The Evolution of Data Warehouse

Ziyu Lin

Department of Computer Science and Technology

Peking University

Outline

- Architectural Beginnings (1978-1988)
- Enterprise Integration (1988)
- DSS Part 1- Enterprise Data Warehouse (1991)
- DSS Part 2 - Data Marts (1994)
- Data Warehouse: Divergence (1996-97)
- Data Warehouse: Synthesis (1998)
- Data Warehouse : Real-Time Data Warehousing (2000)

Architectural Beginnings (1978-1988)

•The origin of the data warehouse can be traced to studies at MIT in the 1970s which were targeted at developing an optimal technical architecture.

•the craft of data processing was evolving into the profession of information management.

•the MIT researchers differentiated between operational systems and analytic applications.

•In the mid- to late-1980s, Digital Equipment Corporation (DEC) planned to be the showcase for a new approach to application architecture.

Architectural Beginnings (1978-1988)

•Engineers combine the MIT design principles with their newfound expertise in networks and relational databases to create the Technical Architecture 2 (TA-2) specification. TA-2 defined four services categories: data capture, data access, directory and user services.

•User services contained the human-machine interfaces for all the other services.

Enterprise Integration (1988)

•companies faced the daunting task of integrating data from many separate systems with dissimilar coding schemes.

•In 1988, Barry Devlin and Paul Murphy of IBM Ireland tackled the problem of enterprise integration head-on. They introduced the term "information warehouse" for the first time

•Where was the architecture? How do you go about building one of these things?

DSS Part 1- Enterprise Data Warehouse (1991)

•In 1991, Bill Inmon published his first book on data warehousing. Inmon provided the first widely available how-to guide on building a data warehouse. This book contains the most widely published definition of a data warehouse.

•A Data Warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.

- Subject-oriented: Focus on natural data groups, not applications boundaries.
- Integrated: Provide consistent formats and encodings.
- Time-variant: Data is organized by time and is stored in diverse time slices.
- Nonvolatile: No updates are allowed. Only load and retrieval operations.

DSS Part 1- Enterprise Data Warehouse (1991)

- Inmon lays out several other principles that are not a component of his definition.
  - Enterprise Model
  - Enterprise Scope
  - Decision Support
  - Atomic Detail
  - Snap Shot Capture
  - Relational Versus Star

DSS Part 2 - Data Marts (1994)

•The first evidence of a split in the data warehousing ranks came with the introduction of the concept of data marts.

•Vendors and analysts, concerned about losing their meal ticket, took an architectural concept out of context and sold it as a justifiable stand alone solution.

•The term "data mart" was introduced into architectural frameworks as a spin-off of the data warehouse optimized for a particular department or function.

•However, this violated the most fundamental principle of data warehousing: creating a single point of distribution.

•Even with this narrowed scope, many data mart projects also began to fail.

•Luckily, Ralph Kimball's first book, *The Data Warehouse Toolkit*, hit the market just as the data mart craze really took off. This best-selling book provided detailed design guidance on how to optimize data for analysis.

DSS Part 2 - Data Marts (1994)

•The year 1994 featured the launch of the OLAP offensive and followed by the ROLAP counter-offensive.

•The debate was widened from relational vs. dimensional to which form of dimensional do you want: relational OLAP-based on normalized tables, a star-schema form of denormalized design or a true multidimensional database.

Data Warehouse: Divergence (1996-97)

- Enterprise warehouse versus department marts.
- Relational versus dimensional.
- OLAP versus ROLAP.
- The most universally excepted term is the operational data store (ODS)

Data Warehouse: Synthesis (1998)

•What these frameworks have in common is the inclusion of multiple layers optimized for different tasks. Each has a layer for staging, a layer for historical detail and a multifaceted layer for optimized access and delivery of data.

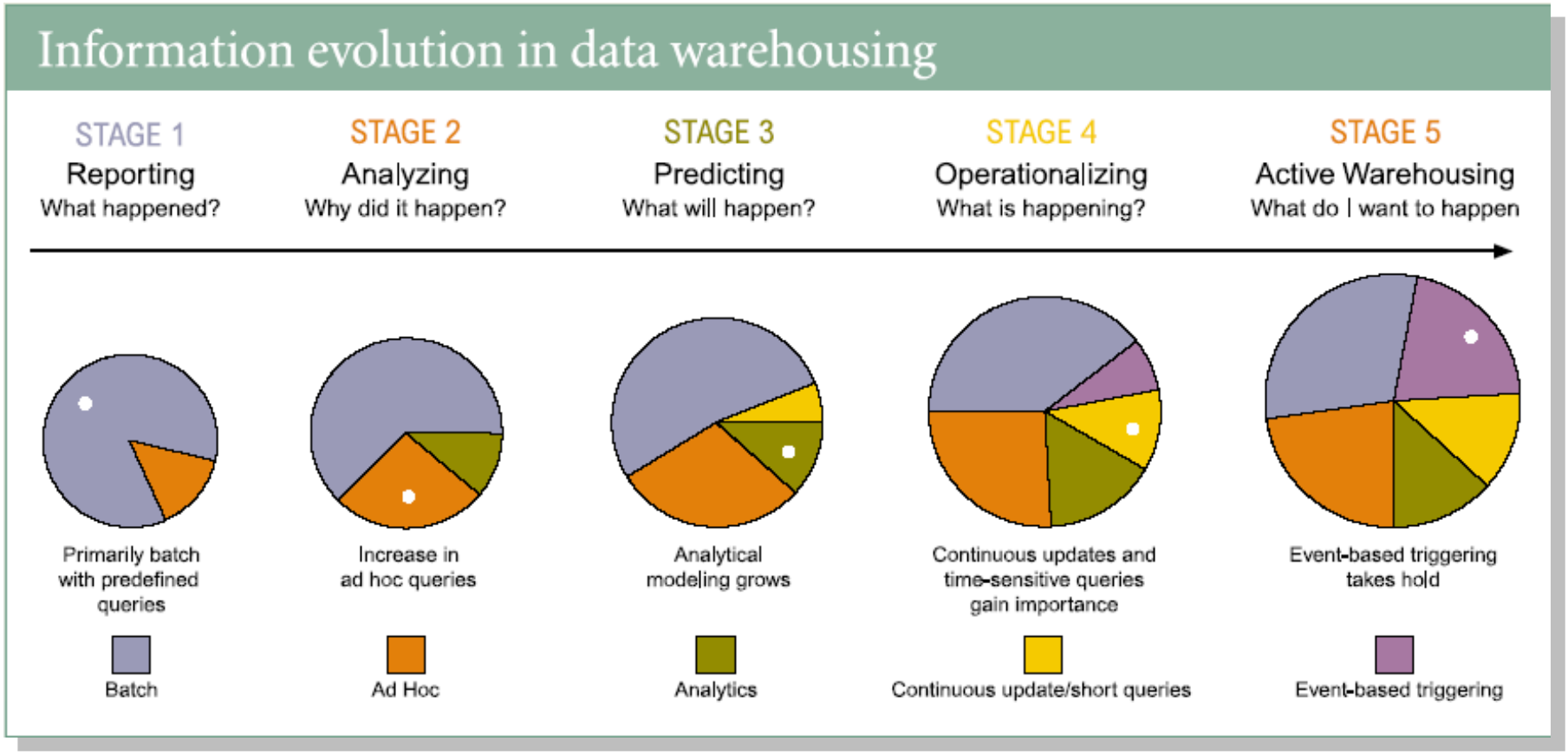
•The evolution of independent data marts to dependent data marts with common source acquisition and reusable data objects is underway.

Data Warehouse: Real-Time Data Warehousing(2000)

•The next step in the data warehouse saga is to eliminate the snapshot concept and the batch ETL mentality that has dominated since the very beginning.

数据仓库的5个发展阶段

1. 报表阶段 Reporting
2. 分析阶段 Analysis
3. 预测阶段 Prediction
4. 实时决策阶段 Operationalize
5. 主动决策阶段 Activate



数据仓库的5个发展阶段

1 报表阶段 Reporting

2 分析阶段 Analysis

3 预测阶段 Prediction

4 实时决策阶段 Operationalize

5 主动决策阶段 Activate

主要任务	主要分析	性能	挑战
• 完成数据的集中	• 定制报表 • 预先定制的查询	• 数据结构已知 •系统性能可以优化和提高	• 高质量的数据集成 •提供一致、干净的数据 •为后继阶段的数据仓库的部署提供好的数据基础

数据仓库的5个发展阶段

1. 报表阶段 Reporting
2. 分析阶段 Analysis
3. 预测阶段 Prediction
4. 实时决策阶段 Operationalize
5. 主动决策阶段 Activate

主要任务	主要分析	主要手段	挑战
• 根据问题，探察原因	• OLAP分析 • 即席查询	•切片 •上卷 •下钻	•性能问题是关键 •问题预先未知，性能优化困难 •采用索引，保证响应时间

数据仓库的5个发展阶段

1 报表阶段 Reporting

2 分析阶段 Analysis

3 预测阶段 Prediction

4 实时决策阶段 Operationalize

5 主动决策阶段 Activate

主要任务	主要分析	主要手段	挑战
•预测什么将发生	• 数据挖掘	•基于抽样数据，训练数据挖掘模型	•模型训练比较复杂 •耗费大量的系统性能

数据仓库的5个发展阶段

1. 报表阶段 Reporting
2. 分析阶段 Analysis
3. 预测阶段 Prediction
4. 实时决策阶段 Operationalize
5. 主动决策阶段 Activate

前面阶段进行战略决策

- 进行长远规划和宏观决策
- 主要基于历史数据之上，对当前实时数据不敏感
- 执行效率低，反应时间长

本阶段支持战术决策

- 在战术决策的引导下，主要侧重如何去执行
- 基于当前最新的数据基础上进行
- 决策结果需实时响应

•Operationalization typically means providing access to information for immediate decision-making in the field.[6]

数据仓库的5个发展阶段

1 报表阶段 Reporting

2 分析阶段 Analysis

3 预测阶段 Prediction

4 实时决策阶段 Operationalize

5 主动决策阶段 Activate

系统主动（自动）决策

- 系统根据用户对系统的访问自动制定战术决策
- 无需或者减少人知识的参与

实　施

- 采用时间触发的方式进行自主决策
- 结合已有的决策知识进行战术决策

Department of Computer Science and Technology, Peking University, Oct, 2006