Data Warehousing
Data Rich, but Information Poor

- Data is *stored*, not *explored*: by its volume and complexity it represents a burden, not a support
- *Data overload* results in uninformed decisions, contradictory information, higher overhead, wrong decisions, increased costs
- Data is not designed and is not structured for *successful management decision making*
Improving Decision Making

Decisions

Data Warehouse

Data
Data Warehouse Concepts
What’s a Data Warehouse?

A data warehouse is a single, integrated source of decision support information formed by collecting data from multiple sources, internal to the organization as well as external, and transforming and summarising this information to enable improved decision making.

A data warehouse is designed for easy access by users to large amounts of information, and data access is typically supported by specialized analytical tools and applications.
Data Warehouse Characteristics

- Key Characteristics of a Data Warehouse
  - Subject-oriented
  - Integrated
  - Time-variant
  - Non-volatile
Subject Oriented

• Example for an insurance company:

Applications Area

- Commercial and Life Insurance Systems
- Auto and Fire Policy Processing Systems
- Accounting System
- Billing System
- Claims Processing System

Data Warehouse

- Customer
- Policy
- Losses
- Premium
Integrated

- Data is stored once in a single integrated location (e.g. insurance company)

Customer data stored in several databases

Data Warehouse Database

Subject = Customer
Time - Variant

- Data is stored as a series of snapshots or views which record how it is collected across time.

Data Warehouse Data

- Data is tagged with some element of time - creation date, as of date, etc.
- Data is available on-line for long periods of time for trend analysis and forecasting. For example, five or more years
Non-Volatile

- Existing data in the warehouse is not overwritten or updated.
Transaction System vs. Data Warehouse
Transaction-Based Reporting System

Day-to-day operations

On-line, real time update into disparate systems

System Experts

Data Manipulation

Users

Unix

VMS

MVS

Other

Users
Warehouse-Based Reporting System

BENEFIT: Reduce data processing costs

Data Staging, Transformation and Cleansing

Data Warehouse

BENEFIT: Integrated, consistent data available for analysis

Executive Reporting and On-Line Analysis

BENEFIT: Improve Network Reporting processes and analytical capabilities

Unix
VMS
MVS
Other

Environment

Summarization

OLAP
Transaction - Warehouse Process

“Transaction Based Process”
- Day-to-day operations
- On-line, real time update.
- Detailed Information to operational systems.

“Warehouse Based Process”
- Decision support for management use.
- Summarize & Refine
- Transform

Batch Load
Transaction System vs. Data Warehouse

♦ **Transaction System**
  - Supports day-to-day operational processes
  - Contains raw, detailed data that has not been refined or cleansed
  - Volatile -- data changes from day-to-day, with frequent updates
  - Technical issues drive the data structure and system design
  - Disparate data structures, physical locations, query types, etc.
  - Users rely on technical analysts for reporting needs
  - Operational processes impacted by queries run off of system

♦ **Data Warehouse**
  - Supports management analysis and decision-making processes
  - Contains summarized, refined, and cleansed information
  - Non-volatile -- provides a data “snapshot”; adjustments are not permitted, or are limited
  - Business analysis requirements drive the data structure and system design
  - Integrated, consistent information on a single technology platform
  - Users have direct, fast access via On-line Analytical Processing tools
  - Minimal impact on operational processes
Data Warehouse Architecture
Data Warehouse Architecture

Operational System Characteristics

- Systems are widely dispersed
- Systems are organized for on-line transaction processing (OLTP)
- Functionality and data definitions are typically duplicated across many systems
Data Warehouse Architecture

Conversion and Cleansing Activities

- Map source data to target
- Data scrubbing
- Derive new data
- Data Extraction
- Transform / convert data
- Create / modify metadata
Data Warehouse Architecture

**ODS vs. DWH Staging Area**

- **ODS**
  - Contains Current and near current data
  - Contains almost all detail data
  - Data is updated frequently
  - Used to report a status continuously and ask specific questions – not flexible

- **DWH Staging Area**
  - Contains historical data
  - Contains summarized and detailed data
  - Data is non-volatile
  - Used to populate the DWH, which makes OLAP possible - flexible
Data Warehouse Architecture

Data Staging Area vs. Presentation Area

- Back room
- Sequential Processing: clean, combine, sort, archive, remove duplicates, add keys
- Off limits to the end users

- Front room
- ROLAP – OLAP: subject oriented, locally implemented, user group driven
- Available for end user inquiry
Data Warehouse Architecture

Data Warehouse Components

- Ranges from detailed to summarized data
- Contains metadata
- Many views of the data
- Subject-Oriented
- Time-variant
Data Warehouse Model
Requirements Gathering Process

**Business Measure Definition**

- Standard definition and related business rules and formulas
- Source data element(s), including quality constraints
- Data granularity levels (e.g., county detail for state)
- Data retention (e.g., one month, one quarter, one year, multiple years)
- Priority of the information (For example, is the information necessary to derive other business measures?)
- Data load frequency (e.g., monthly, quarterly, etc.)
Data Modeling Process

**Fact Table and Dimension**

- **Fact Table**
  - Each subject area (e.g. Business Unit) has its own Fact Table.
  - Fact tables relate or link dimensions.
  - Each attribute of the fact table is a measure or foreign key.
  - “The best facts are numeric, additive and continuously valued.”
  - Fact tables never contain direct links to other fact tables.

- **Dimension**
  - Best defined in focus sessions and interviews
  - Business Unit specific and overall perspectives
  - Typically, hierarchical in nature
Star Join Schema

Region_Dimension_Table

<table>
<thead>
<tr>
<th>region_id</th>
<th>region_doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE</td>
<td>Northeast</td>
</tr>
<tr>
<td>NW</td>
<td>Northwest</td>
</tr>
<tr>
<td>SE</td>
<td>Southeast</td>
</tr>
<tr>
<td>SW</td>
<td>Southwest</td>
</tr>
</tbody>
</table>

Product_Dimension_Table

<table>
<thead>
<tr>
<th>prod_grp_id</th>
<th>prod_id</th>
<th>prod_grp_desc</th>
<th>prod_desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
<td>Fewer devices</td>
<td>Power supply</td>
</tr>
<tr>
<td>20</td>
<td>140</td>
<td>Circuit boards</td>
<td>Motherboard</td>
</tr>
<tr>
<td>30</td>
<td>220</td>
<td>Components</td>
<td>Co-processor</td>
</tr>
</tbody>
</table>

Account_Dimension_Table

<table>
<thead>
<tr>
<th>account_id</th>
<th>account_doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>100000</td>
<td>ABC Electronics</td>
</tr>
<tr>
<td>110000</td>
<td>Midway Electric</td>
</tr>
<tr>
<td>120000</td>
<td>Victor Components</td>
</tr>
<tr>
<td>130000</td>
<td>Washburn, Inc.</td>
</tr>
<tr>
<td>140000</td>
<td>Zerox</td>
</tr>
</tbody>
</table>

Fact Table

<table>
<thead>
<tr>
<th>month</th>
<th>prod_id</th>
<th>region_id</th>
<th>account_id</th>
<th>vend_id</th>
<th>net-sales</th>
<th>gross_sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-1996</td>
<td>100</td>
<td>SW</td>
<td>100000</td>
<td>100</td>
<td>30,000</td>
<td>50,000</td>
</tr>
<tr>
<td>02-1996</td>
<td>140</td>
<td>NE</td>
<td>110000</td>
<td>200</td>
<td>23,000</td>
<td>42,000</td>
</tr>
<tr>
<td>03-1996</td>
<td>220</td>
<td>SW</td>
<td>100000</td>
<td>300</td>
<td>32,000</td>
<td>49,000</td>
</tr>
</tbody>
</table>

Time_Dimension_Table

<table>
<thead>
<tr>
<th>month</th>
<th>month_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-1996</td>
<td>January</td>
</tr>
<tr>
<td>02-1996</td>
<td>February</td>
</tr>
<tr>
<td>03-1996</td>
<td>March</td>
</tr>
</tbody>
</table>

Vendor_Dimension_Table

<table>
<thead>
<tr>
<th>vend_id</th>
<th>vendor_desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>PowerAge, Inc.</td>
</tr>
<tr>
<td>200</td>
<td>Advanced Micro Devices</td>
</tr>
<tr>
<td>300</td>
<td>Farad Incorporated</td>
</tr>
</tbody>
</table>
Storage of cubes

- Rolap (Relational On-line Analytical Processing)
  - Fact table is stored in relational data bases using keys
  - Building is faster, consulting is slower
  - Less storage space
  - Used for very large amounts of data

- Molap (Multi-dimensional On-line Analytical Processing)
  - Cube is stored using multidimensional tables
  - Data is stored in the cube, using sparsity
  - Longer building time, faster consulting time of the cube
  - More storage space needed
  - Used for smaller amounts of data

- Holap (Hybrid On-line Analytical Processing)
  - Cube is stored using a combination of both techniques
  - Detailed data is stored using ROLAP
  - Summarized data is stored using MOLAP
  - Synergy between storage and efficiency
Multi-Dimensional Analysis

Geography Dimension
- Zip Code
- County
- Region
- State

Customer Dimension
- Class of Trade
- Client Type
- Account
- Store

Product Dimension
- Product Family
- Product Line
- Brand
- Category
- Group
- Item

Business Measure: Net Sales

Net Sales by Brand by
Region by Client Type
Application of a Data Warehouse
Application Solution Classes

- Executive information system (EIS):
  - Present information at the highest level of summarization using corporate business measures. They are designed for extreme ease-of-use and, in many cases, only a mouse is required. Graphics are usually generously incorporated to provide at-a-glance indications of performance.

- Decision Support Systems (DSS):
  - They ideally present information in graphical and tabular form, providing the user with the ability to drill down on selected information. Note the increased detail and data manipulation options presented.
Data Mining

- Data Mining provides techniques to:
  - Detect trends or patterns, find correlations
  - Exploratory data analysis
  - Forecasting and business modeling
- Intelligent agents are coupled to the data warehouse using different techniques:
  - Neural networks
  - Expert systems
  - Advanced statistics
- The volume and complexity of information may not become a barrier
- Applications: Early warning systems, Fraud detection, market research, direct mail.