Data Vault Modeling
The Next Generation DW Approach

Presented by: Siva Govindarajan

HTTP://WWW.GLOBYTES.COM
SIVA02@GLOBYTES.COM
Data Vault Modeling

Agenda

- Introduction
- Data Vault Place in Evolution
- Data Warehousing Architecture
- Data Vault Components
- Case Study
  - Typical 3NF / Star Schema
  - Data Vault Approach
  - Temporal Data + Auditability
- Implementation of Data Vault
- Conclusion
Introduction

- Data Vault concept originally conceived by Dan Linstedt
- Enterprise Data Warehouse Modeling approach.
- Hybrid approach - 3NF & Star Schema
- Adaptable to changes
- Auditable
- Timed right with Technology
Data Vault Place in Evolution

- 1960s - Codd, Date et. al Normal Forms
- 1980s - Normal Forms adapted to DWs
- 1985+ - Star Schema for OLAP
- 1990s - Data Vault concept developed Dan Linstedt
- 2000+ - Data Vault concept published by Dan Linstedt
Data Warehousing Architecture

*Bill Inmon’s Data Warehouse Architecture*

*Kimball’s Data Warehouse Bus Architecture*
Data Vault Components

- **Hub Entities**
  - Candidate Keys + Load Time + Source

- **Link Entities**
  - FKs from Hub + Load Time + Source

- **Satellite Entities**
  - Descriptive Data + Load Time + Source + End Time
    - Dimension = Hub + Satellite
    - Fact = Satellite + Link [+ Hub]
Case Study

- Modeling Approaches
  - Typical 3NF / Star Schema
  - Data Vault approach
- Scenario
  - Typical Fact and Dimension
  - New Dimension data
  - Reference Data change - 1:M to M:M
  - Additional attributes to Dimension
Typical 3NF / Star Schema

1. Typical Fact and Dimension
Typical 3NF / Star Schema

2. New Dimension Data
Typical 3NF / Star Schema

2. New Dimension Data
Typical 3NF / Star Schema

3. Reference 1:M to M:M Change
Typical 3NF / Star Schema

4. Additional attributes to Dimension
Data Vault Approach

1. Typical Hub, Satellite and Link
   (Replacing Fact/Dimension)
Data Vault Approach

2. New Dimension Data
Data Vault Approach

3. Reference 1:M to M:M Change

- **HUB_PRODUCT**
  - hub_product_id
  - hub_record_source
  - hub_load_date
  - product_id
  - product_code

- **SAT_PRODUCT01**
  - sat_product01_id
  - hub_product_id
  - sat_load_date
  - sat_end_date
  - sat_record_source
  - product_desc
  - product_name

- **HUB_ORDER**
  - hub_order_id
  - hub_record_source
  - hub_load_date
  - tr_order_id
  - system_order_id
  - system_root_order_id
  - global_root_order_id

- **SAT_ORDER01**
  - sat_order01_id
  - hub_order_id
  - sat_load_date
  - sat_end_date
  - sat_record_source
  - order_quantity
  - total_executed_qty

- **HUB_CATEGORY**
  - hub_category_id
  - hub_record_source
  - hub_load_date
  - product_category_id
  - product_category_code

- **SAT_CATEGORY01**
  - sat_category01_id
  - hub_category_id
  - sat_load_date
  - sat_end_date
  - sat_record_source
  - product_category_name
  - product_category_desc
  - is_strategy
Data Vault Approach

4. Additional attributes to Dimension
Temporal Data + Auditability

HUB_PRODUCT
- hub_product_id
- hub_record_source
- hub_load_date
- product_id
- product_code

SAT_PRODUCT01
- sat_product01_id
- hub_product_id
- sat_load_date
- sat_end_date
- sat_record_source
- product_desc
- product_name

HUB_CATEGORY
- hub_category_id
- hub_record_source
- hub_load_date
- product_category_id
- product_category_code

SAT_CATEGORY01
- sat_category01_id
- hub_category_id
- sat_load_date
- sat_end_date
- sat_record_source
- product_category_name
- product_category_desc
- is_strategy

HUB_ORDER
- hub_order_id
- hub_record_source
- hub_load_date
- t_order_id
- system_order_id
- system_root_order_id
- global_root_order_id

SAT_ORDER01
- sat_order01_id
- hub_order_id
- sat_load_date
- sat_end_date
- sat_record_source
- order_quantity
- total_executed_qty
- event_timestamp

LNK_PRODUCT_X_CATEGORY
- link_product_x_category_id
- link_load_date
- link_record_source
- hub_product_id
- hub_category_id

LNK_ORDER
- link_order_id
- link_load_date
- link_record_source
- hub_order_id
- hub_product_id
- hub_category_id
Implementation of Data Vault

- Make it Simple
  - View for Dimension
  - View For Fact
- Data Loading
- Take advantage of Technology
  - DW Appliances
  - High-throughput storage devices
  - RDBMS Features
Make it Simple

- **Create views for Dimension**

From our Demo model, Join Product related Hubs and Satellites to build View for Product Dimension as shown below:

- HUB_PRODUCT
- SAT_PRODUCT01
- SAT_PRODUCT02
Make it Simple

- Create Views For Fact

Join Order related Hubs, Satellites and Links to build View for ORDER Fact using ORDER related Hubs/Satellites and Links.

- HUB_ORDER
- SAT_ORDER01
- LNK_ORDER
Data Loading

Typically data loading for a Data Vault is in the following sequence.

- Hubs for Dimensions
- Links for Dimensions
- Satellites for Dimensions
- Hubs for Fact (if any)
- Links for Fact
- Satellites for Fact

Refer to Data Vault Series article 5 of Linstedt for further details
Next wave of IT Data Warehouse hardware solutions are the Data Warehouse Appliances. Take advantage of the technology where possible. Few DW Appliances in the market:

- Oracle Exadata
- Netezza
- Teradata
- RedBrick
- GreenPlum
High-throughput storage devices

Utilize Storage devices designed for DW / OLAP applications. Following are some examples only and would change over time. Please do your research based on your sizing requirements.

- EMC CLARiiON Storage family
- Texas Memory systems RamSan – Solid State devices
- Hitachi USP / VSP Storage family
- Other Hybrid solutions with High performance storage and Solid State devices
Some Oracle Features catering DW / OLAP applications:

- Exadata Smart Scans
- OLAP Based Materialized views
- Partitioning Reference partition
- Advanced data compression
- Automatic degree of parallelism (ADOP).
- Star query optimization
- Oracle OLAP/DWH Features.
Conclusion

In this presentation we have addressed high level overview of Data Vault Architecture. Topics discussed are

- Data Vault concept and architecture
- Data Vault Components such as Hubs, Satellites and Link tables
- Typical modeling challenges with traditional modeling approaches
- How those challenges could be handled using Data Vault Modeling Approach.
- Auditing and Temporal data capture using DV Approach.
- And finally, some implementation details

If interested in learning more, try
Dan Linstedt's special coaching area at:
Questions ?
References

- Data Vault Series articles by Dan E. Linstedt -- http://www.tdan.com
- Referred few articles from Genesee Academy -- http://geneseeacademy.com
- Articles and books by Ralph Kimball and Bill Inmon from multiple sources.
- Technical Documents from http://www.oracle.com/technetwork
Special Thanks to

- Dan Linstedt for review and corrections to the article.
- My colleague Mark Bruscke for his assistance in preparing the article.
The End

Contact Email:
siva02@globytes.com