Data Vault 2.0 – A balanced approach to modelling data warehouse

**Why do we need change data modelling approach for data warehouse?**

We are living the phase where technologies are driving the process of the business. Business is more focused not only solving the process related issues using technologies but optimize and digitize most of the manual tasks. Thanks to technologies like AI, ML, Big Data etc. that are proving to be catalyst to this revolutionising journey. Business has also realized that digitization is the way forward and are not hesitating to experiment with new technologies and new process. The world is changing at fast rate so is the process and so is the technology. This has shaken the data management teams in organization at the core. The approach adopted by organization since last five decades are not helping to hold, manage and process data to support this dynamically changing environment. Lots of effort are going on to maintain the historical data and effort to merge new data coming in from rapidly changing environment adopted to support business. Is there a more efficient way to store, manage and process data for business use?

Yes, Data Architects is also trying to keep up the pace with the demands of data storage and processing. The approach that we are going to talk about is Data Vault, data modelling method based on mathematical principles devised by Dan Linstedt to solve the current challenges with 3-NF and de-normalised modelling for EDW. The modelling approach is efficient enough to accommodate data generated by both process and technology/platform changes.

**Mapping data vault modelling technique to Simon Sinek’s famous golden circle.**

**WHY:** Better approach to tackle needs of future DW demands and fix the challenges in current setup

The business is changing, the source systems are changing and adding, and advancement of Big data technologies wants to pump data into the data warehouses. The current DW modelling methods are somewhat not robust enough to support that. Any new source changes/introduction is having tsunami impact on the current DW infrastructure for storage and process. 3-NF approach have following challenges with tides are high; time depended primary key leading to parent-child complexities, difficulties in near real time data loading, complex query formation for analytical needs, challenges in scalability. All these challenges are because the approach of 3-NF is technology focused approach. On the other hand, with denormalized conformed data mart, we face difficulties with subject oriented, data redundancy, synchronization issues. One of the major challenges, we face with granularity issues as conformed data mart inherits no robustness to adopt to changes in the granularity requirements. In nutshell, the 3-NF approach is technology platform focused while conformed data mart is business focused.

**HOW:** Devise data modelling approach to resolve issues with traditional approach (3-NF and Denormalized)

The approach was devised by Dan Linstedt that was focused on approach that takes best of 2 world and club it one. The focus was not solving the concerns of traditional method instead a positive approach to take best of both world and combine it. This led to an approach that will help both technology and business teams of the organization.

**WHAT:** What are the components if Data Vault

There are 3 foundational entities to Data Vault namely HUB, LINKS and SATELLITE. The approach is devised focused on functional areas of business with HUB representing the primary key. LINK entities provide the transactions integration between the HUBS while SATELLITE entities support with the contextual information around the HUBS.

HUBS: Hub entities are single table carrying at minimum a unique list of business keys. These are the keys that the businesses utilize in everyday operations. For example, invoice number, employee number, customer number, part number etc. This makes sure the approach must be business function focused. The other attribute in HUB includes.

* Surrogate Key (Optional).
* Load Date Time Stamp: to record the key landing date time in the warehouse.
* Record source: To track the source system from which the key was loaded to help data traceability.

For example, the requirement is to capture customer number across the company. Accounting may have a customer number (12345) represented in a numeric style and contracts may have the same customer  
number prefixed with an alpha (AC12345).

LINKS: Link entities is the representation of many to many relationships (as in 3-NF). The links represents the transaction of relationship between the business components. The LINKS contains the following attributes,

* Surrogate key (optional)
* HUB 1 key to HUB N key: The keys to migrated into the LINKS that represents the relationship between the HUBS.
* Load date time stamp: recording the relation/transaction when it was created for the first time.
* Record Source: The source system that defined the relation/transaction.

SATELLITE ENTITIES: Satellite entities are the entities that gives descriptive context to the HUBS. All the information that is subject to change over a time. This makes model robust enough to capture an store the altered values at most granular level. The satellite holds the following attributes,

* Satellite Primary Key: Hub primary key migrated to satellite from hub or a link.
* Satellite Primary Key: Load date time stamp recording when the context available in the warehouse (The new row is recorded SCD 2)
* Satellite Optional Primary Key: Sequence Surrogate Number – utilized for Satellites that have multiple values (such as a billing and home address), or line-item numbers, used to keep the Satellites sub-grouped and in order.
* Record Source – The source system that defined the relation/transaction.

**Data Vault’s data conceptual representation and few considerations**

Based on the above discussion**,** the conceptual model looks as show below. HUBS acting as centre view driving the business-critical entities and LINKS defining how are HUBS related or any transaction conducted between the HUBS.

Diagram

Description automatically generated

There is a standard step that has been defined to come up with the conceptual Data Vault model. The steps are as follows,

* Model the Hubs. This requires an understanding of business keys and their usage across the designated scope.
* Model the Links. Forming the relationships between the keys – formulating an understanding of how the business operates today in context to each business key.
* Model the Satellites. Providing context to each of the business keys as well as the transactions (Links) that connect the Hubs together. This begins to provide the complete picture of the  
  business.
* Model the point-in-time tables. This is a Satellite derivative, of which the structure and definition is outside the scope of this document (due to space constraints).

Reference rules to follow while building Data Vaults,

* Hub keys cannot migrate into other Hubs (no parent/child like Hubs). To model in this manner breaks the flexibility and extensibility of the Data Vault modelling technique.
* Hubs must be connected through Links.
* More than two Hubs can be connected through Links.
* Links can be connected to other Links.
* Links must have at least two Hubs associated with them to be instantiated.
* Surrogate keys may be utilized for Hubs and Links.
* Surrogate keys may not be utilized for Satellites.
* Hub keys always migrate outward.
* Hub business keys never change, Hubs primary keys never change.
* Satellites may be connected to Hubs or Links.
* Satellites always contain either a load date-time stamp, or a numeric reference to a stand-alone load date-time stamp sequence table.
* Stand-alone tables such as calendars, time, code, and description tables may be utilized.
* Links may have a surrogate key.
* If a hub has two or more satellites, a point-in-time table may be constructed for ease of joins.
* Satellites are always change driven; duplicate rows should not appear.
* Data is separated into Satellite structures based on:
  + type of information
  + rate of change.

**USEREADY’s approach**

We at Useready are helping our customer to adopt the Data Vault modelling in their data warehouses to leverage the advantage and making them future ready. We are helping customer to evaluate and implement Data Vault on Cloud as well as on Prem. Also, most of our work is migrating away from 3-NF and Denormalized EDW setup to Data Vault. The implementation is agile to leverage the Data Vault as earliest.

Specifically, with snowflake we are helping customers to implement Data Vault using their prescribed reference architecture. This enables the implementations with best practices.

Diagram

Description automatically generated

The architecture is not only aligned to Data Vault 1.0 principles but also covers Data Vault 2.0 supporting Agile approach to entire project execution.

I hope you enjoyed reading the blog. This is the first in the series talking about 30000 ft view at data vault. Stay tuned for more detailed readings around Data Vault.