Chapter 13

Business Intelligence and Data Warehouses
Objectives

In this chapter, you will learn:

• How business intelligence provides a comprehensive business decision support framework

• About business intelligence architecture, its evolution, and reporting styles

• About the relationship and differences between operational data and decision support data

• What a data warehouse is and how to prepare data for one
Objectives (cont’d.)

• What star schemas are and how they are constructed
• About data analytics, data mining, and predictive analytics
• About online analytical processing (OLAP)
• How SQL extensions are used to support OLAP-type data manipulations
The Need for Data Analysis

- Managers track daily transactions to evaluate how the business is performing
- Strategies should be developed to meet organizational goals using operational databases
- Data analysis provides information about short-term tactical evaluations and strategies
Business Intelligence

• Comprehensive, cohesive, integrated tools and processes
  – Capture, collect, integrate, store, and analyze data
  – Generate information to support business decision making

• Framework that allows a business to transform:
  – Data into information
  – Information into knowledge
  – Knowledge into wisdom
Business Intelligence Architecture

• Composed of data, people, processes, technology, and management of components
• Focuses on strategic and tactical use of information
• Key performance indicators (KPI)
  – Measurements that assess company’s effectiveness or success in reaching goals
• Multiple tools from different vendors can be integrated into a single BI framework
Business Intelligence Benefits

• Main goal: improved decision making
• Other benefits
  – Integrating architecture
  – Common user interface for data reporting and analysis
  – Common data repository fosters single version of company data
  – Improved organizational performance
Business Intelligence Evolution

FIGURE 13.2 Evolution of BI information dissemination formats

1979 1989 1999 2009
1970s 1980s 1990s 2000s 2010

Centralized Reporting

Spreadsheets

Enterprise Reporting

Dashboards

OLAP

Mobile BI

Credit: Oleksiy Mark / Shutterstock.com
SOURCE: Course Technology/Cengage Learning
<table>
<thead>
<tr>
<th>SYSTEM TYPE</th>
<th>DATA SOURCE</th>
<th>DATA EXTRACTION/INTEGRATION PROCESS</th>
<th>DATA STORE</th>
<th>END-USER QUERY TOOL</th>
<th>END USER PRESENTATION TOOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional mainframe-based online transaction processing (OLTP)</td>
<td>Operational data</td>
<td>None</td>
<td>None</td>
<td>Very basic</td>
<td>Very basic</td>
</tr>
<tr>
<td>Managerial information system (MIS)</td>
<td>Operational data</td>
<td>Basic extraction and aggregation</td>
<td>Lightly aggregated data in RDBMS</td>
<td>Same as above, in addition to some ad hoc reporting using SQL</td>
<td>Same as above, in addition to some ad hoc columnar report definitions</td>
</tr>
<tr>
<td>First-generation departmental decision support system (DSS)</td>
<td>Operational data</td>
<td>Data extraction and integration process populates DSS data store</td>
<td>First DSS database generation Usually RDBMS</td>
<td>Query tool with some analytical capabilities and predefined reports</td>
<td>Spreadsheet style Advanced presentation tools with plotting and graphics capabilities</td>
</tr>
<tr>
<td>First-generation BI</td>
<td>Operational data</td>
<td>Advanced data extraction and integration</td>
<td>Data warehouse RDBMS technology Optimized for query purposes Star schema model</td>
<td>Same as above</td>
<td>Same as above, in addition to multidimensional presentation tools with drill-down capabilities</td>
</tr>
<tr>
<td>Second-generation BI Online analytical processing (OLAP)</td>
<td>Same as above</td>
<td>Same as above</td>
<td>Data warehouse stores data in MDBMS Cubes with multiple dimensions</td>
<td>Adds support for end-user-based data analytics</td>
<td>Same as above, but uses cubes and multidimensional matrices; limited by terms of cube size Dashboards Scorecards Portals</td>
</tr>
<tr>
<td>Third-generation Mobile BI and cloud-based</td>
<td>Same as above</td>
<td>Same as above Cloud-based</td>
<td>Same as above Cloud based</td>
<td>Advanced analytics Limited ad hoc interactions</td>
<td>Mobile devices: iPhone, iPad, Blackberry, Android</td>
</tr>
</tbody>
</table>
Business Intelligence Technology Trends

• Data storage improvements
• Business intelligence appliances
• Business intelligence as a service
• Big Data analytics
• Personal analytics
Decision Support Data

- BI effectiveness depends on quality of data gathered at operational level
- Operational data seldom well-suited for decision support tasks
- Need reformat data in order to be useful for business intelligence
Operational Data vs. Decision Support Data

• Operational data
  – Mostly stored in relational database
  – Optimized to support transactions representing daily operations

• Decision support data differs from operational data in three main areas:
  – Time span
  – Granularity
  – Dimensionality
Operational data have a narrow time span, low granularity, and single focus. Such data are usually presented in tabular format, in which each row represents a single transaction. This format often makes it difficult to derive useful information.

Decision support system (DSS) data focus on a broader timespan, tend to have high levels of granularity, and can be examined in multiple dimensions. For example, note these possible aggregations:

- Sales by product, region, agent, and so on
- Sales for all years or only a few selected years
- Sales for all products or only a few selected products

**Operational Data**

<table>
<thead>
<tr>
<th>Year</th>
<th>Region</th>
<th>Agent</th>
<th>Product</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>East</td>
<td>Carlos</td>
<td>Erasers</td>
<td>50</td>
</tr>
<tr>
<td>2010</td>
<td>East</td>
<td>Tere</td>
<td>Erasers</td>
<td>12</td>
</tr>
<tr>
<td>2010</td>
<td>North</td>
<td>Carlos</td>
<td>Widgets</td>
<td>120</td>
</tr>
<tr>
<td>2010</td>
<td>North</td>
<td>Tere</td>
<td>Widgets</td>
<td>100</td>
</tr>
<tr>
<td>2010</td>
<td>North</td>
<td>Carlos</td>
<td>Widgets</td>
<td>30</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Victor</td>
<td>Balls</td>
<td>145</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Victor</td>
<td>Balls</td>
<td>34</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Victor</td>
<td>Balls</td>
<td>80</td>
</tr>
<tr>
<td>2010</td>
<td>West</td>
<td>Mary</td>
<td>Pencils</td>
<td>89</td>
</tr>
<tr>
<td>2010</td>
<td>West</td>
<td>Mary</td>
<td>Pencils</td>
<td>56</td>
</tr>
<tr>
<td>2010</td>
<td>East</td>
<td>Carlos</td>
<td>Pencils</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>East</td>
<td>Carlos</td>
<td>Pencils</td>
<td>55</td>
</tr>
<tr>
<td>2010</td>
<td>North</td>
<td>Mary</td>
<td>Pencils</td>
<td>60</td>
</tr>
<tr>
<td>2010</td>
<td>North</td>
<td>Victor</td>
<td>Erasers</td>
<td>20</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Carlos</td>
<td>Widgets</td>
<td>30</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Mary</td>
<td>Widgets</td>
<td>75</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Mary</td>
<td>Widgets</td>
<td>50</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Tere</td>
<td>Balls</td>
<td>70</td>
</tr>
<tr>
<td>2010</td>
<td>South</td>
<td>Tere</td>
<td>Erasers</td>
<td>90</td>
</tr>
<tr>
<td>2011</td>
<td>West</td>
<td>Carlos</td>
<td>Widgets</td>
<td>25</td>
</tr>
<tr>
<td>2011</td>
<td>West</td>
<td>Tere</td>
<td>Balls</td>
<td>100</td>
</tr>
</tbody>
</table>

**Decision Support Data**

<table>
<thead>
<tr>
<th>Year</th>
<th>Region</th>
<th>North</th>
<th>South</th>
<th>West</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>East</td>
<td>55</td>
<td>70</td>
<td>100</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>Erasers</td>
<td>20</td>
<td>90</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pencils</td>
<td>45</td>
<td>60</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Widgets</td>
<td>155</td>
<td>25</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>100</td>
<td>80</td>
<td>315</td>
<td>620</td>
</tr>
</tbody>
</table>

Interest and Attention
Operational Data vs. Decision Support Data

• Operational data
  – Mostly stored in relational database
  – Optimized to support transactions representing daily operations

• Decision support data differs from operational data in three main areas:
  – Time span
  – Granularity
  – Dimensionality
### Table 13.5
Contrasting Operational and Decision Support Data Characteristics

<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>OPERATIONAL DATA</th>
<th>DECISION SUPPORT DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data currency</td>
<td>Current operations</td>
<td>Historic data</td>
</tr>
<tr>
<td></td>
<td>Real-time data</td>
<td>Snapshot of company data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time component (week/month/year)</td>
</tr>
<tr>
<td>Granularity</td>
<td>Atomic-detailed data</td>
<td>Summarized data</td>
</tr>
<tr>
<td>Summarization level</td>
<td>Low; some aggregate yields</td>
<td>High; many aggregation levels</td>
</tr>
<tr>
<td>Data model</td>
<td>Highly normalized</td>
<td>Non-normalized</td>
</tr>
<tr>
<td></td>
<td>Mostly relational DBMSs</td>
<td>Complex structures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Some relational, but mostly multidimensional DBMSs</td>
</tr>
<tr>
<td>Transaction type</td>
<td>Mostly updates</td>
<td>Mostly query</td>
</tr>
<tr>
<td>Transaction volumes</td>
<td>High update volumes</td>
<td>Periodic loads and summary calculations</td>
</tr>
<tr>
<td>Transaction speed</td>
<td>Updates are critical</td>
<td>Retrievals are critical</td>
</tr>
<tr>
<td>Query activity</td>
<td>Low to medium</td>
<td>High</td>
</tr>
<tr>
<td>Query scope</td>
<td>Narrow range</td>
<td>Broad range</td>
</tr>
<tr>
<td>Query complexity</td>
<td>Simple to medium</td>
<td>Very complex</td>
</tr>
<tr>
<td>Data volumes</td>
<td>Hundreds of gigabytes</td>
<td>Terabytes to petabytes</td>
</tr>
</tbody>
</table>

**Data Sources**

- Single centralized source
- Many: includes files, RDBs, non-RDBs
Decision Support

Database Requirements

• Specialized DBMS tailored to provide fast answers to complex queries
• Three main requirements
  – Database schema
  – Data extraction and loading
  – Database size
Decision Support
Database Requirements (cont’d.)

• Database schema
  – Complex data representations
  – Aggregated and summarized data
  – Queries extract multidimensional time slices

• Data extraction and filtering
  – Supports different data sources
    • Flat files
    • Hierarchical, network, and relational databases
    • Multiple vendors
  – Checking for inconsistent data
Decision Support
Database Requirements (cont’d.)

• Database size
  – In 2005, Wal-Mart had 260 terabytes of data in its data warehouses
  – DBMS must support very large databases (VLDBs)
The Data Warehouse

- Integrated, subject-oriented, time-variant, and nonvolatile collection of data
  - Provides support for decision making
- Usually a read-only database optimized for data analysis and query processing
- Requires time, money, and considerable managerial effort to create
FIGURE 13.4  The ETL process

Operational data

Transformation
- Filter
- Transform
- Integrate
- Classify
- Aggregate
- Summarize

Data warehouse
- Integrated
- Subject-oriented
- Time-variant
- Nonvolatile

SOURCE: Course Technology/Cengage Learning
Data Marts

• Small, single-subject data warehouse subset
• More manageable data set than data warehouse
• Provides decision support to small group of people
• Typically lower cost and lower implementation time than data warehouse
### Twelve Rules for a Data Warehouse

<table>
<thead>
<tr>
<th>RULE NO.</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The data warehouse and operational environments are separated.</td>
</tr>
<tr>
<td>2</td>
<td>The data warehouse data are integrated.</td>
</tr>
<tr>
<td>3</td>
<td>The data warehouse contains historical data over a long time.</td>
</tr>
<tr>
<td>4</td>
<td>The data warehouse data are snapshot data captured at a given point in time.</td>
</tr>
<tr>
<td>5</td>
<td>The data warehouse data are subject oriented.</td>
</tr>
<tr>
<td>6</td>
<td>The data warehouse data are mainly read-only with periodic batch updates from operational data. No online updates are allowed.</td>
</tr>
<tr>
<td>7</td>
<td>The data warehouse development life cycle differs from classical systems development. Data warehouse development is data-driven; the classical approach is process-driven.</td>
</tr>
<tr>
<td>8</td>
<td>The data warehouse contains data with several levels of detail: current detail data, old detail data, lightly summarized data, and highly summarized data.</td>
</tr>
<tr>
<td>9</td>
<td>The data warehouse environment is characterized by read-only transactions to very large data sets. The operational environment is characterized by numerous update transactions to a few data entities at a time.</td>
</tr>
<tr>
<td>10</td>
<td>The data warehouse environment has a system that traces data sources, transformations, and storage.</td>
</tr>
<tr>
<td>11</td>
<td>The data warehouse’s metadata are a critical component of this environment. The metadata identify and define all data elements. The metadata provide the source, transformation, integration, storage, usage, relationships, and history of each data element.</td>
</tr>
<tr>
<td>12</td>
<td>The data warehouse contains a chargeback mechanism for resource usage that enforces optimal use of the data by end users.</td>
</tr>
</tbody>
</table>
Star Schemas

• Data-modeling technique
  – Maps multidimensional decision support data into relational database

• Creates near equivalent of multidimensional database schema from relational data

• Easily implemented model for multidimensional data analysis while preserving relational structures

• Four components: facts, dimensions, attributes, and attribute hierarchies
FIGURE 13.10 Star schema for SALES

- **LOCATION**
  - LOC_ID
  - LOC_DESCRIPTION
  - REGION_ID
  - LOC_STATE
  - LOC_CITY
  - 25 records

- **CUSTOMER**
  - CUST_ID
  - CUST_LNAME
  - CUST_FNAME
  - CUST_INITIAL
  - CUSTDOB
  - 125 records

- **PRODUCT**
  - PROD_ID
  - PROD_DESCRIPTION
  - PROD_TYPE_ID
  - PROD_BRAND
  - PROD_COLOR
  - PROD_SIZE
  - PROD_PACKAGE
  - PROD_PRICE
  - 3,000 records

- **SALES**
  - TIME_ID
  - LOC_ID
  - CUST_ID
  - PROD_ID
  - SALES_QUANTITY
  - SALES_PRICE
  - SALES_TOTAL
  - 3,000,000 records

- **TIME**
  - TIME_ID
  - TIME_YEAR
  - TIME_QUARTER
  - TIME_MONTH
  - TIME_DAY
  - TIME_CLOCKTIME
  - 365 records

Daily sales aggregates by store, customer, and product

SOURCE: Course Technology/Cengage Learning
Facts

- Numeric measurements that represent specific business aspect or activity
  - Normally stored in fact table that is center of star schema
- Fact table contains facts linked through their dimensions
- Metrics are facts computed at run time
Dimensions

- Qualifying characteristics provide additional perspectives to a given fact
- Decision support data almost always viewed in relation to other data
- Study facts via dimensions
- Dimensions stored in dimension tables
Attributes

• Use to search, filter, and classify facts
• Dimensions provide descriptions of facts through their attributes
• No mathematical limit to the number of dimensions
• Slice and dice: focus on slices of the data cube for more detailed analysis
Attribute Hierarchies

- Provide top-down data organization
- Two purposes:
  - Aggregation
  - Drill-down/roll-up data analysis
- Determine how the data are extracted and represented
- Stored in the DBMS’s data dictionary
- Used by OLAP tool to access warehouse properly
The attribute hierarchy allows the end user to perform drill-down and roll-up searches.
Attribute hierarchies in multidimensional analysis

- Time dimension
  - Year
  - Quarter
  - Month
  - Week

- Product dimension
  - All products
  - By product type
  - One product

- Location hierarchy
  - Total of quarters
  - Region
  - State
  - City
  - Store

SOURCE: Course Technology/Cengage Learning
Star Schema Representation

• Facts and dimensions represented in physical tables in data warehouse database

• Many fact rows related to each dimension row
  – Primary key of fact table is a composite primary key
  – Fact table primary key formed by combining foreign keys pointing to dimension tables

• Dimension tables are smaller than fact tables

• Each dimension record is related to thousands of fact records
Performance-Improving Techniques for the Star Schema

• Four techniques to optimize data warehouse design:
  – Normalizing dimensional tables
  – Maintaining multiple fact tables to represent different aggregation levels
  – Denormalizing fact tables
  – Partitioning and replicating tables
Performance-Improving Techniques for the Star Schema (cont’d.)

• Dimension tables normalized to:
  – Achieve semantic simplicity
  – Facilitate end-user navigation through the dimensions

• Denormalizing fact tables improves data access performance and saves data storage space

• Partitioning splits table into subsets of rows or columns

• Replication makes copy of table and places it in different location
Normalized dimension tables

**REGION**
- REGION_ID
- REGION_NAME

**STATE**
- STATE_ID
- STATE_NAME
- REGION_ID

**LOCATION**
- LOC_ID
- LOC_DESCRIPTION
- CITY_ID

**CITY**
- CITY_ID
- CITY_NAME
- STATE_ID

**SALES**
- TIME_ID
- LOC_ID
- CUST_ID
- PROD_ID
- SALES_QUANTITY
- SALES_PRICE
- SALES_TOTAL

SOURCE: Course Technology/Cengage Learning
Data Analytics

• Subset of BI functionality
• Encompasses a wide range of mathematical, statistical, and modeling techniques
  – Purpose of extracting knowledge from data
• Tools can be grouped into two separate areas:
  – Explanatory analytics
  – Predictive analytics
Data Mining

• Data-mining tools do the following:
  – Analyze data
  – Uncover problems or opportunities hidden in data relationships
  – Form computer models based on their findings
  – Use models to predict business behavior

• Runs in two modes
  – Guided
  – Automated
Figure 13.15: Data-mining phases

- **Data preparation phase**
  - Identify data set
  - Clean data set
  - Integrate data set

- **Data analysis and classification phase**
  - Classification analysis
  - Clustering and sequence analysis
  - Link analysis
  - Trend and deviation analysis

- **Knowledge acquisition phase**
  - Select and apply algorithms
  - Neural networks
  - Inductive logic
  - Decision trees
  - Clustering
  - Regression tree
  - Nearest neighbor
  - Visualization, etc.

- **Prognosis phase**
  - Modeling
  - Forecasting
  - Prediction

*Source: Course Technology/Cengage Learning*
Online Analytical Processing

• Three main characteristics:
  – Multidimensional data analysis techniques
  – Advanced database support
  – Easy-to-use end-user interfaces
Multidimensional Data Analysis Techniques

• Data are processed and viewed as part of a multidimensional structure

• Augmented by the following functions:
  – Advanced data presentation functions
  – Advanced data aggregation, consolidation, and classification functions
  – Advanced computational functions
  – Advanced data modeling functions
Advanced Database Support

• Advanced data access features include:
  – Access to many different kinds of DBMSs, flat files, and internal and external data sources
  – Access to aggregated data warehouse data
  – Advanced data navigation
  – Rapid and consistent query response times
  – Maps end-user requests to appropriate data source and to proper data access language
  – Support for very large databases
Easy-to-Use End-User Interface

• Advanced OLAP features are more useful when access is simple
• Many interface features are “borrowed” from previous generations of data analysis tools
  – Already familiar to end users
  – Makes OLAP easily accepted and readily used
FIGURE 13.19  OLAP server with local miniature data marts

OLAP “server”

Analytical processing logic

Data processing logic

Multiple OLAP clients accessing the OLAP server

Sales Dept.
OLAP GUI

Marketing Dept.
OLAP GUI

Manufacturing Dept.
OLAP GUI

Procurement Dept.
OLAP GUI

Local data marts

Customers

Marketing

Production

Vendors

Data extracted from the data warehouse to local data marts, which provides faster processing

SOURCE: Course Technology/Cengage Learning
Relational OLAP

• Relational online analytical processing (ROLAP) provides the following extensions:
  – Multidimensional data schema support within the RDBMS
  – Data access language and query performance optimized for multidimensional data
  – Support for very large databases (VLDBs)
Multidimensional OLAP

- Multidimensional online analytical processing (MOLAP) extends OLAP functionality to multidimensional database management systems (MDBMSs)
  - MDBMS end users visualize stored data as a 3D data cube
  - Data cubes can grow to $n$ dimensions, becoming hypercubes
  - To speed access, data cubes are held in memory in a cube cache
Relational vs. Multidimensional OLAP

• Selection of one or the other depends on evaluator’s vantage point
• Proper evaluation must include supported hardware, compatibility with DBMS, etc.
• ROLAP and MOLAP vendors working toward integration within unified framework
• Relational databases use star schema design to handle multidimensional data
<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>ROLAP</th>
<th>MOLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schema</td>
<td>Uses star schema</td>
<td>Uses data cubes</td>
</tr>
<tr>
<td></td>
<td>Additional dimensions can be added dynamically</td>
<td>Multidimensional arrays, row stores, column stores</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additional dimensions require re-creation of the data cube</td>
</tr>
<tr>
<td>Database size</td>
<td>Medium to large</td>
<td>Large</td>
</tr>
<tr>
<td>Architecture</td>
<td>Client/server standards-based</td>
<td>Client/server</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Open or proprietary, depending on vendor</td>
</tr>
<tr>
<td>Access</td>
<td>Supports ad hoc requests</td>
<td>Limited to predefined dimensions</td>
</tr>
<tr>
<td></td>
<td>Unlimited dimensions</td>
<td>Proprietary access languages</td>
</tr>
<tr>
<td>Speed</td>
<td>Good with small data sets; average for medium-sized to large data sets</td>
<td>Faster for large data sets with predefined dimensions</td>
</tr>
</tbody>
</table>
SQL Extensions for OLAP

• Proliferation of OLAP tools fostered development of SQL extensions
• Many innovations have become part of standard SQL
• All SQL commands will work in data warehouse as expected
• Most queries include many data groupings and aggregations over multiple columns
The ROLLUP Extension

• Used with GROUP BY clause to generate aggregates by different dimensions
• GROUP BY generates only one aggregate for each new value combination of attributes
• ROLLUP extension enables subtotal for each column listed except for the last one
  – Last column gets grand total
• Order of column list important
### ROLLUP extension

**Figure 13.21**

The SQL query to calculate the total sales for different product codes is as follows:

```
SELECT V_CODE, P_CODE, SUM(SALE_UNITS*SALE_PRICE) AS TOTSALES
FROM DW_DAYSALESFACT NATURAL JOIN DW_PRODUCT NATURAL JOIN DW_VENDOR
GROUP BY ROLLUP (V_CODE, P_CODE)
ORDER BY V_CODE, P_CODE;
```

#### Result Table

<table>
<thead>
<tr>
<th>V_CODE</th>
<th>P_CODE</th>
<th>TOTSALES</th>
</tr>
</thead>
<tbody>
<tr>
<td>21225</td>
<td>23109-HB</td>
<td>99.5</td>
</tr>
<tr>
<td>21225</td>
<td>PVC23DRT</td>
<td>199.58</td>
</tr>
<tr>
<td>21225</td>
<td>SM-18277</td>
<td>41.94</td>
</tr>
<tr>
<td>21225</td>
<td></td>
<td>341.02</td>
</tr>
<tr>
<td>21344</td>
<td>13-Q2/P2</td>
<td>239.84</td>
</tr>
<tr>
<td>21344</td>
<td>54778-2T</td>
<td>59.88</td>
</tr>
<tr>
<td>21344</td>
<td></td>
<td>299.72</td>
</tr>
<tr>
<td>23119</td>
<td>1546-002</td>
<td>79.9</td>
</tr>
<tr>
<td>23119</td>
<td></td>
<td>79.9</td>
</tr>
<tr>
<td>24288</td>
<td>2232/QTY</td>
<td>219.84</td>
</tr>
<tr>
<td>24288</td>
<td>89-WRE-Q</td>
<td>513.98</td>
</tr>
<tr>
<td>24288</td>
<td></td>
<td>733.82</td>
</tr>
<tr>
<td>25595</td>
<td>2238/QPD</td>
<td>77.9</td>
</tr>
<tr>
<td>25595</td>
<td>WR3/TT3</td>
<td>719.7</td>
</tr>
<tr>
<td>25595</td>
<td></td>
<td>797.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2252.06</td>
</tr>
</tbody>
</table>

16 rows selected.
The CUBE Extension

- CUBE extension used with GROUP BY clause to generate aggregates by listed columns
  - Includes the last column
- Enables subtotal for each column in addition to grand total for last column
  - Useful when you want to compute all possible subtotals within groupings
- Cross-tabulations are good candidates for application of CUBE extension
### SQL Query

```sql
SQL> SELECT TM_MONTH, P_CODE, SUM(SALE_UNITS*SALE_PRICE) AS TOTSALES
2  FROM DWWAYSALEFACT NATURAL JOIN DVPRODUCT NATURAL JOIN DWTIME
3  GROUP BY CUBE (TM_MONTH, P_CODE)
4  ORDER BY TM_MONTH, P_CODE;
```

### Table

<table>
<thead>
<tr>
<th>TM_MONTH</th>
<th>P_CODE</th>
<th>TOTSALES</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>13-Q2/P2</td>
<td>134.91</td>
</tr>
<tr>
<td>9</td>
<td>1546-QQ2</td>
<td>79.9</td>
</tr>
<tr>
<td>9</td>
<td>2232/QTY</td>
<td>109.92</td>
</tr>
<tr>
<td>9</td>
<td>2238/QPD</td>
<td>77.9</td>
</tr>
<tr>
<td>9</td>
<td>23109-HB</td>
<td>59.7</td>
</tr>
<tr>
<td>9</td>
<td>54778-2T</td>
<td>39.92</td>
</tr>
<tr>
<td>9</td>
<td>89-WRE-Q</td>
<td>256.99</td>
</tr>
<tr>
<td>9</td>
<td>PUC23DRT</td>
<td>99.79</td>
</tr>
<tr>
<td>9</td>
<td>SM-18277</td>
<td>20.97</td>
</tr>
<tr>
<td>9</td>
<td>WR3/TT3</td>
<td>359.85</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>1239.85</td>
</tr>
<tr>
<td>10</td>
<td>13-Q2/P2</td>
<td>104.93</td>
</tr>
<tr>
<td>10</td>
<td>2232/QTY</td>
<td>109.92</td>
</tr>
<tr>
<td>10</td>
<td>23109-HB</td>
<td>39.8</td>
</tr>
<tr>
<td>10</td>
<td>54778-2T</td>
<td>19.96</td>
</tr>
<tr>
<td>10</td>
<td>89-WRE-Q</td>
<td>256.99</td>
</tr>
<tr>
<td>10</td>
<td>PUC23DRT</td>
<td>99.79</td>
</tr>
<tr>
<td>10</td>
<td>SM-18277</td>
<td>20.97</td>
</tr>
<tr>
<td>10</td>
<td>WR3/TT3</td>
<td>359.85</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>1012.21</td>
</tr>
</tbody>
</table>

#### Subtotals
- **Subtotals by Month**
  - Month 9: 1239.85
  - Month 10: 1012.21

#### Subtotals by Product
- Product 13-Q2/P2: 239.84
- Product 1546-QQ2: 79.9
- Product 2232/QTY: 219.84
- Product 2238/QPD: 77.9
- Product 23109-HB: 99.5
- Product 54778-2T: 59.88
- Product 89-WRE-Q: 513.08
- Product PUC23DRT: 199.58
- Product SM-18277: 41.94
- Product WR3/TT3: 719.7

#### Grand Total
- Total for all products and months: 2252.06

31 rows selected.
Materialized Views

• A dynamic table that contains SQL query command to generate rows
  – Also contains the actual rows
• Created the first time query is run and summary rows are stored in table
• Automatically updated when base tables are updated
Summary

• Business intelligence generates information used to support decision making
• BI covers a range of technologies, applications, and functionalities
• Decision support systems were the precursor of current generation BI systems
• Operational data not suited for decision support
Summary (cont’d.)

• Data warehouse provides support for decision making
  – Usually read-only
  – Optimized for data analysis, query processing
• Star schema is a data-modeling technique
  – Maps multidimensional decision support data into a relational database
• Star schema has four components:
  – Facts, dimensions, attributes, and attribute hierarchies
Summary (cont’d.)

• Data analytics
  – Provides advanced data analysis tools to extract knowledge from business data

• Data mining
  – Automates the analysis of operational data to find previously unknown data characteristics, relationships, dependencies, and trends

• Predictive analytics
  – Uses information generated in the data-mining phase to create advanced predictive models
Summary (cont’d.)

• Online analytical processing (OLAP)
  – Advanced data analysis environment that supports decision making, business modeling, and operations research

• SQL has been enhanced with extensions that support OLAP-type processing and data generation
Database Systems: Design, Implementation, and Management

Chapter 15 (Section 15.3)
XML
Extensible Markup Language (XML)

• Companies use Internet to create new systems that integrate their data
  – Increase efficiency and reduce costs
• Electronic commerce enables organizations to market to millions of users
• Most e-commerce transactions take place between businesses
• HTML Web pages display in the browser
  – Tags describe how something looks on the page
Extensible Markup Language (XML) (cont’d.)

• Extensible Markup Language (XML)
  – Meta-language to represent and manipulate data elements
  – Self-descriptive (semi-structured)
  – Hierarchical (vs. flat tables)
  – Facilitates exchange of structured documents over the Web
  – Allows definition of new tags (hence the name: Extensible)
    • Case sensitive
    • Must be well-formed and properly nested
    • Comments indicated with <- and ->
    • XML and xml prefixes reserved for XML tags only
<table>
<thead>
<tr>
<th>P_CODE</th>
<th>P_DESCRIPT</th>
<th>P_INDATE</th>
<th>P_QOH</th>
<th>P_MIN</th>
<th>P_PRICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>23109-HB</td>
<td>Claw hammer</td>
<td>08/19/2009</td>
<td>23</td>
<td>10</td>
<td>5.95</td>
</tr>
<tr>
<td>23114-A4</td>
<td>Sledge Hammer, 12 lb.</td>
<td>09/01/2009</td>
<td>8</td>
<td>5</td>
<td>14.40</td>
</tr>
</tbody>
</table>
E.g. ProductList Table & Corresponding XML Document

<table>
<thead>
<tr>
<th>P_CODE</th>
<th>P_DESCRIP</th>
<th>P_INDATE</th>
<th>P_QOH</th>
<th>P_MIN</th>
<th>P_PRICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>23109-HB</td>
<td>Claw hammer</td>
<td>08/19/2009</td>
<td>23</td>
<td>10</td>
<td>5.95</td>
</tr>
<tr>
<td>23114-A4</td>
<td>Sledge Hammer, 12 lb.</td>
<td>09/01/2009</td>
<td>8</td>
<td>5</td>
<td>14.40</td>
</tr>
</tbody>
</table>

```xml
<?xml version ="1.0"?>
<ProductList>
    <Product>
        <P_CODE>23109-HB</P_CODE>
        <P_DESCRIP>Claw hammer</P_DESCRIP>
        <P_INDATE>08/19/2009</P_INDATE>
        <P_QOH>23</P_QOH>
        <P_MIN>10</P_MIN>
        <P_PRICE>5.95</P_PRICE>
    </Product>
    <Product>
        <P_CODE>23114-A4</P_CODE>
        <P_DESCRIP>Sledge Hammer, 12 lb.</P_DESCRIP>
        <P_INDATE>09/01/2009</P_INDATE>
        <P_QOH>8</P_QOH>
        <P_MIN>5</P_MIN>
        <P_PRICE>14.40</P_PRICE>
    </Product>
</ProductList>
```
Document Type Definitions (DTD) and XML Schemas

• Document Type Definition (DTD)
  – File with .dtd extension that describes elements
  – Provides composition of database’s logical model
  – Defines the syntax rules or valid tags for each type of XML document

• Companies engaging in e-commerce transaction must develop and share DTDs

• DTD referenced from inside XML document
  – XML document said to adhere to DTD
Nothing: exactly 1
? : means 0 or 1
* : means 0 or more
+ : means 1 or more

“+” sign indicates one or more ORD_PRODS elements

Two ORD_PRODS elements in XML document
Document Type Definitions (DTD) and XML Schemas (cont’d.)

• XML schema
  – Advanced data definition language
  – Describes the structure of XML data documents

• Advantage of XML schema:
  – More closely maps to database terminology and features

• XML schema definition (XSD) file uses syntax similar to XML document
<?xml version="1.0" ?>
<ProductList>
  <Product>
    <P_CODE>23109-HB</P_CODE>
    <P_DESCRIP>Claw hammer</P_DESCRIP>
    <P_DATE>08/19/2009</P_DATE>
    <P_QOH>23</P_QOH>
    <P_MIN>10</P_MIN>
    <P_PRICE>5.95</P_PRICE>
  </Product>
  <Product>
    <P_CODE>23114-AA</P_CODE>
    <P_DESCRIP>Sledge Hammer, 12 lb.</P_DESCRIP>
    <P_DATE>09/01/2009</P_DATE>
    <P_QOH>8</P_QOH>
    <P_MIN>5</P_MIN>
    <P_PRICE>14.40</P_PRICE>
  </Product>
</ProductList>
### XML data binding

#### HTML Code

```html
<html>
<head>
<title>BINDING THE PRODUCTLIST XML DATA TO HTML TABLE</title>
</head>
<body>
<xml id="PRODLIST" src="PRODUCTLIST.XML"></xml>
<table border="1" dataSRC="#PRODLIST">
  <tr>
    <td><span datafld="P_CODE"></span></td>
    <td><span datafld="P_DESCRIP"></span></td>
    <td><span datafld="P_QOH"></span></td>
    <td><span datafld="P_PRICE"></span></td>
  </tr>
</table>
</body>
</html>
```

#### HTML Output

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>QOH</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>23109-HB</td>
<td>Claw hammer</td>
<td>23</td>
<td>5.95</td>
</tr>
<tr>
<td>23114-AA</td>
<td>Sledge Hammer, 12 lb</td>
<td>8</td>
<td>14.40</td>
</tr>
</tbody>
</table>