Data Warehouse

▶ A data warehouse is a historical database and should be considered as the long term memory of an organization. Historical data is not to be tampered with; no insertion, up-dation and deletion are to be made. Usually, it is used only for retrieval such as verification and data analysis. Data warehouses are typically used for Online Analytical Processing (OLAP) to support management queries.
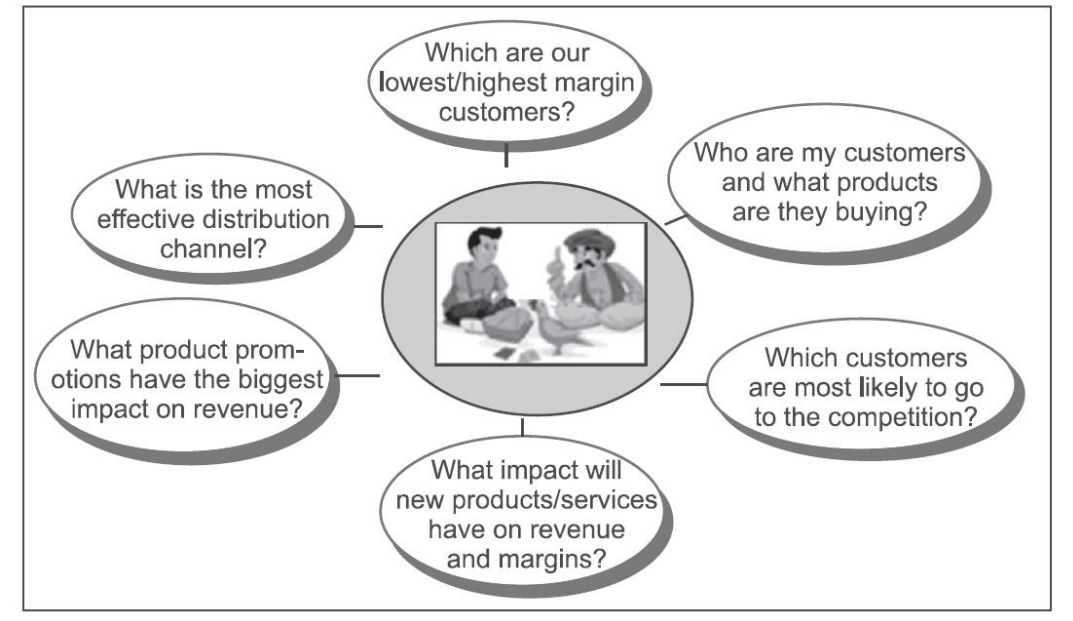
Answering management queries

Defining Data Warehousing

- Data warehouse could be defined as, 'A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.'

- **Subject-Oriented:** Data warehouses are subject oriented - they are built around the major data entity or subjects of an organization.

- **Integrated:** A data warehouse integrates (combines) data from multiple systems to provide a wide view of any enterprise's data.

- **Time variant:** Data warehouses data is not always up to date as it contains historical data which is valid or accurate till some point of time (or time interval).

- **Non-volatile:** Data warehouse is described as a long term enterprise memory due to its non-volatile nature because the data is not updated in real time but is rather refreshed on a regular basis.

Data warehouse architecture
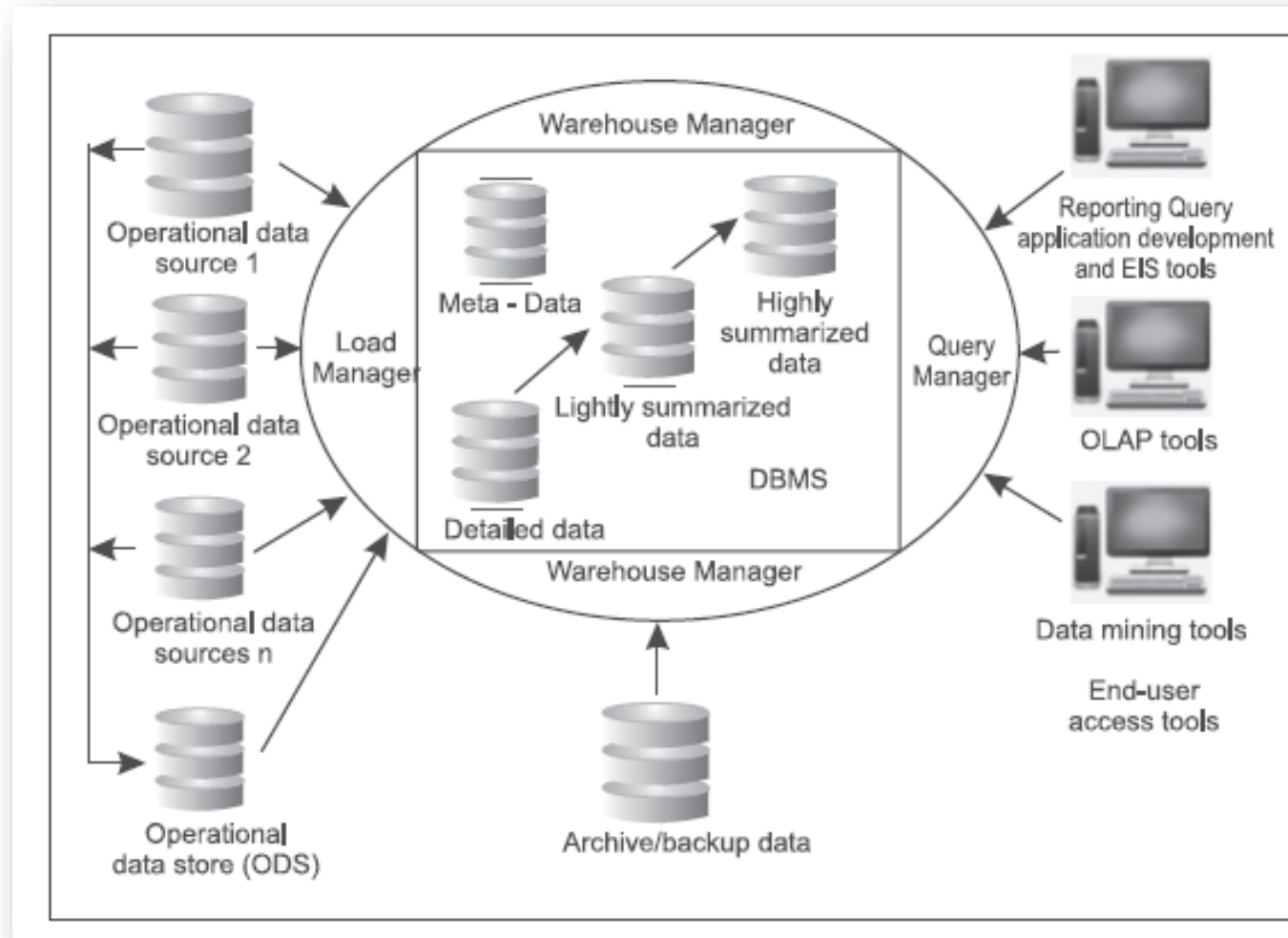
Data warehouse architecture

▶ **Load manager**

The Load manager is responsible for Data collection from operational systems.

▶ **Warehouse Manager**

The Warehouse manager is the main part of Data Warehousing system as it holds the massive amount of information from myriad sources. It organizes data in a way so it becomes easy for anyone to analyze or find the required information.

▶ **Query manager**

Finally the Query manager is that interface which connects the end users with the information stored in data warehouse through the usage of specialized end-user tools. These tools are known as Data mining access tools.

Benefits of data warehousing

▶ Some of its advantages and benefits are given below:

• Potential high ROI (Return on Investment)

• Unbeatable competitive advantage

• High Productivity in corporate decision making and business intelligence

• Cost effective

• Enhanced customer service

Limitations of data warehousing

- Underestimation of resources for data ETL
- Erroneous source systems
- Required data not captured
- Increased end user queries or demands
- Loss of information during data homogenization
- High demand of resources
- Data ownership
- Necessary maintenance
- Long-duration projects
- Complexity of integration

Data Marts

▸ Data mart can be defined as 'a specialized, subject-oriented, integrated, time-variant, volatile data store in support of specific subset of management's decisions'.
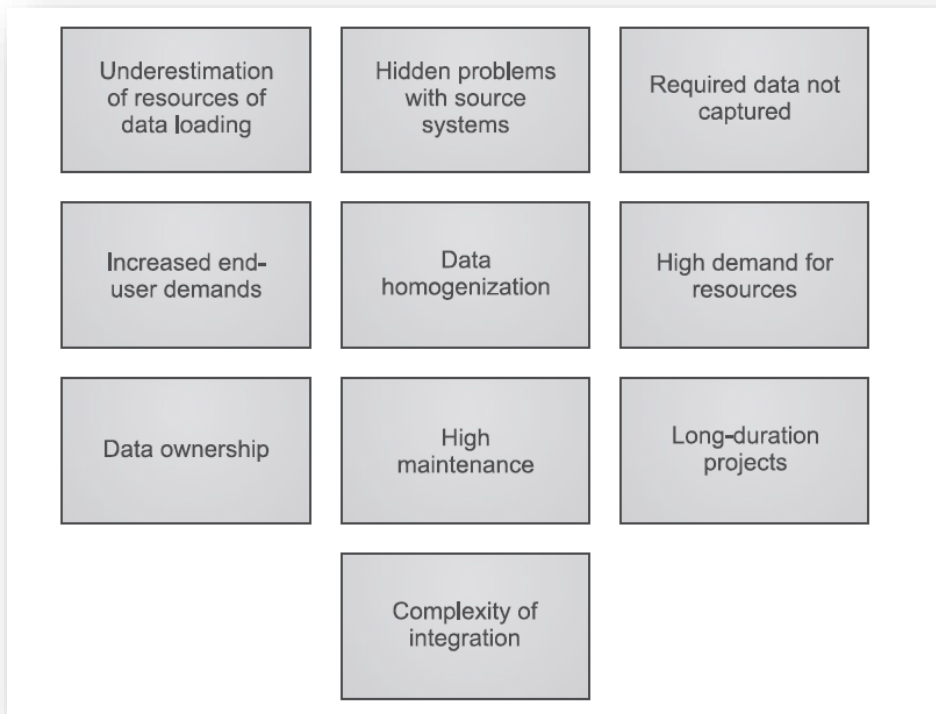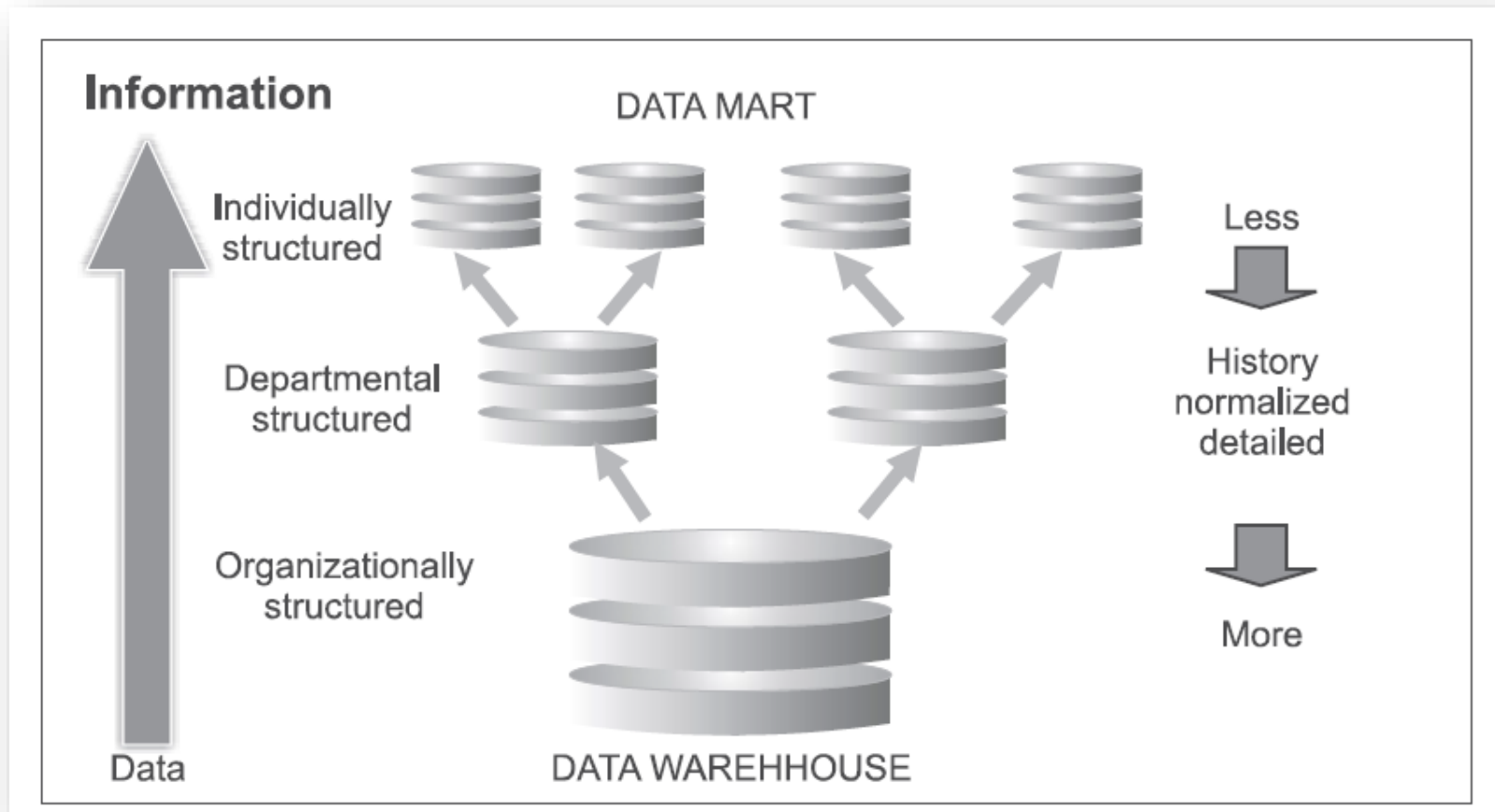

Limitations of data warehousing
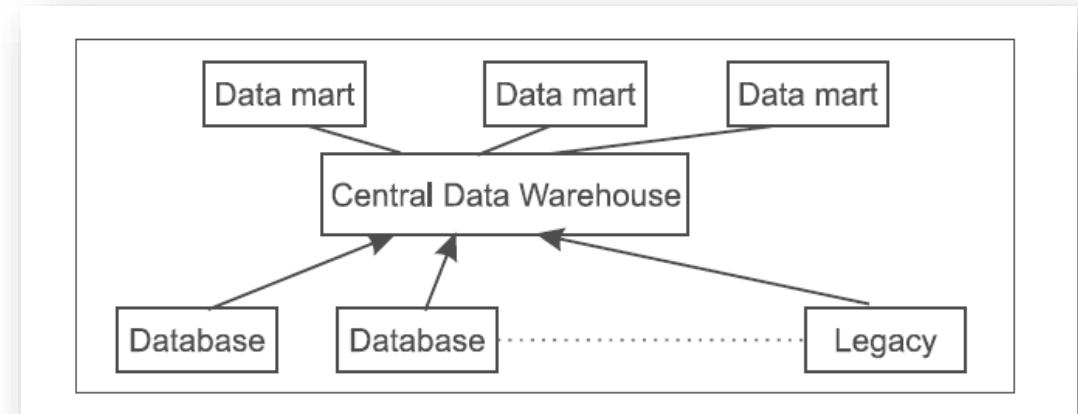
Data mart and data warehouse

Data Marts

Relationship between data mart and data warehouse

Comparative Study of Data Warehouse with OLTP and ODS

ODS	Data Warehouse
High quality of data and easy availability of data is assured.	Imperfect Data which is sufficient for doing strategic analysis. Availability of data is not a primary concern.
Contains both Real Time and Current data.	Contains Historical data
Updated at the Data field level.	Normally, batch data loads. Data is appended, and not updated .
Typically detailed data only.	Contains summarized and detailed data.
Modeled to support rapid data updates (3NF).	Variety of modeling techniques used, typically multi-dimensional for data marts to optimize query performance.
Transactions resemble OLTP system transactions.	Queries are quite complex and even the data volumes are larger.
Aids in detailed decision making and operational reporting and hence it is used mostly at an operational level.	Aids in long term planning and decision making along with management reporting.

Data warehouses versus OLTP: similarities and distinction

OLTP System	Data Warehousing System
They hold Current data.	Designed to hold only Historical data.
Detailed data is stored.	It is flexible enough to store detailed, light or profoundly summarized data.
Nature of data is Dynamic.	Nature of data is Static.
Processing is recursive in nature.	Processing is rather unstructured and ad-hoc.
Transaction output is of high level.	Here, it is low or medium mostly.
Pattern of usage is predictable.	Pattern of usage is unpredictable.
Transaction-driven	Analysis driven
OLTP systems are application oriented.	They focus more on subject.
Day-to-day decisions are supported.	Long term or strategic decisions are taken through them.
Massive numbers of operational users are served.	Mostly for managerial users.