Integration of Data Warehousing and Operations Analytics

Zhen Liu, PhD
Daniel L. Goodwin College of Business
Benedictine University

Email: zliu@ben.edu
https://www.linkedin.com/in/zhenliu/

Joint work with Erica Arnold and Daniel Kreuger
Datawarehouse Architecture

ETL -> Metadata
  Summary Data
  Raw Data
  OLAP
  Reporting

Operational System
SAP
ERP System
Salesforce
CM System
Flat Files
Third Party Data

MIS674 Database Management Systems

Chapter 4 Logical Database Design and the Relational Model

Zhen Liu, PhD
Goodwin College of Business
Benedictine University

Email: zliu@ben.edu
https://www.linkedin.com/in/zhenliu/

16th IT and Business Analytics Teaching Workshop, June 1, 2018
• Me: I am offering two courses this quarter: Database Management Systems and Data Warehousing
• Friend (CS professor): they are CS courses
• Me: well, they are our MSBA core courses
Motivation/Theme

• My **understanding** of data warehousing
  – MSBA provides a unique perspective
  – Why are we different from Computer Science

• Justify by examples in Operations Analytics
  – Newsvendor problem
  – Inventory pooling under fat-tail demands
Business Analytics Domain

- **Descriptive Statistics**
  - Sampling
  - Mean
  - Mode
  - Median
  - Standard Deviation
  - Range & Variance
  - Stem & Leaf Diagram
  - Histogram
  - Interquartile Range
  - Quartiles
  - Frequency Distributions

- **Forecasting**
  - Time Series
  - Causal Relationships

- **Data Mining**
  - Cluster Analysis
  - Association Analysis
  - Multiple Regression
  - Logistic Regression
  - Decision Tree Methods
  - Neural Networks
  - Text Mining

- **Management Science**
  - Linear Programming
  - Sensitivity Analysis
  - Integer Programming
  - Goal Programming
  - Nonlinear Programming
  - Transportation
  - Logistics
  - Optimization Heuristics
  - Simulation Modeling

**Descriptive Analytics**

**Predictive Analytics**

**Prescriptive Analytics**

**Databases & Data Warehousing**
- Relational Database Modeling
- Structured Query Languages
- Report Generation and Data Visualization
- Dimensional Modeling
- Extract-Transform-Load
- Data Warehousing Schemas
- Online Analytical Processing
- Nonstructured Query Languages
- Distributed File Systems
- Map-Reduce

Benedictine University
Topics

- Introduction to Data Warehousing
- Inventory Management
- Customer Relationship Management (CRM)
- Challenges and Future work
Intro to Data Warehousing

• The term "Data Warehouse" was first coined by Bill Inmon in 1990.

• According to Inmon, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data.
  – This data helps analysts to take informed decisions in an organization.
key features of a data warehouse

• **Subject Oriented** – A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations.
  – These subjects can be product, customers, suppliers, sales, revenue, etc.
  – A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.

• **Integrated** – A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc.

• **Time Variant** – The data collected in a data warehouse is identified with a particular time period.
  – The data in a data warehouse provides information from the historical point of view.

• **Non-volatile** – Non-volatile means the previous data is not erased when new data is added to it.
  – A data warehouse is kept separate from the operational database and therefore frequent changes in operational database is not reflected in the data warehouse.

**Note** – A data warehouse does not require transaction processing, recovery, and concurrency controls, because it is physically stored and separate from the operational database.
Data Warehouse Architecture

Operational System
SAP
ERP System
salesforce
CRM System
Flat Files
Third Party Data

ETL

Metadata
Summary Data
Raw Data

OLAP
Reporting

Benedictine University
## Data Warehouse vs Database

<table>
<thead>
<tr>
<th>Data Warehouse (OLAP)</th>
<th>Operational Database (OLTP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>It involves <strong>historical</strong> processing of information.</td>
<td>It involves <strong>day-to-day</strong> processing.</td>
</tr>
<tr>
<td>OLAP systems are used by knowledge workers such as executives, managers, and analysts.</td>
<td>OLTP systems are used by <strong>clerks, DBAs, or database professionals</strong>.</td>
</tr>
<tr>
<td>It contains historical data.</td>
<td>It contains <strong>current</strong> data.</td>
</tr>
<tr>
<td>It provides <strong>summarized</strong> and <strong>consolidated</strong> data.</td>
<td>It provides primitive and highly detailed data.</td>
</tr>
<tr>
<td>The number of users is in hundreds.</td>
<td>The number of users is in thousands.</td>
</tr>
<tr>
<td>The number of records accessed is in millions.</td>
<td>The number of records accessed is in tens.</td>
</tr>
<tr>
<td>The database size is from 100GB to 100 TB.</td>
<td>The database size is from 100 MB to 100 GB.</td>
</tr>
</tbody>
</table>
Topics

• Introduction to Data Warehousing
• Inventory Management
• Customer Relationship Management (CRM)
• Challenges and Future work
Newsvendor Model in Inventory Management

Adopted from McGraw Hill
Inventory Models

• The inventory *periodic snapshot* where product inventory levels are measured at regular intervals and placed as separate rows in a fact table.
  – These periodic snapshot rows appear over time as a series of data layers in the dimensional model, much like geologic layers represent the accumulation of sediment over long periods of time.

• Second inventory model where
  – every transaction that impacts inventory levels as products move through the warehouse is recorded.

• The *inventory accumulating snapshot* where
  – a fact table row is inserted for each product delivery and then the row is updated as the product moves through the warehouse. Each model tells a different story.

Note that for some analytic requirements, two or even all three models may be appropriate simultaneously.
Inventory Periodic Snapshot

• The right product is in the right store at the right time minimizes out-of-stocks (penalty costs) and reduces overall inventory holding costs.
• The retailer wants to analyze daily quantity-on-hand inventory levels by product and store.
• This leads to an exceptionally clean dimensional design, as shown in next slide.
Figure 4-2: Store Inventory periodic snapshot schema.
• a serious challenge
  • The sales fact table was reasonably **sparse** because you don’t sell every product in every shopping cart.
  • Inventory generates **dense** snapshot tables.
    ✓ Because the retailer strives to avoid out-of-stock situations in which the product is not available, there may be a row in the fact table for every product in every store every day.
• Reduce the snapshot frequencies over time.
  • keep the **last 60 days of inventory** at the daily level and then revert to less granular **weekly snapshots** for historical data.
    ✓ instead of retaining 1,095 snapshots during a 3-year period, the number could be reduced to 208 total snapshots; the 60 daily and 148 weekly snapshots should be stored in two separate fact tables given their unique periodicity.
My Methodology

• How to calibrate the model under this snapshot frequencies?
  – Are there any fat-tail other than Gaussian distribution
  – Shall we combine daily and weekly data for calibration

• Inventory pooling
  – Would a central warehouse reduce the costs?
Topics

• Introduction to Data Warehousing
• Inventory Management
• Customer Relationship Management (CRM)
• Challenges and Future work
Customer Relationship Management (CRM)

- Customer Relationship Management (CRM)
- Customer dimension
  - Behavioral attributes
- Techniques for very large dimensions
  - Mini-dimensions
Customer Relationship Management (CRM)

• Currently a hot topic in business data analysis
• Idea: Gain better understanding of customer behavior by integrating data from various sources
  – Multiple interaction types
    • Orders
    • Returns
    • Customer support
    • Billing
    • Service / repairs
  – Multiple interaction channels
    • Retail store
    • E-mail
    • Call center (Inbound / Outbound)
    • Web site
CRM questions

• Customer retention
  – Which customers are most likely to defect to a competitor?
  – Which retention measures work best?

• Customer acquisition
  – Which prospects are most promising?
  – What offers will entice them to become customers?
Behavioral Attributes

• Customers can be segmented based on past behavior
• Aggregated fact data converted to dimensional attributes
• Examples:
  – Scores based on predictive models
    • Propensity to churn: subscribers to a service who discontinue their subscriptions to that service within a given time period
    • Probability of default
  – Segmentation based on clustering algorithms
  – Raw aggregated data
    • Total dollar sales in past year
**Figure 5-12:** Type 4 mini-dimension with type 1 outrigger in customer dimension.
Topics

• Introduction to Data Warehousing
• Inventory Management
• Customer Relationship Management (CRM)
• Challenges and Future work
Help Needed

- **Data** from industry
  - Hard to get: proprietary
    - e.g., retails, inventory, procurement, finance, and healthcare managements
  - Working with industry firms with interests in BA and machine learning

- **Open access data for machine learning**
  - [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets)
Challenges and Future Work

• Unstructured and complex data in Health Analytics
  – Prescription data
  – Personal health data
    • blood pressure, cholesterol, heart rate, etc.
  – May contain sensitive information

• Big data, Hadoop and Prescriptive Analytics (Fall 2018)
  – [Prescriptive Analytics Is the Future of Big Data](#)