Data Vault

- Detail oriented, historical tracking, and uniquely linked set of normalized tables that support one or more functional areas of business
- 3rd normal form + Star schema
- Based on mathematical principles that support the normalized data models

Problems of existing Data Warehouse Data Modeling Architectures

- 3NF
  - Date time stamp placed as a primary key
- DataMart
  - Aggregation must stay consistent for all time (during the life of the relationship). The structure of each fact table must not change (i.e., no new dimensions will be added to either fact table). It limits design, scalability, and flexibility of the data model.

Importance of Architecture and Design for Enterprise Data warehousing

A data warehouse should be:

- Top-down architecture
- Bottom-up implementation

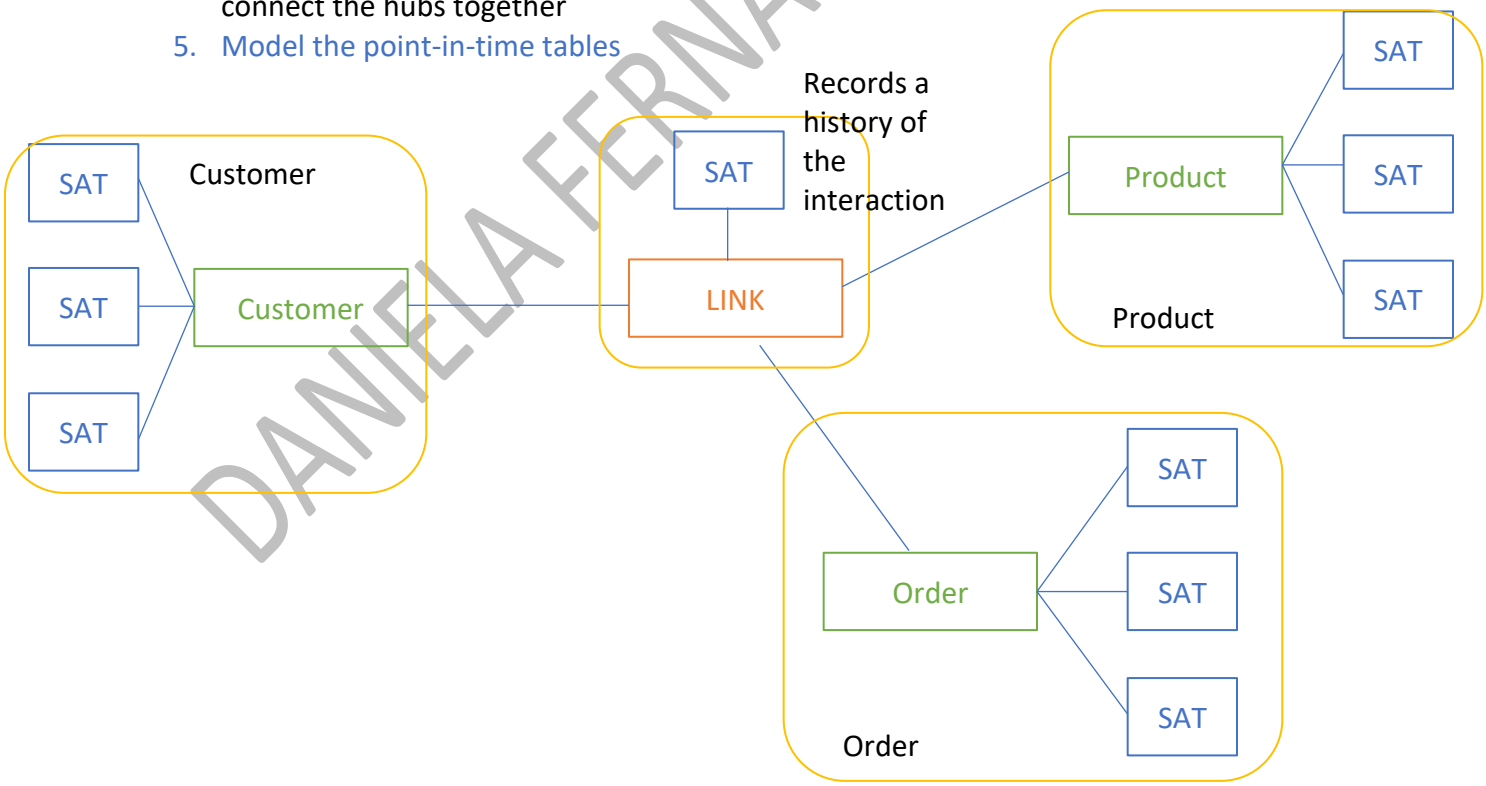
Data vault components:

- Hub Entities
  - A single table carrying at a minimum a unique list of business keys. These keys are used by business in everyday operations.
    - Surrogate key (optional) – a smart key or sequential number
    - Load Date Time stamp – recording when the key itself first arrived in the warehouse
    - Record Source – a recording of the source system utilized for data traceability
- Link Entities
  - It represents the relationship or transaction between two or more business components (two or more business keys)
    - Surrogate key (optional) – possibly a smart key or a sequential number. Only utilized if there are more than two hubs through this link, or the composite primary key might cause performance problems
    - Hub 1 key to hub N key – hub keys migrated into Link to represent the composite key or relationship between two hubs
    - Load date time stamp – recording when the relationship/transaction was first created in the warehouse

        - Record Source – a recording of the source system utilized for data traceability
- Satellite Entities
    - Hub key context (description) information
        - Satellite Primary Key
            - Hub or Link Primary key – Migrated into the satellite from the Hub or Link
            - Load Date time stamp – Recording when context information is available in the warehouse (the new row is always inserted)
        - Satellite Optional Primary Key
            - Sequence Surrogate number – utilized for satellites that have multiple values, or line items numbers, used to keep the satellites sub-grouped and in order
            - Record source – A recording of the source system utilized for data traceability

Building a data vault

1. Model the Hubs
    a. Understand business keys and usage across the designated scope
2. Model the Links
    a. Forming the relationships between the keys (how business operates)
3. Model the satellites
4. Providing context to each of the business keys as well as the transactions (links) that connect the hubs together
5. Model the point-in-time tables



HUBS = Unique Business keys
LINKS = Relationships and associations
SATELLITES = descriptive data

Reference
https://tdan.com/data-vault-series-1-data-vault-overview/5054