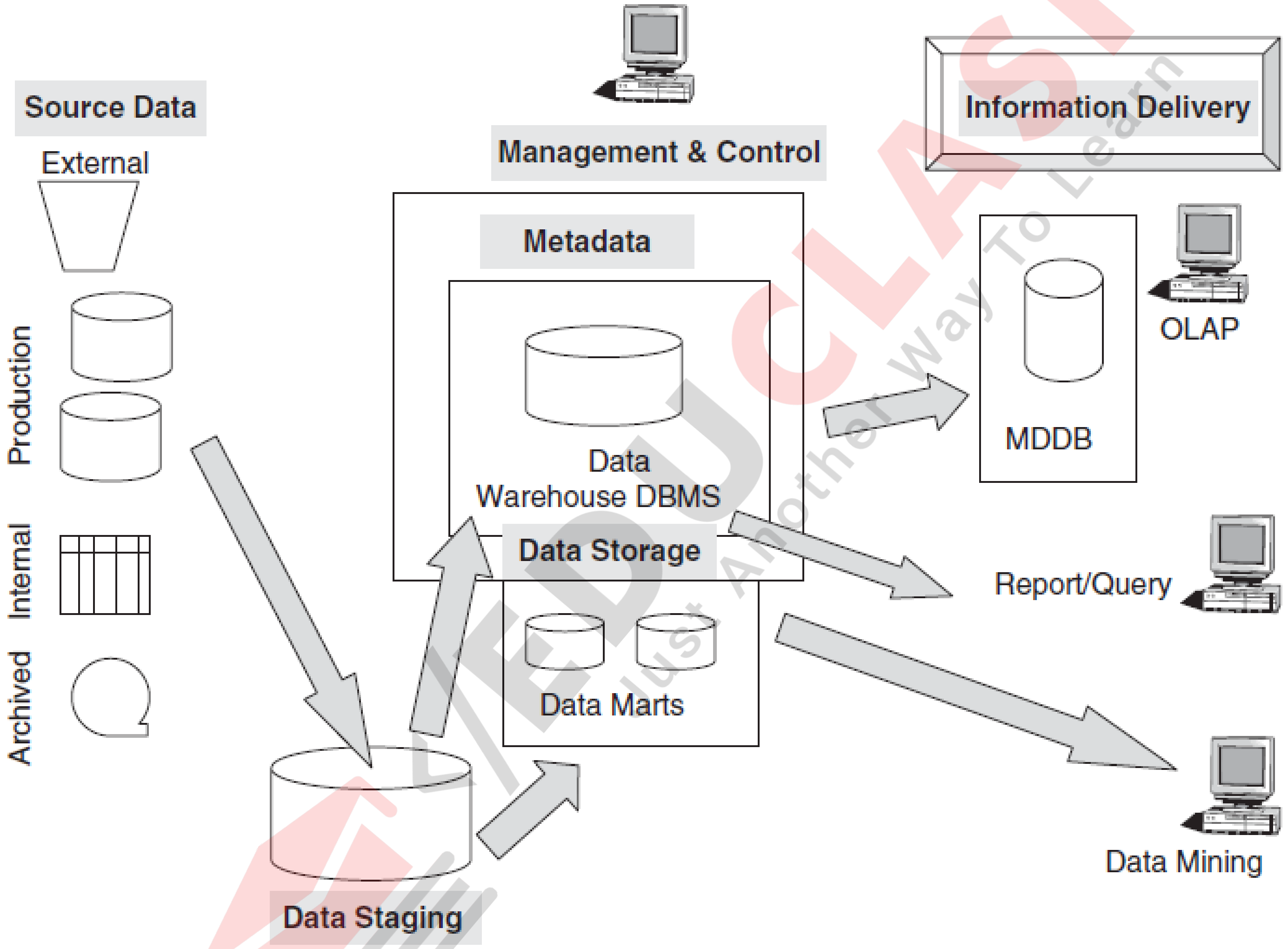


Figure 7-2 Architectural framework supporting the flow of data.

Extraction

Data Extraction Issues

- **Source identification-** identify source applications and source structures.
- **Method of extraction-** for each data source, define whether the extraction process is manual or tool-based.
- **Extraction frequency-** for each data source, establish how frequently the data extraction must by done—daily, weekly, quarterly, and so on.
- **Time requirement-** for each data source, denote the time window for the extraction process.
- **Job sequencing-** determine whether the beginning of one job in an extraction job stream has to wait until the previous job has finished successfully.

2. Transformation

- Transformation: Transform extracted data into appropriate format, structure and values that are required by data warehouse.

- Transformation: changing the units of measure or converting data to a different schema by joining data from multiple source relations.

- It is the process of dealing with inconsistencies.

Transformation

Transformation types

- Format revisions. Eg: field length.
- Decoding fields. Eg: M / Male.
- Calculated & derived values. Eg: cost, profit, average daily balance
- Splitting of single fields – Eg: name.
- Merging of information- combination of pdt code,description,packg types & cost into a single entity.
- Character set conversion- Eg:EBCDIC to ASCII format
- Conversion of unit of measure. Eg: weight.
- Date/ time conversion - To standard format.
- Summarization-avoiding detailed data

3. Cleansing/cleaning

- The task of correcting & preprocessing data is called data cleansing.

- This deals with many types of possible errors like missing data and incorrect data in one source.

- Inconsistent data & conflicting data when 2 or more sources are involved etc.

4. Loading

- Data loaders load transformed data into the data warehouse.
- It is the process of moving data into the DW repository.

It can be done through

- **Initial Load** —populating all the data warehouse tables for the very first time.
- **Incremental Load** —applying ongoing changes as necessary in a periodic manner.
- **Full Refresh** —completely erasing the contents of one or more tables and reloading with fresh data

Data Mart

- From a data warehouse data flows to various dept for their customized DSS(Decision support system) usage. These individual dept components are called data marts.
- A data warehouse is only a collection of data marts.
- Data mart is loaded with data from a data warehouse by means of a load program.

Advantages of data marts

1. Easily customize, summarize & analyze.
2. The processing load or overhead is very limited.
3. Cost of processing data is reduced.

Metadata

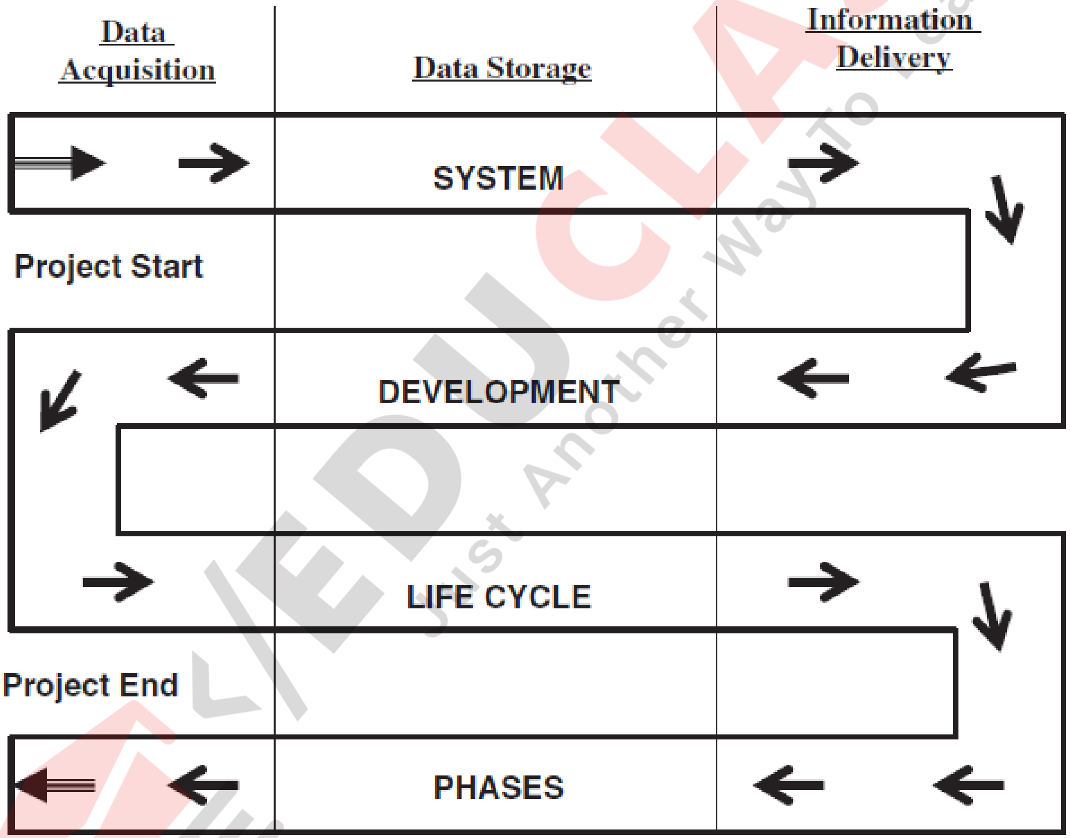
Categories

- **Operational metadata** -contains infor (like field length, data type) about the operational data sources.

- **Extraction & transformation metadata-** contains data about extraction methods of data from source systems(extraction frequencies, extraction methods & business rules for data extraction).

- **End-user metadata-** A navigational map of DW. Enables users to find information from DW. Allows to use their own terminology for looking information.

DW Life Cycle



Data Acquisition

Data Storage

Information Delivery

SYSTEM

Project Start

DEVELOPMENT

LIFE CYCLE

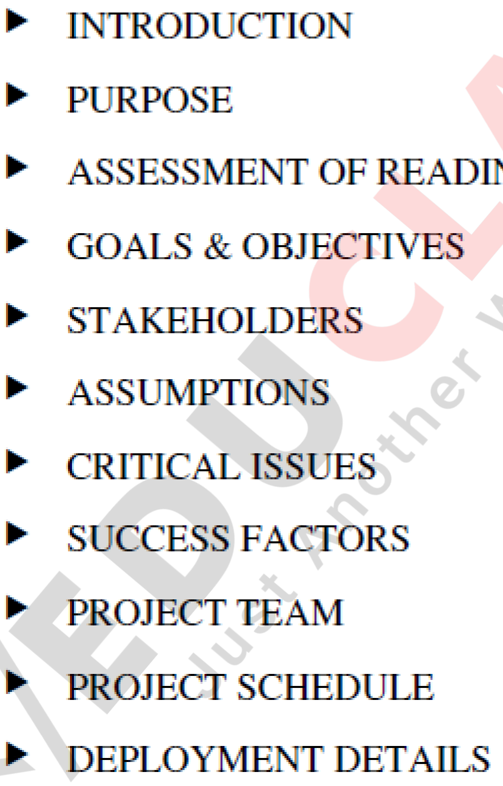
Project End

PHASES

Figure 4-3 Date warehouse functional components and SDLC.

DW Project

The Development Phases



- INTRODUCTION
- PURPOSE
- ASSESSMENT OF READINESS
- GOALS & OBJECTIVES
- STAKEHOLDERS
- ASSUMPTIONS
- CRITICAL ISSUES
- SUCCESS FACTORS
- PROJECT TEAM
- PROJECT SCHEDULE
- DEPLOYMENT DETAILS

Figure 4-4 Data warehouse project plan: sample outline.

DW Project

The Development Phases

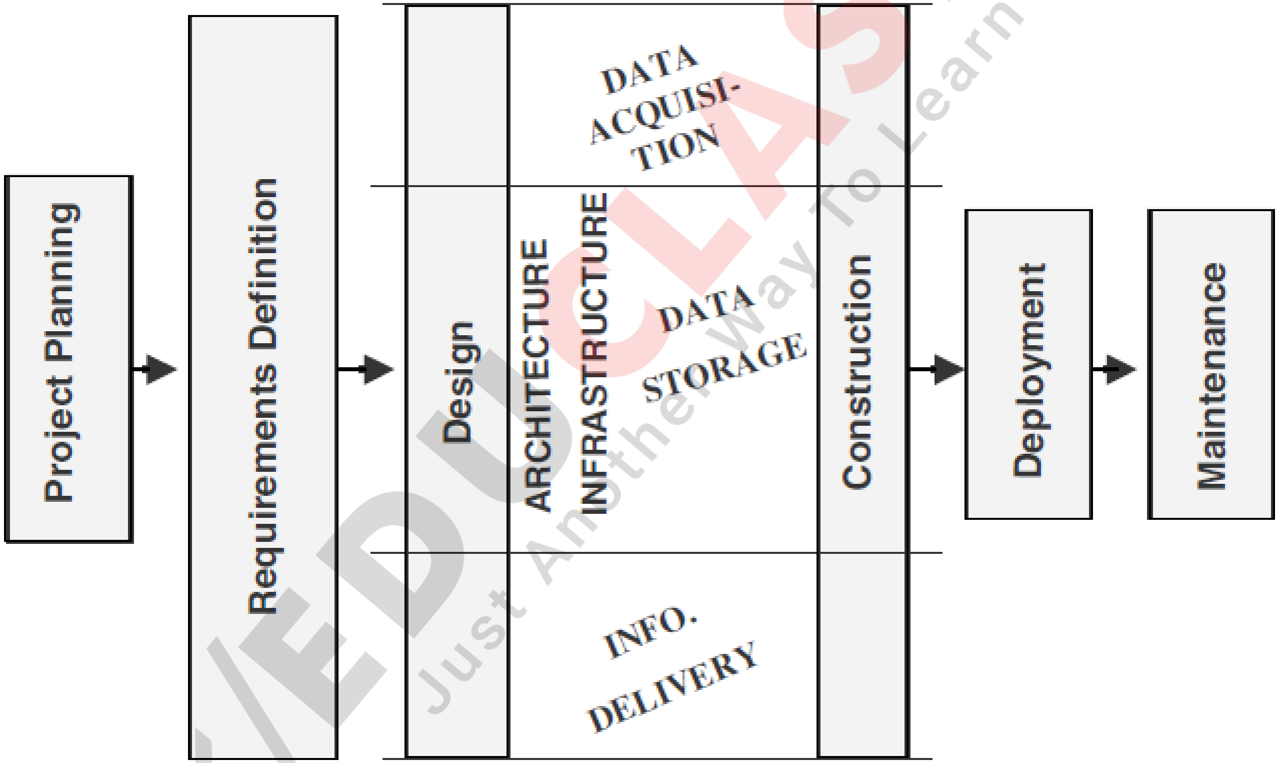


Figure 4-5 Data warehouse development phases.

- Project plan
- Requirements definition
- Design
- Construction
- Deployment
- Growth and maintenance

Creating and Maintaining a Warehouse

- After data is loaded into a warehouse, additional measures must be taken to ensure that the data in the warehouse is periodically **refreshed to reflect updates to the data** sources and to periodically **remove data that is too old from the warehouse.**
- An important task in maintaining a warehouse is keeping track of the data currently stored in it.
- This is done by storing information about the warehouse data in the system catalogs.
- The system catalogs associated with a warehouse are very large and are often stored and managed in a separate database called a **metadata repository.**

5. Summarization

E.g. storing sales by product by store by day.

- Once DW database has been loaded it is possible to create summaries.
- It must usually be re-created after every incremental update, as any changes in the underlying data may impact them.

Advantages:

- ➢ Queries that use the prestored summaries can be answered quickly.

Disadvantages:

- ➢ Calculating summaries requires computer time & resource.
- ➢ The summaries occupy space on a mass storage device.
- ➢ Someone has to figure out which summaries to prestore.
- ➢ Someone must define the summaries to the DW software.

comparision of OLTP systems and data warehousing system

OLTP systems	Data warehousing systems
Hold current data Stores detailed data Data is dynamic Repetitive processing High level of transaction throughput Predictable pattern of usage Transaction-driven Application-oriented Supports day-to-day decisions Serves large number of clerical/operation users	Holds historical data Stores detailed and highly summarized data Data is largely static Ad hoc, unstructured, and heuristic processing Unpredictable pattern of usage Analysis driven Subject-oriented supports strategic decisions Serves relatively Low number of managerial users