Data Warehouse: The Choice of Inmon versus Kimball

Ian Abramson
IAS Inc.
Agenda

- The 2 Approaches
  - Bill Inmon – Enterprise Warehouse (CIF)
  - Ralph Kimball – Dimensional Design
- Similarities
- Differences
- Choices
DW History

- **1990**
  - Inmon publishes “Building the Data Warehouse”

- **1996**
  - Kimball publishes “The Data Warehouse Toolkit”

- **2002**
  - Inmon updates book and defines architecture for collection of disparate sources into detailed, time variant data store.
    - The top down approach
  - Kimball updates book and defines multiple databases called data marts that are organized by business processes, but use enterprise standard data bus.
    - The bottom-up approach
The Data Warehouse Is:

- **Bill Inmon**, an early and influential practitioner, has formally defined a data warehouse in the following terms:
  - Subject-oriented
    - The data in the database is organized so that all the data elements relating to the same real-world event or object are linked together;
  - Time-variant
    - The changes to the data in the database are tracked and recorded so that reports can be produced showing changes over time;
  - Non-volatile
    - Data in the database is never over-written or deleted - once committed, the data is static, read-only, but retained for future reporting; and
  - Integrated
    - The database contains data from most or all of an organization’s operational applications, and that this data is made consistent

- **Ralph Kimball**, a leading proponent of the dimensional approach to building data warehouses, provides a succinct definition for a data warehouse:
  - “A copy of transaction data specifically structured for query and analysis.”

Ref: wikipedia
What are they saying?

- These two influential data warehousing experts represent the current prevailing views on data warehousing.

- Kimball, in 1997, stated that
  - "...the data warehouse is nothing more than the union of all the data marts",
  - Kimball indicates a bottom-up data warehousing methodology in which individual data marts providing thin views into the organizational data could be created and later combined into a larger all-encompassing data warehouse.

- Inmon responded in 1998 by saying,
  - "You can catch all the minnows in the ocean and stack them together and they still do not make a whale,"
  - This indicates the opposing view that the data warehouse should be designed from the top-down to include all corporate data. In this methodology, data marts are created only after the complete data warehouse has been created.
What is a Data Warehouse:

- The single organizational repository of enterprise wide data across many or all lines of business and subject areas.
  - Contains massive and integrated data
  - Represents the complete organizational view of information needed to run and understand the business
What is a Data Mart?

- The specific, subject oriented, or departmental view of information from the organization. Generally these are built to satisfy user requirements for information
  - Multiple data marts for one organization
  - A data mart is built using dimensional modeling
  - More focused
  - Generally smaller, selected facts and dimensions
  - Integrated
Data Warehouses vs. Data Marts

- **Data Warehouses:**
  - **Scope**
    - Application independent
    - Centralized or Enterprise
    - Planned
  - **Data**
    - Historical, detailed, summary
    - Some denormalization
  - **Subjects**
    - Multiple subjects
  - **Source**
    - Many internal and external sources
  - **Other**
    - Flexible
    - Data oriented
    - Long life
    - Single complex structure

- **Data Marts:**
  - **Scope**
    - Specific application
    - Decentralized by group
    - Organic but may be planned
  - **Data**
    - Some history, detailed, summary
    - High denormalization
  - **Subjects**
    - Single central subject area
  - **Source**
    - Few internal and external sources
  - **Other**
    - Restrictive
    - Project oriented
    - Short life
    - Multiple simple structures that may form a complex structure
The Inmon Model

- Consists of all databases and information systems in an organization.....
  - The CIF (Corporate Information Factory)
- Defines overall database environment as:
  - Operational
  - Atomic data warehouse
  - Departmental
  - Individual
- The Warehouse is part of the bigger whole (CIF)
## The Data Warehouse

<table>
<thead>
<tr>
<th>Operational (Day-to-Day Operations) * Transactions *</th>
<th>Customer Credit Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic Data Warehouse (Data manipulated &amp; moved) * Transactions *</td>
<td>Customer Credit History</td>
</tr>
<tr>
<td>Departmental (Focused) * Source is ADW *</td>
<td>Customer by Postal Code</td>
</tr>
<tr>
<td>Individual (Ad hoc) * Source is ADW *</td>
<td>Delinquent Customers</td>
</tr>
</tbody>
</table>
Inmon Modeling

- Three levels of data modeling
  - ERD (Entity Relationship Diagram)
    - Refines entities, attributes and relationships
  - Mid-Level model (*DIS*)
    - Data Item Sets
    - Data sets by department
    - Four constructs:
      - Primary data groupings
      - Secondary data groupings
      - Connectors
      - “Type of” data
  - Physical data model
    - Optimize for performance (de-normalize)
Relationship between Levels One and Two of Inmon's Data model (Inmon, 2002)
The Warehouse Architecture
The Inmon Warehouse

Data Sources

Staging

The Data Warehouse

Data Access

Source DB 1

Source DB 2

Landing Staging Area

Data Marts

Cubes

File or External Data
The Kimball Approach

The Dimensional Data Model

- Starts with tables
  - Facts
  - Dimensions
- Facts contain metrics
- Dimensions contain attributes
  - May contain repeating groups
- Does not adhere to normalization theory
- User accessible
The Kimball Data Lifecycle

Data Sources
Source DB 1
Source DB 2

File or External Data

Staging
Landing Staging Area

The Data Warehouse

Data Access
Workstation Group
End Users
Cubes

End Users
The Dimensional Model
The Kimball Data Bus

- Data is moved to staging area
  - Data is scrubbed and made consistent
- From Staging Data Marts are created
- Data Marts are based on a single process
- Sum of the data marts can constitute an Enterprise Data Warehouse
- Conformed dimensions are the key to success
The Kimball Design Approach

- Select business process
- Declare the grain
- Choose dimensions
- Identify facts (metrics)
Kimball’s Philosophy

- Make data easily accessible
- Present the organization’s information consistently
- Be adaptive and resilient to change
- Protect information
- Service as the foundation for improved decision making.
Getting Started with Choices

- Kimball
  - Will start with data marts
  - Focused on quick delivery to users

- Inmon
  - Will focus on the enterprise
  - Organizational focus
Kimball vs. Inmon

- Inmon:
  - Subject-Oriented
  - Integrated
  - Non-Volatile
  - Time-Variant
  - Top-Down
  - Integration Achieved via an Assumed Enterprise Data Model
  - Characterizes Data marts as Aggregates

- Kimball
  - Business-Process-Oriented
  - Bottom-Up and Evolutionary
  - Stresses Dimensional Model, Not E-R
  - Integration Achieved via Conformed Dimensions
  - Star Schemas Enforce Query Semantics
The Comparison (Methodology and Architecture)

<table>
<thead>
<tr>
<th></th>
<th>Inmon</th>
<th>Kimball</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall approach</td>
<td>Top-down</td>
<td>Bottom-up</td>
</tr>
<tr>
<td>Architectural structure</td>
<td>Enterprise-wide DW</td>
<td>Data marts model a business process; enterprise is achieved with conformed dims</td>
</tr>
<tr>
<td></td>
<td>feeds departmental DBs</td>
<td></td>
</tr>
<tr>
<td>Complexity of method</td>
<td>Quite complex</td>
<td>Fairly simple</td>
</tr>
</tbody>
</table>

# The Comparison (Data Modeling)

<table>
<thead>
<tr>
<th></th>
<th>Inmon</th>
<th>Kimball</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data orientation</strong></td>
<td>Subject or data driven</td>
<td>Process oriented</td>
</tr>
<tr>
<td><strong>Tools</strong></td>
<td>Traditional (ERDs and DIS)</td>
<td>Dimensional modeling; departs from traditional relational modeling</td>
</tr>
<tr>
<td><strong>End user accessibility</strong></td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

# The Comparison (Dimensions)

<table>
<thead>
<tr>
<th></th>
<th>Inmon</th>
<th>Kimball</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Timeframe</strong></td>
<td>Continuous &amp; Discrete</td>
<td>Slowly Changing</td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td>Timestamps</td>
<td>Dimension keys</td>
</tr>
</tbody>
</table>

Inmon Continuous & Discrete Dimension Management

- Define data management via dates in your data
  - Continuous time
    - When is a record active
    - Start and end dates
  - Discrete time
    - A point in time
    - Snapshot
Kimball Slowly Changing Dimension Management

- Define data management via versioning
  - Type I
    - Change record as required
    - No History
  - Type II
    - Manage all changes
    - History is recorded
  - Type III
    - Some history is parallel
    - Limit to defined history
# The Comparison (Philosophy)

<table>
<thead>
<tr>
<th></th>
<th>Inmon</th>
<th>Kimball</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary Audience</strong></td>
<td>IT</td>
<td>End Users</td>
</tr>
<tr>
<td><strong>Place in the Organization</strong></td>
<td>Integral part of the Corporate Information Factory (CIF)</td>
<td>Transformer and retainer of operational data</td>
</tr>
<tr>
<td><strong>Objective</strong></td>
<td>Deliver a sound technical solution based on proven methods</td>
<td>Deliver a solution that makes it easy for end users to directly query data and still have reasonable response rate</td>
</tr>
</tbody>
</table>

How to Choose?

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Favors Kimball</th>
<th>Favors Inmon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature of the organization's decision support</td>
<td>Tactical</td>
<td>Strategic</td>
</tr>
<tr>
<td>requirements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data integration requirements</td>
<td>Individual business areas</td>
<td>Enterprise-wide integration</td>
</tr>
<tr>
<td>Structure of data</td>
<td>Business metrics, performance measures, and scorecards</td>
<td>Non-metric data and for data that will be applied to meet multiple and varied information needs</td>
</tr>
<tr>
<td>Scalability</td>
<td>Need to adapt to highly volatile needs within a limited scope</td>
<td>Growing scope and changing requirements are critical</td>
</tr>
</tbody>
</table>
How to Choose?

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Favors Kimball</th>
<th>Favors Inmon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistency of data</td>
<td>Source systems are relatively stable</td>
<td>High rate of change from source systems</td>
</tr>
<tr>
<td>Staffing and skills requirements</td>
<td>Small teams of generalists</td>
<td>Larger team(s) of specialists</td>
</tr>
<tr>
<td>Time to delivery</td>
<td>Need for the first data warehouse application is urgent</td>
<td>Organization's requirements allow for longer start-up time</td>
</tr>
<tr>
<td>Cost to deploy</td>
<td>Lower start-up costs, with each subsequent project costing about the same</td>
<td>Higher start-up costs, with lower subsequent project development costs</td>
</tr>
</tbody>
</table>
References

- The Data Warehouse Toolkit, Kimball, 2002
Thanks and Questions?

- Ian Abramson
  IAS Inc
  416-407-2448
  ian@abramson.ca