Commercial Intelligence Practices

資料倉儲
(Data Warehousing)

Min-Yuh Day
Assistant Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2015-03-18
<table>
<thead>
<tr>
<th>週次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2015/02/25</td>
<td>商業智慧導論 (Introduction to Business Intelligence)</td>
</tr>
<tr>
<td>2</td>
<td>2015/03/04</td>
<td>管理決策支援系統與商業智慧 (Management Decision Support System and Business Intelligence)</td>
</tr>
<tr>
<td>3</td>
<td>2015/03/11</td>
<td>企業績效管理 (Business Performance Management)</td>
</tr>
<tr>
<td>4</td>
<td>2015/03/18</td>
<td>資料倉儲 (Data Warehousing)</td>
</tr>
<tr>
<td>5</td>
<td>2015/03/25</td>
<td>商業智慧的資料探勘 (Data Mining for Business Intelligence)</td>
</tr>
<tr>
<td>6</td>
<td>2015/04/01</td>
<td>教學行政觀摩日 (Off-campus study)</td>
</tr>
<tr>
<td>7</td>
<td>2015/04/08</td>
<td>商業智慧的資料探勘 (Data Mining for Business Intelligence)</td>
</tr>
<tr>
<td>8</td>
<td>2015/04/15</td>
<td>資料科學與巨量資料分析 (Data Science and Big Data Analytics)</td>
</tr>
<tr>
<td>週次</td>
<td>日期</td>
<td>內容（Subject/Topics）</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>9</td>
<td>2015/04/22</td>
<td>期中報告 (Midterm Project Presentation)</td>
</tr>
<tr>
<td>10</td>
<td>2015/04/29</td>
<td>期中考試週 (Midterm Exam)</td>
</tr>
<tr>
<td>11</td>
<td>2015/05/06</td>
<td>文字探勘與網路探勘 (Text and Web Mining)</td>
</tr>
<tr>
<td>12</td>
<td>2015/05/13</td>
<td>意見探勘與情感分析 (Opinion Mining and Sentiment Analysis)</td>
</tr>
<tr>
<td>13</td>
<td>2015/05/20</td>
<td>社會網路分析 (Social Network Analysis)</td>
</tr>
<tr>
<td>14</td>
<td>2015/05/27</td>
<td>期末報告 (Final Project Presentation)</td>
</tr>
<tr>
<td>15</td>
<td>2015/06/03</td>
<td>畢業考試週 (Final Exam)</td>
</tr>
</tbody>
</table>
Business Intelligence

Data Mining, Data Warehouses

Increasing potential to support business decisions

Decision Making

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

End User
Business Analyst
Data Analyst
DBA

Source: Han & Kamber (2006)
A High-Level Architecture of BI

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Chapter 8:
Data Warehousing
Learning Objectives

• Definitions and concepts of data warehouses
• Types of data warehousing architectures
• Processes used in developing and managing data warehouses
• Data warehousing operations
• Role of data warehouses in decision support
• Data integration and the extraction, transformation, and load (ETL) processes
• Data warehouse administration and security issues

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Main Data Warehousing (DW) Topics

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

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Warehouse Defined

• A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format

• “The data warehouse is a collection of integrated, subject-oriented databases design to support DSS functions, where each unit of data is non-volatile and relevant to some moment in time”

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Characteristics of DW

• Subject oriented
• Integrated
• Time-variant (time series)
• Nonvolatile
• Summarized
• Not normalized
• Metadata
• Web based, relational/multi-dimensional
• Client/server
• Real-time and/or right-time (active)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
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

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Warehousing Definitions

• **Operational data stores (ODS)**
  A type of database often used as an interim area for a data warehouse

• **Oper marts**
  An operational data mart.

• **Enterprise data warehouse (EDW)**
  A data warehouse for the enterprise.

• **Metadata**
  Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its acquisition and use

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
A Conceptual Framework for DW

Data Sources
- ERP
- Legacy
- POS
- Other OLTP/WEB
- External data

ETL Process
- Select
- Extract
- Transform
- Integrate
- Load

Enterprise Data warehouse
- Metadata
- Replication

Data Sources
- ERP
- Legacy
- POS
- Other OLTP/WEB
- External data

Access
- Data mart (Marketing)
- Data mart (Engineering)
- Data mart (Finance)
- Data mart (...)

Applications (Visualization)
- Routine Business Reporting
- Data/text mining
- OLAP, Dashboard, Web
- Custom built applications

No data marts option

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Generic DW Architectures

• **Three-tier architecture**
  1. Data acquisition software (back-end)
  2. The data warehouse that contains the data & software
  3. Client (front-end) software that allows users to access and analyze data from the warehouse

• **Two-tier architecture**
  First 2 tiers in three-tier architecture is combined into one
  ... sometime there is only one tier?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Generic DW Architectures

3-tier architecture

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

2-tier architecture

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

1-tier Architecture?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
DW Architecture Considerations

• Issues to consider when deciding which architecture to use:
  – Which database management system (DBMS) should be used?
  – Will parallel processing and/or partitioning be used?
  – Will data migration tools be used to load the data warehouse?
  – What tools will be used to support data retrieval and analysis?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
A Web-based DW Architecture

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Alternative DW Architectures

(a) Independent Data Marts Architecture

Source Systems → Staging Area → Independent data marts (atomic/summarized data) → End user access and applications

(b) Data Mart Bus Architecture with Linked Dimensional Datamarts

Source Systems → Staging Area → Dimensionalized data marts linked by conformed dimensions (atomic/summarized data) → End user access and applications

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Alternative DW Architectures

(c) Hub and Spoke Architecture (Corporate Information Factory)

(d) Centralized Data Warehouse Architecture

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Alternative DW Architectures

(e) Federated Architecture

Existing data warehouses
Data marts and legacy systems

Data mapping / metadata
Logical/physical integration of common data elements

End user access and applications

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
## Alternative DW Architectures

### Data Mart Centric
- **Sources**
- **Marts**
- **Users**

### Virtual, Distributed, Federated
- **Sources**
- **Middleware**
- **Users**

### Hub-and-Spoke Data Warehouse
- **Sources**
- **Data Warehouse**
- **Marts**
- **Users**

### Enterprise Data Warehouse
- **Sources**
- **Data Warehouse**
- **Users**

<table>
<thead>
<tr>
<th>Independent Data Marts</th>
<th>Leave Data Where it Lies</th>
<th>Dependent Data Marts</th>
<th>Centralized Integrated Data With Direct Access</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td><strong>Cons</strong></td>
<td><strong>Pros</strong></td>
<td><strong>Cons</strong></td>
</tr>
<tr>
<td>Easy to Build Organizationally</td>
<td>Business Enterprise view unavailable</td>
<td>Allows easier customization of user interfaces &amp; reports</td>
<td>Requires corporate leadership and vision</td>
</tr>
<tr>
<td>Easy to Build Technically</td>
<td>Redundant data costs</td>
<td>No need for ETL</td>
<td>Business Enterprise view</td>
</tr>
<tr>
<td></td>
<td>High ETL costs</td>
<td>No need for separate platform</td>
<td>Design consistency &amp; data quality</td>
</tr>
<tr>
<td></td>
<td>High App costs</td>
<td>Only viable for low volume</td>
<td>Data reusability</td>
</tr>
<tr>
<td></td>
<td>High DBA and operational costs</td>
<td>Meta data issues</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Network bandwidth and join complexity issues</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Workload typically placed on workstation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Business Enterprise view challenging</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Redundant data costs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>High DBA and operational costs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data latency</td>
<td></td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Which Architecture is the Best?

- Bill Inmon versus Ralph Kimball
- Enterprise DW versus Data Marts approach

<table>
<thead>
<tr>
<th></th>
<th>Independent Data Marts</th>
<th>Bus Architecture</th>
<th>Hub-and-Spoke Architecture</th>
<th>Centralized Architecture (No Dependent Data Marts)</th>
<th>Federated Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Quality</td>
<td>4.42</td>
<td>5.16</td>
<td>5.35</td>
<td>5.23</td>
<td>4.73</td>
</tr>
<tr>
<td>System Quality</td>
<td>4.59</td>
<td>5.60</td>
<td>5.56</td>
<td>5.41</td>
<td>4.69</td>
</tr>
<tr>
<td>Individual Impacts</td>
<td>5.08</td>
<td>5.80</td>
<td>5.62</td>
<td>5.64</td>
<td>5.15</td>
</tr>
<tr>
<td>Organizational Impacts</td>
<td>4.66</td>
<td>5.34</td>
<td>5.24</td>
<td>5.30</td>
<td>4.77</td>
</tr>
</tbody>
</table>

*Empirical study by Ariyachandra and Watson (2006)*

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Ten factors that potentially affect the architecture selection decision:

1. Information interdependence between organizational units
2. Upper management’s information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Enterprise Data Warehouse
(by Teradata Corporation)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Integration and the Extraction, Transformation, and Load (ETL) Process

• 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

• Service-oriented architecture (SOA)
  A new way of integrating information systems

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Integration and the Extraction, Transformation, and Load (ETL) Process

Extraction, transformation, and load (ETL) process

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
ETL

• Issues affecting the purchase of and ETL tool
  – Data transformation tools are expensive
  – Data transformation tools may have a long learning curve

• Important criteria in selecting an ETL tool
  – Ability to read from and write to an unlimited number of data sources/architectures
  – Automatic capturing and delivery of metadata
  – A history of conforming to open standards
  – An easy-to-use interface for the developer and the functional user

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Benefits of DW

• Direct benefits of a data warehouse
  – Allows end users to perform extensive analysis
  – Allows a consolidated view of corporate data
  – Better and more timely information
  – Enhanced system performance
  – Simplification of data access

• Indirect benefits of data warehouse
  – Enhance business knowledge
  – Present competitive advantage
  – Enhance customer service and satisfaction
  – Facilitate decision making
  – Help in reforming business processes

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Warehouse Development

• Data warehouse development approaches
  – Inmon Model: EDW approach (top-down)
  – Kimball Model: Data mart approach (bottom-up)
  – 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

• Real-time data warehousing?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
## DW Development Approaches

<table>
<thead>
<tr>
<th>Effort</th>
<th>Data Mart Approach</th>
<th>EDW Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scope</strong></td>
<td>One subject area</td>
<td>Several subject areas</td>
</tr>
<tr>
<td><strong>Development time</strong></td>
<td>Months</td>
<td>Years</td>
</tr>
<tr>
<td><strong>Development cost</strong></td>
<td>$10,000 to $100,000+</td>
<td>$1,000,000+</td>
</tr>
<tr>
<td><strong>Development difficulty</strong></td>
<td>Low to medium</td>
<td>High</td>
</tr>
<tr>
<td><strong>Data prerequisite for sharing</strong></td>
<td>Common (within business area)</td>
<td>Common (across enterprise)</td>
</tr>
<tr>
<td><strong>Sources</strong></td>
<td>Only some operational and external systems</td>
<td>Many operational and external systems</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>Megabytes to several gigabytes</td>
<td>Gigabytes to petabytes</td>
</tr>
<tr>
<td><strong>Time horizon</strong></td>
<td>Near-current and historical data</td>
<td>Historical data</td>
</tr>
<tr>
<td><strong>Data transformations</strong></td>
<td>Low to medium</td>
<td>High</td>
</tr>
<tr>
<td><strong>Update frequency</strong></td>
<td>Hourly, daily, weekly</td>
<td>Weekly, monthly</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hardware</strong></td>
<td>Workstations and departmental servers</td>
<td>Enterprise servers and mainframe computers</td>
</tr>
<tr>
<td><strong>Operating system</strong></td>
<td>Windows and Linux</td>
<td>Unix, Z/OS, OS/390</td>
</tr>
<tr>
<td><strong>Databases</strong></td>
<td>Workgroup or standard database servers</td>
<td>Enterprise database servers</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Start Schema Example for an Automobile Insurance Data Warehouse

Dimensions:
How data will be sliced/diced (e.g., by location, time period, type of automobile or driver)

Facts:
Central table that contains (usually summarized) information; also contains foreign keys to access each dimension table.

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
**Dimensional Modeling**

**Data cube**
A two-dimensional, three-dimensional, or higher-dimensional object in which each dimension of the data represents a measure of interest.
- Grain
- Drill-down
- Slicing

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Best Practices for Implementing DW

- The project must fit with corporate strategy
- There must be complete buy-in to the project
- It is important to manage user expectations
- The data warehouse must be built incrementally
- Adaptability must be built in from the start
- The project must be managed by both IT and business professionals (a business–supplier relationship must be developed)
- Only load data that have been cleansed/high quality
- Do not overlook training requirements
- Be politically aware.

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Risks in Implementing DW

• No mission or objective
• Quality of source data unknown
• Skills not in place
• Inadequate budget
• Lack of supporting software
• Source data not understood
• Weak sponsor
• Users not computer literate
• Political problems or turf wars
• Unrealistic user expectations

(Continued ...)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Risks in Implementing DW – Cont.

- Architectural and design risks
- Scope creep and changing requirements
- Vendors out of control
- Multiple platforms
- Key people leaving the project
- Loss of the sponsor
- Too much new technology
- Having to fix an operational system
- Geographically distributed environment
- Team geography and language culture

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Things to **Avoid** for Successful Implementation of DW

- Starting with the wrong sponsorship chain
- Setting expectations that you cannot meet
- Engaging in politically naive behavior
- Loading the warehouse with information just because it is available
- Believing that data warehousing database design is the same as transactional DB design
- Choosing a data warehouse manager who is technology oriented rather than user oriented

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Real-time DW
(a.k.a. Active Data Warehousing)

• Enabling real-time data updates for real-time analysis and real-time decision making is growing rapidly
  – 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

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Evolution of DSS & DW

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Active Data Warehousing
(by Teradata Corporation)

**Active Access**
Front-Line operational decisions or services supported by near-real-time (NRT) access; Service Level Agreements of 5 seconds or less

**Active Load**
Intra-day data acquisition; Mini-batch to NRT trickle data feeds measured in minutes or seconds

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

**Active Workload Management**
Dynamically manage system resources for optimum performance and resource utilization supporting a mixed-workload environment

**Active Enterprise Integration**
Integration into the Enterprise Architecture for delivery of intelligent decisioning services

**Active Availability**
Business Continuity to support the requirements of the business (up to 7X24)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
## Comparing Traditional and Active DW

<table>
<thead>
<tr>
<th>Traditional Data Warehouse Environment</th>
<th>Active Data Warehouse Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategic decisions only</td>
<td>Strategic and tactical decisions</td>
</tr>
<tr>
<td>Results sometimes hard to measure</td>
<td>Results measured with operations</td>
</tr>
<tr>
<td>Daily, weekly, monthly data currency acceptable; summaries often appropriate</td>
<td>Only comprehensive detailed data available within minutes is acceptable</td>
</tr>
<tr>
<td>Moderate user concurrency</td>
<td>High number (1,000 or more) of users accessing and querying the system simultaneously</td>
</tr>
<tr>
<td>Highly restrictive reporting used to confirm or check existing processes and patterns; often uses predeveloped summary tables or data marts</td>
<td>Flexible ad hoc reporting, as well as machine-assisted modeling (e.g., data mining) to discover new hypotheses and relationships</td>
</tr>
<tr>
<td>Power users, knowledge workers, internal users</td>
<td>Operational staffs, call centers, external users</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Warehouse Administration

• Due to its **huge size** and its intrinsic nature, a DW requires especially strong monitoring in order to sustain its efficiency, productivity and security.

• The successful administration and management of a data warehouse entails skills and proficiency that go past what is required of a traditional database administrator.
  – Requires expertise in high-performance software, hardware, and networking technologies

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
**DW Scalability and Security**

- **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

- **Security**
  - Emphasis on security and privacy

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Summary

• Definitions and concepts of data warehouses
• Types of data warehousing architectures
• Processes used in developing and managing data warehouses
• Data warehousing operations
• Role of data warehouses in decision support
• Data integration and the extraction, transformation, and load (ETL) processes
• Data warehouse administration and security issues
References