Introduction

- The Past and The Problem
- What is a Data Warehouse?
- Components of a Data Warehouse
- OLAP, Metadata, Data Mining
- Getting the Data in
- Benefits vs. Costs
- Conclusion & Questions

Data Warehousing

DBMS

- stores data in the form of tables
- ER model
- ACID

Data Warehouse

- stores a huge amount of data
- collected from multiple heterogeneous sources
  - Files
  - DBMS
- help in decision makings

Data Warehouse - Why?

- Database can store MBs to GBs of data
  - specific purpose
- The storage shifted to Data Warehouse storage
  - stores TBs of data

Data Warehouse - Benefits

- Business analytics
- Faster Queries
- Improved data Quality
- Historical Insight

SNO	Database Management System (DBMS)	Data Warehouse
1	Transaction processing	Analytical Processing
2	Data for daily operations are stored	Historical data are stored
3	Application Specific	
4	Not Expensive	Expensive
5	Runs the business	How to run the business

Applications of Data Warehousing

- Social Media websites -
- Banking – Analyse spending patterns
- Government – Analyse tax payments ,detect tax theft

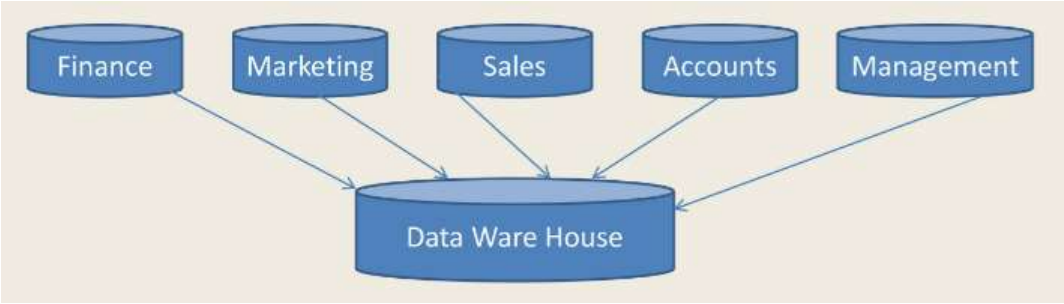
Data Warehousing

- Process of transforming data into information
- Use information for decision making

OLTP- online transaction processing

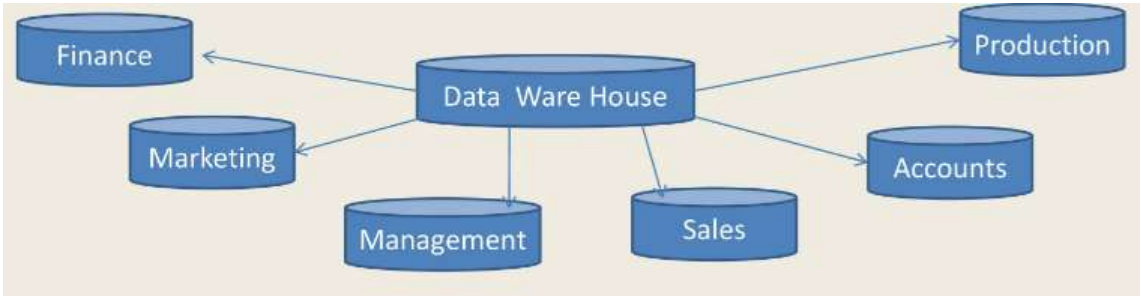
Data Warehouse design- Botton up Approach

- Create small data marts
- Combine data marts to large business



Data Warehouse design- Top Down Approach

- Create data warehouse
- As per specific business needs create data marts



The Past and The Problem

- Only had scattered transactional systems in the organization – data spread among different systems
- Transactional systems were not designed for decision support analysis
- Data constantly changes on transactional systems
- Lack of historical data
- Often resources were taxed with both needs on the same systems

The Past and The Problem

- Operational databases are designed to keep transactions from daily operations. It is optimized to efficiently update or create individual records
- A database for analysis on the other hand needs to be geared toward flexible requests or queries (Ad hoc, statistical analysis)

What is a Data Warehouse?

Data warehousing is an architectural model designed to gather data from various sources into *a single unified data model for analysis purposes.*

What Is a Data Warehouse?

Term was introduced in 1990 by William Immon

A managed database in which the data is:

- Subject Oriented
- Integrated
- Time Variant
- Non Volatile

Subject Oriented

- Organized around major subject areas in the enterprise (Sales, Inventory, Financial, etc.)
- Only includes data which is used in the decision making processes
  - Elements used for transactional processing are removed

Integrated

- Data from different sources are brought together and consolidated
- The data is cleaned and made consistent

Example – Bank Systems using Different Codes

Loan Department – COMM

Transactional System - C

Time Variant

- Data in a Data Warehouse contains both current and historical information
- Operational Systems contain only current data

Systems typically retain data:

Operational Systems – 60 to 90 Days

Data Warehouse – 5 to 10 Years

Non Volatile

- Operational systems have continually changing data
- Data Warehouses continually absorb current data and integrates it with its existing data (Aggregate or Summary tables)

Example of volatile data would be an account balance at a bank

What Is a Data Warehouse?

- Not a product, it is a process
- Combination of hardware and software
- Concept of a Data Warehouse is not new, but the technology that allows it is

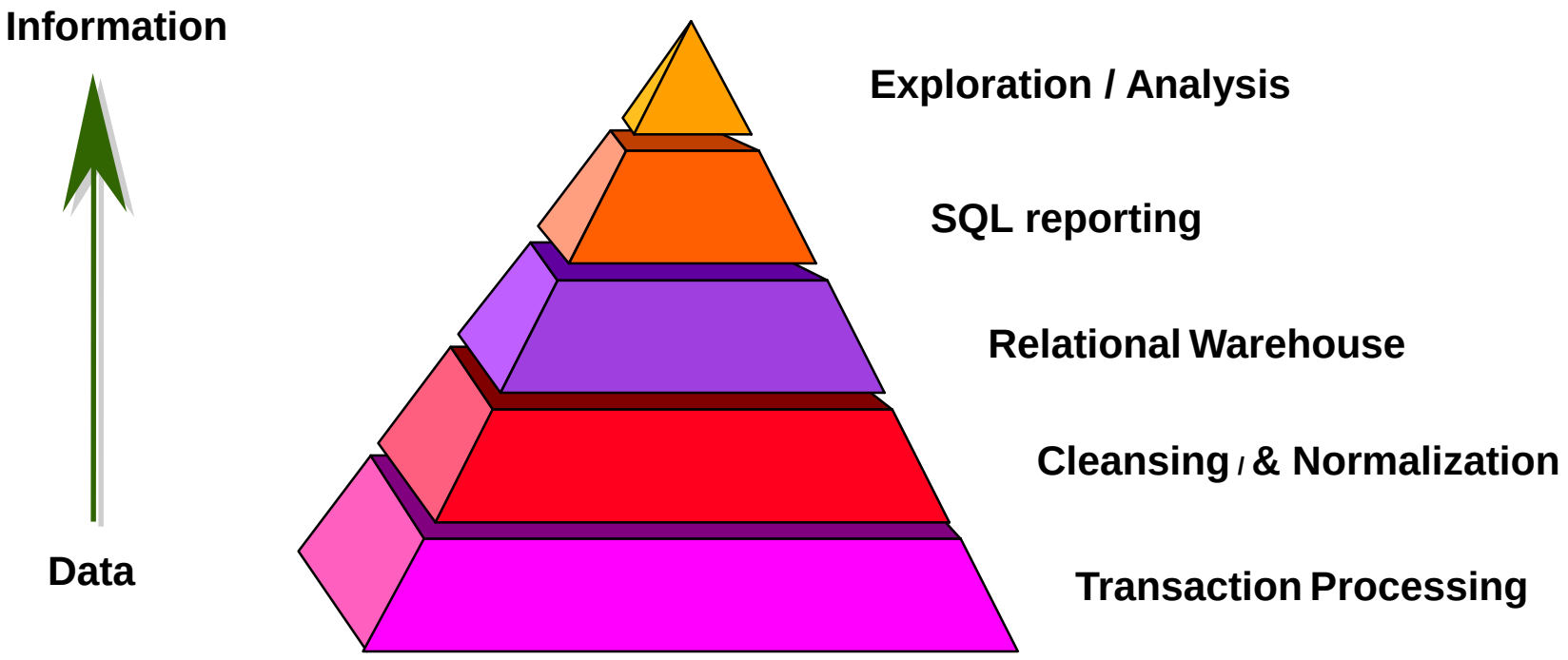
What Is a Data Warehouse?

Can often be set up as one VLDB (Very Large Database) or a collection of subject areas called Data Marts.

There are now tools which “unify” these Data Marts and make it appear as a single database.

What Is a Data Warehouse?

Transformation of Data to Information



Components of a Data Warehouse

Four General Components:

- Hardware
- DBMS - Database Management System
- Front End Access Tools
- Other Tools

In all components scalability is vital

Scalability is the ability to grow as your data and processing needs increase

Components of a Data Warehouse - Hardware

- Power - # of Processors, Memory, I/O Bandwidth, and Speed of the Bus
- Availability – Redundant equipment
- Disk Storage - Speed and enough storage for the loaded data set
- Backup Solution - Automated and be able to allow for incremental backups and archiving older data

Components of a Data Warehouse - DBMS

- Physical storage capacity of the DBMS
- Loading, indexing, and processing speed
- Availability
- Handle your data needs
- Operational integrity, reliability, and manageability

Components of a Data Warehouse - Front End & Other Tools

- Query Tools (SQL & GUI based)
- Report Writers
- Metadata Repositories
- OLAP (Online Analytical Processing)
- Data Mining Products

Components of a Data Warehouse – Metadata Repositories

Metadata is Data about Data. Users and Developers often need a way to find information on the data they use. Information can include:

- Source System(s) of the Data, contact information
- Related tables or subject areas
- Programs or Processes which use the data
- Population rules (Update or Insert and how often)
- Status of the Data Warehouse's processing and condition

Components of a Data Warehouse – OLAP Tools

OLAP - Online Analytical Processing. It works by aggregating detail data and looks at it by dimensions

- Gives the ability to "Drill Down" in to the detail data
- Decision Support Analysis Tool
- Multidimensional DB focusing on retrieval of precalculated data
- Ends the "big reports" with large amounts of detailed data
- These tools are often graphical and can run on a "thin client" such as a web browser

Components of a Data Warehouse – Data Mining

- Answers the questions you didn’t know to ask
- Analyzes great amounts of data (usually contained in a Data Warehouse) and looks for trends in the data
- Technology now allows us to do this better than in the past

Components of a Data Warehouse – Data Mining

- Most famous example is the Huggies - Heineken case
- Used in Retail sector to analyze buying habits
- Used in financial areas to detect fraud
- Used in the stock market to find trends
- Used in scientific research
- Used in national security

Getting the Data In

- Data will come from multiple databases and files within the organization
- Also can come from outside sources
  - Examples:
    - Weather Reports
    - Demographic information by Zip Code

Getting the Data In

Three Steps :

1. Extraction Phase
2. Transformation Phase
3. Loading Phase

Getting the Data In

Extraction Phase:

- Source systems export data via files or populates directly when the databases can “talk” to each other
- Transfers them to the Data Warehouse server and puts it into some sort of staging area

Getting the Data In

Transformation Phase:

- Takes data and turns it into a form that is suitable for insertion into the warehouse
- Combines related data
- Removes redundancies
- Common Codes (Commercial Customer)
- Spelling Mistakes (Lozenges)
- Consistency (PA,Pa,Penna,Pennsylvania)
- Formatting (Addresses)

Getting the Data In

Loading Phase:

- Places the cleaned data into the DBMS in its final, useable form
- Compare data from source systems and the Data Warehouse
- Document the load information for the users

Benefits

- Creates a single point for all data
- System is optimized and designed specifically for analysis
- Access data without impacting the operational systems
- Users can access the data directly without the direct help from IT dept

Costs

- Cost of implementation & maintenance (hardware, software, and staffing)
- Lack of compatibility between components
- Data from many sources are hard to combine, data integrity issues
- Bad designs and practices can lead to costly failures

Conclusion

- What is a Data Warehouse?
- Components of a Data Warehouse
- How the Data Gets In
- OLAP, Metadata, and Data Mining
- Benefits vs. Costs