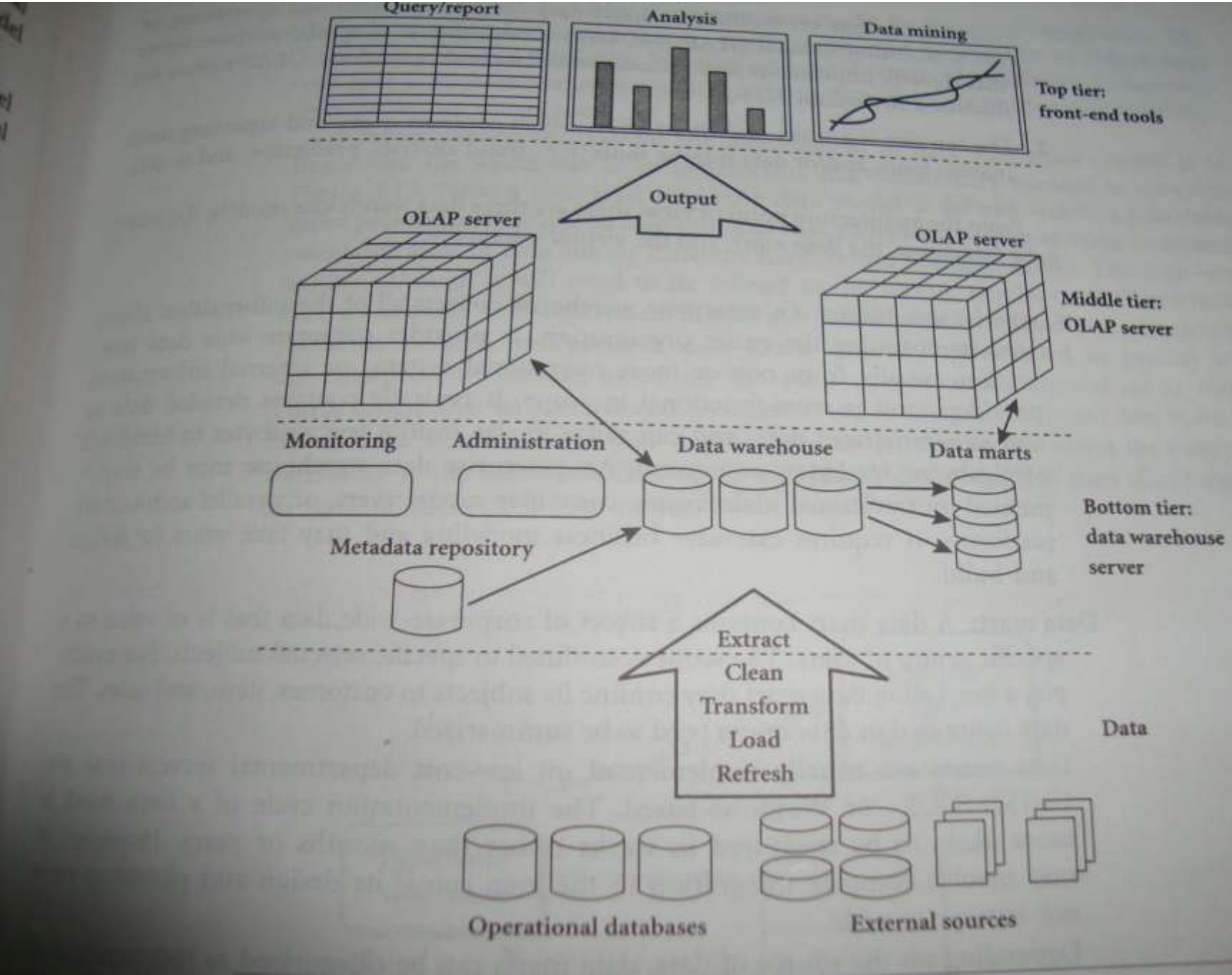
Data Warehouse Architecture

A Three-Tier Data Warehouse Architecture



Warehouse database server

- The bottom tier is a **warehouse database server** that is almost always a **relational database** system.
- **Back-end tools and utilities** are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning, and transformation (e.g. to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse.
- The data are **extracted** using application program interfaces known as **gateways**. A gateway is supported by the underlying DBMS and **allows client programs to generate** SQL code to be executed at a server. Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection). The tier also contains a metadata repository, which stores information about the data warehouse and its contents.

OLAP Server

The middle tier is an **OLAP server** that is typically implemented using either

- a **relational OLAP (ROLAP)** model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations.

  It can handle large amount of data but performance can be slow and limited by SQL Functionalities. or
- a **multidimensional OLAP (MOLAP)** model, that is a special-purpose server that directly implements multidimensional data through array based and operations. It allows fast indexing for summarized data. More traditional way of OLAP Analysis.

  It can perform complex calculations and excellent performance but limited in the amount of data it can handle and requires additional investment.

Front-end Client Layer

- The top tier is a **front-end client layer**, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g. trend analysis, prediction, and so on).

Data Warehouse Models

- From the architecture point of view there are three data warehouse models;
  1. the ***enterprise warehouse***,
  2. the ***data mart***,
  3. the ***virtual warehouse***.

Enterprise warehouse

Enterprise warehouse:

- An enterprise warehouse collects all of the information **about subjects** spanning the entire organization.
- It provides corporate-wide data integration, usually from one or more operational systems or external information providers.
- It typically contains **detailed data** as well as **summarized data**, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.
- An enterprise data warehouse may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms.
- It requires extensive business modeling and may take years to design and build.

Data mart

Data mart:

- A data mart contains a **subset of corporate-wide data** that is of value to a **specific group of users**.
- The **scope is confined to specific selected subjects.** For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in the data marts tend to be summarized.
- Data marts are usually implemented on low-cost departmental servers that are **UNIX/LINUX or Windows-based.** The implementation cycle of a data mart is more likely to be measured in **weeks rather than months or years.** However, it may involve complex integration in the long run if its design and planning were not enterprise-wide.
- Depending on the source of data, data marts can be categorized as independent or dependent. ***Independent*** data marts are sourced from one or more operational systems or external information providers, or from data generated locally within a particular department or geographical area. ***Dependent*** data marts are sourced directly from enterprise data warehouses.

Virtual warehouse

- A virtual warehouse is a **set of views** over operational databases.
- For efficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.