Chapter 2: Data Warehousing

Jason C. H. Chen, Ph.D.
Professor of MIS
School of Business Administration
Gonzaga University
Spokane, WA 99258
chen@jepson.gonzaga.edu

Learning Objectives

• Understand the basic definitions and concepts of data warehouses
• Learn different types of data warehousing architectures; their comparative advantages and disadvantages
• Describe the processes used in developing and managing data warehouses
• Explain data warehousing operations

(Continued...)

Learning Objectives

• Explain the role of data warehouses in decision support
• Explain data integration and the extraction, transformation, and load (ETL) processes
• Describe real-time (a.k.a. right-time and/or active) data warehousing
• Understand data warehouse administration and security issues

Opening Vignette…

Isle of Capri Casinos Is Winning with Enterprise Data Warehouse

• Company background
• Problem description
• Proposed solution
• Results
• Answer & discuss the case questions.

Questions for the Opening Vignette

1. Why is it important for Isle to have an EDW?
2. What were the business challenges or opportunities that Isle was facing?
3. What was the process Isle followed to realize EDW? Comment on the potential challenges Isle might have had going through the process of EDW development.
4. What were the benefits of implementing an EDW at Isle? Can you think of other potential benefits that were not listed in the case?
5. Why do you think large enterprises like Isle in the gaming industry can succeed without having a capable data warehouse/business intelligence infrastructure?

• 1. Why is it important for Isle to have an EDW?

• In the gaming industry, companies distinguish themselves based on their customer relationships. An enterprise data warehouse (EDW) gathers and provides the data needed to tell Isle of Capri what customers respond to, so the casinos can adapt their offerings.

• The information provided by the EDW lets Isle deepen its understanding of customers, so it can efficiently give them more of the kinds of entertainment they are looking for.
2. What were the business challenges or opportunities that Isle was facing?

- Isle of Capri Casinos is one of the largest publicly traded gaming companies in the United States, but the gaming business is competitive. Other casinos offer essentially the same games, so Isle must find ways to make its entertainment and hospitality atmosphere one that exceeds customer expectations. Before Isle implemented the EDW, casino managers had to wait to review monthly data until the second week of the following month.
- The time lag made it difficult for casinos to identify what actions were appealing to customers in time to respond. Adding to these business challenges, Hurricane Katrina set back initial efforts to set up a data warehouse at the southeastern company.

3. What was the process Isle followed to realize EDW? Comment on the potential challenges Isle might have had going through the process of EDW development.

- Isle of Capri brought in two expert suppliers. Teradata provided the core solution; IBM Cognos provided expertise in business intelligence. Isle hired a management team that understood how EDW could support decision making at Isle.
- That team would be able to help Isle’s managers with the challenge of understanding how they can frame queries and follow-up questions to figure out ways to improve the business. Most likely, the potential of using detailed, up-to-the-minute data would be unfamiliar to many of these managers.

4. What were the benefits of implementing an EDW at Isle? Can you think of other potential benefits that were not listed in the case?

- The implementation of EDW brought several benefits related to the timeliness and detail of the data that became available. Instead of five week-old monthly reports, managers can now study a variety of daily, weekly, and monthly reports. The reports segment data by particular properties and customer groups, so managers can zero in on particular problems and successes, easily making comparisons among properties.
- Managers can submit queries about data sets and receive information within minutes. In this way, managers can find out how particular promotions are affecting customer behavior at particular casinos. The EDW also connects data about casino activity with data about customers’ use of hotels and efforts by Isle’s hosts. This, too, helps the company target promotions and offer customers incentives they value.
- Even decisions as detailed as where to locate slot machines can be adjusted to boost profits based on data from the EDW.

5. Why do you think large enterprises like Isle in the gaming industry can succeed without having a capable data warehouse/business intelligence infrastructure?

- In the past, businesses in the gaming industry could succeed without a capable data warehouse/business intelligence infrastructure because their managers knew as much about customers as their competitors’ management knew.
- They were all testing ideas for promotion or entertainment and responding to customer behavior at about the same pace. However, when one company such as Isle begins to respond to daily, property-by-property data, choosing the marketing and entertainment options that deliver exactly what most profitably lures customers, competitors without that capability will soon begin to suffer.

Main Data Warehousing Topics

- DW definition
- Characteristics of DW
- Data Marts
- ODS, EDW, Metadata
- DW Framework
- DW Architecture & ETL Process
- DW Development
- DW Issues

Key Fields

- Keys are special fields that serve two main purposes:
  - **Primary keys** are unique identifiers of the relation. Examples include employee numbers, social security numbers, etc. This guarantees that all rows are unique.
  - **Foreign keys** are identifiers that enable a dependent relation (on the many side of a relationship) to refer to its parent relation (on the one side of the relationship).
- Keys can be simple (a single field) or composite (more than one field).
- Keys usually are used as indexes to speed up the response to user queries.
Referential Integrity

Can we “delete” customers#5 if orders#1 is still in the database? Why?

<table>
<thead>
<tr>
<th>customers#5</th>
<th>Orders#1</th>
</tr>
</thead>
<tbody>
<tr>
<td>pk: primary key</td>
<td>fk: foreign key</td>
</tr>
<tr>
<td>customer#</td>
<td>OrderID</td>
</tr>
<tr>
<td>1003</td>
<td>MORALES</td>
</tr>
<tr>
<td>1005</td>
<td>GIRARD</td>
</tr>
<tr>
<td>1020</td>
<td>FALAH</td>
</tr>
</tbody>
</table>

Can we “create” orders#1 if customers#5 is not created? Why?

Normalized vs. De-normalized

- We will study the concept and technique of “normalization and de-normalization” (especially for those who do not have background on database management) as well as OLTP and OLAP.

More on OLTP vs. OLAP

• The figure depicts a relational database environment with two tables.
  - The first table contains information about pet owners; the second information about pets. The tables are related by the single column they have in common: Owner_ID.
  - By relating tables to one another, we can reduce redundancy of data and improve database performance.
  - The process of breaking tables apart and thereby reducing data redundancy is called normalization.

Fig. Extra-a: A simple database with a relation between two tables. For those who have database background.

OLTP vs. OLAP (cont.)

• Most relational databases which are designed to handle a high number of reads and writes (updates and retrievals of information) are referred to as OLTP (OnLine Transaction Processing) systems.
  - OLTP systems are very efficient for high volume activities such as cashiering, where many items are being recorded via bar code scanners in a very short period of time.
  - However, using OLTP databases for analysis is generally not very efficient, because in order to retrieve data from multiple tables at the same time, a query containing joins must be used.

Why need Data Warehouses?

- A solution to …
  - A solution to bridging the information gap is the data warehouses which consolidate and integrate information from many different sources and arrange it in a meaningful format for making accurate business decisions.

OLTP vs. OLAP (cont.)

• In order to keep our transactional databases running quickly and smoothly we may wish to create a data warehouse. A data warehouse is a type of large database (including both current and historical data) that has been denormalized and archived.
  - Denormalization is the process of intentionally combining some tables into a single table in spite of the fact that this may introduce duplicate data in some columns.

Fig. Extra-b: A combination of the tables into a single dataset.

Why need Data Warehouses?

- A solution to …
  - A solution to bridging the information gap is the data warehouses which consolidate and integrate information from many different sources and arrange it in a meaningful format for making accurate business decisions.

OLTP vs. OLAP (cont.)

• Most relational databases which are designed to handle a high number of reads and writes (updates and retrievals of information) are referred to as OLTP (OnLine Transaction Processing) systems.
  - OLTP systems are very efficient for high volume activities such as cashiering, where many items are being recorded via bar code scanners in a very short period of time.
  - However, using OLTP databases for analysis is generally not very efficient, because in order to retrieve data from multiple tables at the same time, a query containing joins must be used.

Why need Data Warehouses?

- A solution to …
  - A solution to bridging the information gap is the data warehouses which consolidate and integrate information from many different sources and arrange it in a meaningful format for making accurate business decisions.
What is a Data Warehouse?

- A physical repository where relational data or multidimensional structure (including both current and historical data) are specially organized to provide enterprise-wide, cleansed data in a standardized format for analytical processing activities (i.e., OLAP, data mining, querying, reporting and other decision support apps).
- “The data warehouse is a collection of integrated, subject-oriented databases designed to support DSS functions, where each unit of data is non-volatile and relevant to some moment in time”

Characteristics of DW

- Subject-oriented
  - e.g. customers, patients, students, products
- Integrated
  - Consistent naming conventions, formats, encoding structures; from multiple data sources
- Time variant (time series)
  - Can study trends and changes
- Non-volatile
  - Read-only, periodically refreshed
- Summarized
- Not normalized
- Metadata
- Web based, relational/multi-dimensional
- Client/server
- Real-time and/or right-time (active)

Types of Data Warehouses

- Metadata
  - Data about data. In a data warehouse, metadata describe the structure of and some meaning about data and the manner of its acquisition and their effective or ineffective use.

(Three) Types of Data Warehouses

1. Data Mart
   - A departmental data warehouse that stores only relevant data
     - Dependent data mart
       - A subset that is created directly from a data warehouse
     - Independent data mart
       - A small data warehouse designed for a strategic business unit or a department

2. Operational data stores (ODS)
   - A type of database often used as an interim area for a data warehouse for short-term decisions involving mission-critical apps.

3. Enterprise data warehouse (EDW)
   - A data warehouse for the enterprise for medium- and long-term decisions.
   - Operational marts: An operational data mart and it is created when operational data need to be analyzed multi-dimensionally. Its data come from an ODS.

A Historical Perspective to Data Warehousing

- Mainframe computers
- Simple data entry
- Routine reporting
- Primitive database structures
- Centralized data storage
- Data warehousing was born
- Inmon, Building the Data Warehouse
- Kimball, The Data Warehouse Toolkit
- EDW architecture design
- Big Data analytics
- Social media analytics
- Text and Web Analytics
- In-memory, in-database
- Mini/personal computers (PCs)
- Business applications for PCs
- Distributor DBMS
- Relational DBMS
- Teradata commercial DBs
- Business Data Warehouse scientist

An Example of CUSTOMERS Data File (Part of Database)

Which one is the “Metadata” for CUSTOMERS data file?
Application Case 2.1

A Better Data Plan: Well-Established TELCOs Leverage Data Warehousing and Analytics to Stay on Top in a Competitive Industry

Questions for Discussion
1. What are the main challenges for TELCOs?
2. How can data warehousing and data analytics help TELCOs in overcoming their challenges?
3. Why do you think TELCOs are well suited to take full advantage of data analytics?

• 1. What are the main challenges for TELCOs?
• To stay competitive, TELCOs must continuously refine everything from customer service to plan pricing.
• The major challenges faced by both entrenched and new companies in this industry include: retaining customers, decreasing costs, fine-tuning pricing models, improving customer satisfaction, acquiring new customers, and understanding the role of social media in customer loyalty.

2. How can data warehousing and data analytics help TELCOs in overcoming their challenges?
• Highly targeted data analytics play an ever more critical role in helping carriers secure or improve their standing in an increasingly competitive marketplace.
• Argentina’s Telefónica de Argentina used analytics for its “traceability project,” which tracked the factors involved in customer churn, a big problem among phone service carriers.
• France’s Bouygues Telecom used BI technologies to facilitate cost reduction through automation via its Teradata-based marketing operations management system, which automates marketing/communications collateral production.
• Pakistan’s Mobilink uses BI to help acquire customers and grow their subscriber network, largely aided by social networking.

3. Why do you think TELCOs are well suited to take full advantage of data analytics?
• TELCOs control the telecommunications infrastructure, and acquire much usage data as a result.
• They have the technical expertise to create, deploy, and refine plans to address their business challenges. The industry and mobile technology have expanded and improved over the years, which provides a strong foundation on which to build intelligent solutions.
• The data analytics solutions that have been created to meet these challenges have also improved drastically over the past few years, placing TELCOs in a good position to capitalize on their technological advantages.

2.2 Data Warehousing Process Overview

Fig. 2.2 Data-Driven Decision Making – Business Benefits of the Data Warehouse
**A Generic DW Framework**

![A Generic DW Framework Diagram](image)

**Fig. 2.3 A Data Warehousing Framework and Views**

---

**Application Case 2.2**

**Data Warehousing Helps MultiCare Save More Lives**

**Questions for Discussion**

1. What do you think is the role of data warehousing in healthcare systems?
2. How did MultiCare use data warehousing to improve health outcomes?

---

**DW Architectures**

**3-tier architecture**

1. Tier 1: Client workstation
2. Tier 2: Application server
3. Tier 3: Database server

**2-tier architecture**

1. Tier 1: Client workstation
2. Tier 2: Application & database server

**4-tier Architecture?**

---

**A Web-based DW Architecture**

![A Web-based DW Architecture Diagram](image)

**Advantages:** ease of access, platform independence, and lower cost.
Information interdependence

Strategic view of the data
Nature of end load
Upper management’s transform
Technical issues transform
Compatibility with existing systems
Scrub or data cleansing
Social/political factors cleanse
Perceived ability of the in-house staff

Transformation, and Load (ETL) Process

ETL = Extract, transform, and load

The ETL Process – another perspective and example

- Capture/Extract - E
- Scrub or data cleansing
- Transform - T
- Load and Index - L
Dr. Chen,

Business Intelligence

43

Static extract = capturing a snapshot of the source data at a point in time

Incremental extract = capturing changes that have occurred since the last static extract

Capture/Extract...obtaining a snapshot of a chosen subset of the source data for loading into the data warehouse

Scrub/Cleanse...uses pattern recognition and AI techniques to upgrade data quality

Fixing errors: misspellings, erroneous dates, incorrect field usage, mismatched addresses, missing data, duplicate data, inconsistencies

Also: decoding, reformatting, time stamping, conversion, key generation, merging, error detection/logging, locating missing data

Transform = convert data from format of operational system to format of data warehouse

Record-level:
Selection – data partitioning
Joining – data combining
Aggregation – data summarization

Field-level:
single-field – from one field to one field
multi-field – from many fields to one, or one field to many

Load/Index = place transformed data into the warehouse and create indexes

Refresh mode:
bulk rewriting of target data at periodic intervals

Update mode:
only changes in source data are written to data warehouse

Data Integration and the Extraction, Transformation, and Load (ETL) Process

• ETL = Extract Transform Load

• Data integration
Integration that comprises three major processes: data access, data federation, and change capture.

• Enterprise application integration (EAI)
A technology that provides a vehicle for pushing data from source systems into a data warehouse

• Enterprise information integration (EII)
An evolving tool space that promises real-time data integration from a variety of sources, such as relational or multidimensional databases, Web services, etc.

Information Cleansing or Scrubbing

• An organization must maintain high-quality data in the data warehouse

• Information cleansing or scrubbing – a process that weeds out and fixes or discards inconsistent, incorrect, or incomplete information
Information Cleansing or Scrubbing

- Standardizing Customer name from Operational Systems

Sales

Customer
- Jane Smith
- Bob Lake
- Judy Smith
- Pat Burton

Customer Information

Customer: 10001
- Jane Smith
- Robert P. Lake
- Julie R. Smith
- Patricia Burton

Billing

Customer: 10001
- Berton, Todd
- Smith, Julie
- Berton, Patricia
- Lake, Robert

Information Cleansing or Scrubbing

- Accurate and complete information

Quality Management

100% Completeness

Not very useful - May be a prototype only

Very incomplete but accurate

Perfect information price

Complete but with known errors

Data Warehouse Development

- Data warehouse development approaches
  - (1) Inmon Model: EDW approach (top-down)
    - Adapts traditional RDB tools to the development needs of an EDW
  - (2) Kimball Model: Data mart approach (bottom-up)
    - Employs dimensional data modeling.
  - Which model is best?
    - There is no one-size-fits-all strategy to DW
  - One alternative is the hosted warehouse

- Data warehouse structure:
  - The Star Schema vs. Relational

- Table 2.3 provides a comparative analysis between EDW and Data Mart approach

Application Case 2.5

Starwood Hotels & Resorts Manages Hotel Profitability with Data Warehousing

Questions for Discussion

1. How big and complex are the business operations of Starwood Hotels & Resorts?
2. How did Starwood Hotels & Resorts use data warehousing for better profitability?
3. What were the challenges, the proposed solution, and the obtained results?

Data Warehousing Strategy

- A data warehouse strategy is a blue print for the successful introduction of the DW.
- The strategy should describe
  - Where the company wants to go,
  - Why it wants to go there,
  - What it will do when it gets there
- It needs to take into consideration the organization’s vision, structure, and culture.
Additional Data Warehouse Considerations
Hosted Data Warehouses

- Benefits:
  - Requires minimal investment in infrastructure
  - Frees up capacity on in-house systems
  - Frees up cash flow (why?)
  - Makes powerful solutions affordable
  - Enables solutions that provide for growth
  - Offers better quality equipment and software
  - Provides faster connections
  - … more in the book

Representation of Data in DW

- Dimensional Modeling – a retrieval-based system that supports high-volume query access
  - Not only accommodate but also boost the processing of complex multidimensional queries.
- Two means
  - 1. Star schema – the most commonly used and the simplest style of dimensional modeling
    - Contains a fact table surrounded by and connected to several dimension tables
    - Fact table contains the descriptive attributes (numerical values) needed to perform decision analysis and query reporting, and foreign keys are used to link to dimension table.
    - Dimension tables contain classification and aggregation information about the values in the fact table (i.e., attributes describing the data contained within the fact table).
  - 2. Snowflakes schema – an extension of star schema where the diagram resembles a snowflake in shape

Fact Table vs. Dimensional Table

Many to Many Relationship (M:N)

<table>
<thead>
<tr>
<th>Dimension Table</th>
<th>Fact Table</th>
<th>Dimension Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order ID</td>
<td>Order Date</td>
<td>Products</td>
</tr>
<tr>
<td>PK</td>
<td>CPK</td>
<td>PK</td>
</tr>
<tr>
<td>Order ID</td>
<td>Date</td>
<td>Products</td>
</tr>
</tbody>
</table>

Star versus Snowflake Schema

Star Schema

- Fact tables contain factual (descriptive) or quantitative data (numerical values)
- Dimension tables contain descriptions about the subjects of the business (values in the fact table)
- 1:N relationship between dimension tables and fact tables
- Dimension tables are denormalized to maximize performance

Snowflake Schema

- Dimension tables are denormalized to maximize performance

Figure (extra): Components of a star schema

- Fact tables contain factual (descriptive) or quantitative data (numerical values)
- Dimension tables are denormalized to maximize performance
- 1:N relationship between dimension tables and fact tables
- Dimension tables contain descriptions about the subjects of the business (values in the fact table)
- Excellent for ad-hoc queries, but bad for online transaction processing

Figure (extra) Star schema example

- Fact table provides statistics for sales broken down by product, period and store dimensions
Multidimensionality
- Multidimensionality
  - The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)
  - In a data warehouse and data mart, information is **multidimensional**, it contains layers of columns and rows
- Multidimensional presentation
  - **Dimensions**: a dimension is a particular attribute of information such as products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry
  - **Measures**: money, sales volume, head count, inventory profit, actual versus forecast
  - **Time**: daily, weekly, monthly, quarterly, or yearly

Analysis of Data in DW
- OLTP vs. OLAP…
- OLTP (online transaction processing)
  - Capturing and storing data from ERP, CRM, POS,…
  - The main focus is on efficiency of routine tasks
- OLAP (Online analytical processing)
  - Converting data into information for decision support
  - Data cubes, drill-down / rollup, slice & dice,…
  - Requesting ad hoc reports
  - Conducting statistical and other analyses
  - Developing multimedia-based applications
  - …more in the book

OLAP and its Applications
- What software and function that enable you to create OLAP and its applications?
- **ANSWER**
  - EXCEL with
  - Pivot table

OLTP and OLAP - Rely Heavily on Each Other

<table>
<thead>
<tr>
<th>Criteria</th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td>To carry out day-to-day business functions</td>
<td>To support decision making and provide answers to business and management queries</td>
</tr>
<tr>
<td><strong>Data source</strong></td>
<td>Transaction database (a normalized data repository primarily focused on efficiency and consistency)</td>
<td>Data warehouse or data mart (a nonnormalized data repository primarily focused on accuracy and completeness)</td>
</tr>
<tr>
<td><strong>Reporting</strong></td>
<td>Routine, periodic, narrowly focused reports</td>
<td>Ad-hoc, multidimensional, broadly focused reports and queries</td>
</tr>
<tr>
<td><strong>Resource requirements</strong></td>
<td>Ordinary relational databases</td>
<td>Multiprocessor, large-capacity, specialized databases</td>
</tr>
<tr>
<td><strong>Execution speed</strong></td>
<td>Fast (recording of business transactions and routine reports)</td>
<td>Slow (resource intensive, complex, large-scale queries)</td>
</tr>
</tbody>
</table>
**OLAP Operations**

- **Cube**
  - A multidimensional data structure that allows fast analysis of data.
  - *common term for the representation of multidimensional information*

- **Slice**
  - A subset of a multidimensional array (via rotations)

- **Dice**
  - A slice on more than two dimensions

- **Drill Down/Up**
  - Navigating among levels of data ranging from the most summarized (up) to the most detailed (down) (via aggregation and disaggregation)

- **Roll Up**
  - Computing all of the data relationships for one or more dimensions

- **Pivot**
  - Used to change the dimensional orientation of a report or an ad hoc query

---

**Variations of OLAP**

- **Multidimensional OLAP (MOLAP)**
  - OLAP implemented via a *specialized* multidimensional database (or data store) that summarizes transactions into multidimensional views ahead of time

- **Cube structure**

- **Relational OLAP (ROLAP)**
  - The implementation of an OLAP database on top of an existing relational database and does not require pre-computation and storage of information.

- **Database OLAP and Web OLAP (DOLAP and WOLAP); Desktop OLAP…**

---

**Figure (extra): Slicing a data cube**

**Figure (extra): Multidimensional Analysis**

- **Cube** — common term for the representation of multidimensional information

---

**Figure (extra): Example of drill-down**

**Summary report**

Starting with summary data, users can obtain details for particular cells

**Drill-down with color added**

---
BPM systems allow managers to measure, monitor, and manage key activities and processes to achieve organizational goals. Dashboards are often used to provide an information system in support of BPM.

Charts like these are examples of data visualization, the representation of data in graphical and multimedia formats for human analysis.

Successful DW Implementation - Things to Avoid

- Data warehousing risks are more serious (than other IT projects) because DW are expensive, time-and-resource demanding, large-scale project.
- Therefore, a successful DW implementation should avoid:
  - Starting with the wrong sponsorship chain
  - Setting expectations that you cannot meet
  - Engaging in politically naive behavior
  - Loading the data warehouse with information just because it is available
  - Believing that data warehousing database design is the same as transactional database design
  - Choosing a data warehouse manager who is technology oriented rather than user oriented

Failure Factors in DW Projects

- Lack of executive sponsorship
  - What are the “three” types of people should be involved in any project?
  - (1) (Top) Manager
  - (2) Specialist
  - (3) Users
- Unclear business objectives
- Cultural issues being ignored
- Change management
- Unrealistic expectations
- Inappropriate architecture
- Low data quality / missing information
- Loading data just because it is available

Massive DW and Scalability

- Scalability
  - The main issues pertaining to scalability:
    - The amount of data in the warehouse
    - How quickly the warehouse is expected to grow
    - The number of concurrent users
    - The complexity of user queries
  - Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse

Real-time/Active DW/BI

- Enabling real-time data updates for real-time analysis and real-time decision making is growing rapidly
  - RDW may be successful if the organization develops a sound methodology to handle project risks, incorporates proper planning, and focuses on quality assurance activities.
  - RDW is also known as active data warehouse (ADW)
  - Push vs. Pull (of data)
- Concerns about real-time BI
  - Not all data should be updated continuously
  - Mismatch of reports generated minutes apart
  - May be cost prohibitive
  - May also be infeasible
How are business intelligence applications delivered?

This figure shows the components of a generic BI system. A BI application server delivers results in a variety of formats to devices for consumption by BI users. A BI server provides two functions: 1) management and 2) delivery.

- Operational data
- Data warehouse
- Data mart
- Human interaction
- EFM
- OLAP reports
- Other data mining
- Context indexing
- Web front
- Expert system

Pull – deliver the information ONLY the user really needed

Components of Generic Business Intelligence System (Kroenke, Elizar, My 2013)

Real-Time/Active DW at Teradata

Active Access
- Time: Operational decisions or services supported by real-time (RT) access; Service Level Agreements of 5 seconds or less

Active Load
- Real-time data access; Mini-batch to RT trickle data feeds measured in minutes or seconds

Active Events
- Proactive monitoring of business activity; relating intelligent actions based on rules and context; to systems or users supporting an operational business process

Active Workload
- Management
- Dynamically manages system resources for optimum performance
- Supports a mixed-workload environment

Active Enterprise Integration
- Integration into the Enterprise Architecture
- Delivery of intelligent decisioning services

Active Availability
- Business Continuity
- Support all functions of the business (up to 7x24)

 DW Administration and Security

- Data warehouse administrator (DWA)
  - DWA should…
    - have the knowledge of high-performance software, hardware and networking technologies
    - possess solid business knowledge and insight
    - be familiar with the decision-making processes so as to suitably design/maintain the data warehouse structure
    - possess excellent communications skills
  - Security and privacy is a pressing issue in DW
    - Safeguarding the most valuable assets
    - Government regulations (HIPAA, etc.)
    - Must be explicitly planned and executed

Traditional versus Active DW

Traditional Data Warehouse Environment
- Strategic decisions only
- Results measured with operators
- High number (1,000 or more) of users accessing and querying the system simultaneously
- Flexible ad hoc reporting, as well as machine-assisted modeling (e.g., data mining) to discover new hypotheses and relationships
- Operational staffs, call centers, external users

Active Data Warehouse Environment
- Strategic and tactical decisions
- Only comprehensive detailed data available within minutes is acceptable
- Full-text search of data available
- Advanced analytics

The Future of DW

- Sourcing...
  - Web, social media, and Big Data
  - Open source software
  - SaaS (software as a service)
  - Cloud computing

- Infrastructure... (architectural: hardware and software enhancements)
  - Real-time DW and Columnar (vs. Relational DB – stored as rows)
  - Data warehouse appliances (best-of-breed philosophy in the future)
  - Data management practices/technologies
  - In-database & In-memory processing New DBMS (“super-computing”)
  - Advanced analytics
  - …
Big Data And Data Warehousing

- Two paradigms in BI:
  - Data Warehouse and Big Data.
  - Both are competing each other for turning data into actionable information.
- However, in recent years, the variety and complexity of data made data warehouse incapable of keeping up the changing needs.
- Big Data
  - A new paradigm that the world of IT was forced to develop, not because the volume of the structured data but the variety and the velocity.
- Big data and analytics will be further discussed in chapter 6.

Conclusion

- The future of data warehousing seems to be full of promises and significant challenges. As the world of business becomes more global and complex, the need for business intelligence and DW tools also become more prominent.
- The fast improving IT tools and techniques seem to be moving in the right direction to address the needs of the future BI systems.

End of Chapter Application Case

- Continental Airlines Flies High with its Real-Time Data Warehouse

1. Describe the benefits of implementing the Continental Go Forward strategy.
   - This strategy consisted of a number of interrelated, concurrent actions. The first version of the overall strategy had the benefit of restoring Continental (CO) to profitability and giving it first-place rankings by many airline industry metrics.
   - The second phase of the strategy led to savings of $41 million and a reduction of $7 million in fraud in the first year alone. Its revenue increased by over $300 million in six years. A data warehouse played a critical role in the second phase.

2. Explain why it is important for an airline to use a real-time data warehouse.
   - It’s important for an airline to use real-time data warehouse because many airline decisions cannot be made with week-old, or even day-old, data.
   - An example is frequent flyer award availability on a given flight. Airlines limit these so as not to give away too many seats that would otherwise be sold. Award seat allocation is usually automated. Travelers can check availability online. American Airlines (and probably others) offers expanded award availability to the 30,000 or so people who fly at least 100,000 miles per year with them. When one of these wants an award seat that is not available online, he or she can request it by phone. The agent must decide whether or not to make it available. If the flight is selling slowly, the traveler gets the seat, even if the computer hasn’t allocated it for award. If it is likely to sell out, the seat isn’t offered, even to this select group. To make this decision, telephone agents (and the Yield/Revenue Management staff, which agents can consult) need up-to-the-minute, or at least up-to-the-hour (i.e., real-time), information.
4. Identify the major differences between the traditional data warehouse and a real-time data warehouse, as was implemented at Continental.

• A traditional data warehouse moves data from operational databases to the DW on a scheduled basis, typically daily or weekly. This provides consistent data for analyses performed during one update cycle, but does not make current information available for decisions that require it.
• A real-time DW, as was implemented at CO, moves data (from legacy systems) into the DW on an hourly or even more frequent basis.

5. What strategic advantage can Continental derive from the real-time system as opposed to a traditional information system?

• By having real-time data available through its data warehouse, CO can make decisions using up-to-date information. While data warehousing applications which focus on long-term decisions aren’t affected much by the last hour’s, day’s or even week’s data, lower-level short-term decisions are.
• As the use of the DW is extended to these decisions and down in the organization, current data become necessary. By having real-time (or near-real-time) data in the system, CO obtains a strategic advantage by making better decisions.

End of the Chapter

• Questions, comments