Building an Efficient Data Warehouse: summarisation and partitioning

Scott Anderson
SAS Europe
Topics

- Steps in Building the DW
- Data modelling for the DW
- Alternative data models for the DW
- Sizing considerations
- Data summarisation
- Data partitioning
- Conclusions
The SAS Data Warehouse

Metadata

Management

Operational

Data Extraction

Transformation Engine

Loader

Scheduler

Metadata Manager

Organisation

RDBMS

Legacy

SAS

External

Exploitation

Quality

Risk

Customer

Product

Market

Future

Information Database

EIS

OO RAD

Query

C/S Visualise

DSS OLAP++

Open

Copyright 1995 by SAS Institute
Steps in Building the Data Warehouse

- Data Subject Definition
- **Aggregation Levels**
- Data Acquisition
- Data Transformation
- **Data Partitioning (physical separation)**
- Meta Data Creation and Management
- Production Loading/Scheduling
- Archiving
- Exploitation (new and/or existing apps & tools)
Steps in Building the Data Warehouse

- Data Subject Definition
- Aggregation Levels
- Data Acquisition
- Data Transformation
- Data Partitioning (physical separation)
- Meta Data Creation and Management
- Production Loading/Scheduling
- Archiving
- Exploitation (new and/or existing apps & tools)
Data Subjects

- May be:
  - customer, product, sales, ...
  - derived from multiple (OLTP) data sources
  - at varying levels of detail

- will contain:
  - metadata
  - value data
  - descriptive data
  - lookup, translation data
Process vs Subject-Oriented

Entry
- Sales Rep
- Quantity Sold
- Part Number
- Date
- Customer Name
- Product
  - Description
- Unit Price
- Mail Address

Sales

Customers

Products

Transactional Storage

Warehouse Storage
‘The starting point for the design and development of the data warehouse is the data model’

Understand your data
The ‘Spider’s Web’
Data Modelling - definitions?

- the process of matching business requirements to the data needed to support them. To be considered:
  - business requirements
  - technology requirements
- a methodological approach to analysing the data resource
Data Modelling - DBMS Design

- **Logical modelling**
  - entity-relationship analysis
  - TNF (Third Normal Form) analysis
  - ‘top-down vs bottom-up’

- **Physical modelling**
  - transformation of logical, or ‘conceptual’ model to physical DBMS structure eg IMS, IDMS, DB2, etc.
Data Modelling - DBMS Design

- **logical modelling**
  - entity-relationship analysis
  - TNF (Third Normal Form) analysis
  - ‘top-down vs bottom-up’

- **physical modelling**
  - transformation of logical, or ‘conceptual’ model to physical DBMS structure eg IMS, IDMS, DB2, etc.
Data Modelling for the Data Warehouse

- **logical modelling**
  - high level data analysis
- **physical modelling**
  - mapping of data to structure of physical DBMS
- **transformation modelling**
  - mapping of operational to DW data type
Rapid Warehousing Methodology

1) Justification
2) Requirements Gathering
3) Design/Modelling
4) Implementation
5) Review
X) Project Management
Evolutionary Project Cycles

- Start small / focused with measurable Results
- Reuse positive Experience to extend
- Ultimate Goal: Corporate Warehousing Usage
Requirements-Driven Data Modelling

Requirements

JAD

OLTP data model(s)

External data source(s)

Transformation Model

Logical Business Rules

Data Warehouse data model

Logical data model

METADATA
Data Modelling: OLTP v Data Warehouse

Logical data model

Operational data model

Data Warehouse data model
Data Modelling for OLTP

- tuned for performance
- file system dependent
- for RDBMS:
  - normalised
  - SQL dependent
  - referential integrity constraints
Data Modelling for DW

- creates data of business value only
- adds time dimension
- adds derived data
- subject area grouping (denormalisation)
- prepares for information delivery
Physical Data Models for DW

- dependent upon data storage engine, and tools to be used
- alternative models
  - relational model
  - star schema
  - snowflake schema
  - multidimensional model
  - multi-level summaries
  - ....
- may be mix ‘n match
Entity-Relationship Model

- Divides entities across a large number of small tables
- Seeks to drive all redundancy from the system
- Provides excellent data integrity for transactions
- Need only ‘touch’ the database in one place
Star Schema

- aka dimensional modelling
Star Schema

- Uses a more ‘asymmetrical’ table relationship model
- Contains a highly normalized, large central ‘fact’ table of columns that are additive
- Contains smaller, dimension oriented tables
- which contain dimension descriptions
- Dimension tables contain a primary key
- Fact table contains a composite, foreign key
Star Schema

Time Dimension
- time_key
- day_of_week
- month
- quarter
- year
- week-end_flag
- etc....

Sales Facts
- time_key
- product_key
- amount_sold
- units_sold
- dollar_cost
- other facts....

Product Dimension
- product_key
- description
- brand
- category
- department
- vendor
- vendor
- etc....
Snowflake Schema
Snowflake Schema

- Similar to a star-schema
- Contains a more normalized dimension table
- Reduces disk overhead for the dimension table by employing an ‘out-rigger’ table
- Primary/foreign key relationship is between dimension table & out-rigger table
Multidimensional Model

- Geography
- Products
- Time
Multidimensional Model (MDDB)

- Create and store permanent N-way crossings similar to a fact table
- (Multiple) values held in cells
- Only valid crossings are stored, eliminating ‘sparsity’
- Build time and space requirements minimised ‘drip feed’
- Some summary statistics are computed ‘on the fly’ rather than stored
- the SAS Implementation of MDDB will allow for incremental update
Multidimensional Model (MDDB)

- Packaged OLAP Solution
- **Fast**, Multi-Dimensional Access to data
- GUI Browsers based on Templates
- Large Volumes - two examples
  - Portable PC based - 4.1 Million Cells - 400 Mb
  - UNIX Server based - 28 Million Cells - 1.7 Gb
- Response times under 5 seconds
Multi-Level Summarisation

Product Hierarchy

- Brand
- Product Line
- Product
DW Sizing Considerations

- level of detail required
- access requirements (frequency)
- storage requirement
- performance requirements
- drill-down requirements

Level of Granularity
Enterprise Wide Information Delivery - data summarisation

Directors, Executives

Middle Management

Business Analysts, Specialised Professionals

Operational Users

ACTION

KNOWLEDGE

INFORMATION

DATA
Summary Tables - design considerations

- Takes data at a granular level and ‘rolls’ it up into a more compact form (data density, space)
- Lower overhead requirements than detail level data (performance)
- Higher-level summaries result in ‘loss’ of data detail (recalculation)
- Pre-storing pertinent summaries can reduce demand on end-user resources (usage patterns)
SAS Summary Statistics

- mean, median, and mode
- counts, missing values
- sum, weighted sum
- standard deviation and variance
- max, min, and range
- percentile
- ....
Summarisation - example

Category

Articles (40K)

Weekly Sales
(2 years - 9 mill. records)

45 min query

20K sales/week
Article desc.:
(2 langs x 60 chars)

extra 240 MB if denormalised
Summarisation - example (Contd.)

< 20 sec query

Denormalised YTD Sales Figures (40K records)

no joins
no ad-hoc summarisation

Article desc.:
(2 langs x 60 chars)
extra 5 MB
Summarisation - example (Contd.)

Denormalised YTD Sales Figures (40K records)

< 20 sec query

no joins
no ad-hoc summarisation

Article desc.:
(2 langs x 60 chars)
extra 5 MB

Performance v Space
 Ease of use = business benefit
Loading the Warehouse

Transformation Engine

Metadata

Event Data

Summary
Data Warehouse Architecture

- Highly Summarised
- Lightly Summarised
- Event Data / Summarised Data
- Aged Data
Data Warehouse Usage
Distributed Data Warehouse

- Workstation (Department)
- Client / Server (Countries)
- Mainframe / Unix Server (Headquarters)
- Bulk Storage
Multidimensional Data Marts
Data Partitioning

- time periods
- geographic locations
- lines of business
- organisational units
- platform partitioning (MVA)
- table partitioning
  - horizontal ie. by row
  - vertical ie. by column (stability analysis)
- ...

Data Partitioning - SAS Strengths

- date/time handling
- dataset expression handling (where... processing)
- data merging
- flexible indexing
- flexible data storage
  - data structure
  - indexing
  - alternative physical data models
  - alternative platforms (MVA)
Data Warehousing - current issues

- packaged DW management solutions
- MOLAP & ROLAP
- data marts
- data mining
- internet/intranet access to DW
- data server
- parallel processing
- DW solutions for specific business processes
The SAS Data Warehouse - new developments

- packaged DW management solutions (SAS/Warehouse Administrator)
- MOLAP & ROLAP (MDDB / Motore)
- data marts
- data mining (eg. Neural Networks)
- internet/intranet access to DW
- data server: Orlando Server
- parallel processing (SPDS)
- DW solutions for specific business processes (eg. Risk management, CFO/Vision)
The SAS/Warehouse Administrator
SAS/Warehouse Administrator

- Aimed at the Manage and Organise Activities
- Used primarily by IT staff (DWA)

**Functions:**
- Subject Definition
- Data source organiser
- Scheduler (load & distribution)
- Metadata Browsers
- Management of Multiple Warehouses - test/production

- Beta Release by SEUGI 96
Conclusions

- understand your data
- optimise the efficiency of your DW solutions using strengths of SAS data model / data manipulation tools
- techniques providing optimised tool performance
**References**