DATA VAULT MODELING GUIDE

Introductory Guide to Data Vault Modeling

GENESEE ACADEMY, LLC

2012

Authored by: Hans Hultgren

DATA VAULT MODELING GUIDE

Introductory Guide to Data Vault Modeling

Forward

Data Vault modeling is most compelling when applied to an enterprise data warehouse program (EDW). Several key decisions concerning the type of program, related projects, and the scope of the broader initiative are then answered by this designation. In short, the organization contemplating this initiative is committing to an integrated, non-volatile, time variant and business key driven data warehouse program.

The data vault principles are specifically well suited for such a program and – when applied consistently – can provide the organization with some very compelling benefits. These include auditability, agility, adaptability, alignment with the business, and support for operational data warehousing initiatives.

To gain these benefits however, the organization will need to commit to both EDW program level factors as well as specific data vault modeling patterns, rules and methods. This guide presents data vault modeling in the context of the EDW.

Index

The EDW Program

The Enterprise Data Warehousing (EDW) Program represents the ongoing data warehousing activities of the organization. These activities will include the maintenance functions of the data warehouse in addition to the continuous flow of incremental projects related to the enterprise data warehouse. These incremental projects are comprised of

a) Adapting to new data sources from internal new systems, external integrations, and from acquisitions, and
b) Absorbing changes to existing sources including new tables, new attributes, new domain values, new formats and new rules, and
c) Adapting to new business rules concerning the alignment, grain, cardinality and domain values of business keys as well as changes to the relationships between them, and
d) Accommodating new downstream delivery requirements including new subject areas, new business rules, additional regulatory and other compliance reporting and changes to operational latency requirements.

For this reason, the EDW itself is not designated a "project" (there is no discernable beginning and end, and no pre-determined set of specific goals).

In a broader sense, this program can be defined as the BI Function or BI Program within an organization. To be clear however, this is not simply the group that owns the OLAP tools. This is the higher level view of all data warehousing and business intelligence (DWBI) within the organization which includes the business intelligence competency center or BICC, the EDW or CDW team, the related governance components and the environment both technical and organizational. The success of a DWBI program depends on an organizational commitment and a corporate BI culture.

It is precisely in this context where the data vault approach is the most valuable. So the Data Vault EDW is defined first and foremost by the enterprise wide, long term DWBI program – from a technical architecture perspective and from an organizational cultural alignment perspective as well.

The Data Vault Fundamentals

The data vault consists of three core components, the **Hub, Link** and **Satellite**. Above all other DV Program rules and factors, the commitment to the consistency and integrity of these constructs is paramount to a successful DV Program.

The **Hub** represents a Core Business Concept such as Customer, Vendor, Sale or Product. The Hub table is formed around the Business Key of this concept and is established the first time a new instance of that business key is introduced to the EDW. It may require a multiple part key to assure an enterprise wide unique key however the cardinality of the Hub must be 1:1 with a single instance of the business concept. The Hub contains no descriptive information and contains no FKs. The Hub consists of the business key only, with a warehouse machine sequence id, a load date/time stamp and a record source.

Fig. 1 ***Hub***

A **Link** represents a natural business relationships between business keys and is established the first time this new unique association is presented to the EDW. It can represent an association between several Hubs and sometimes other Links. It does maintain a 1:1 relationship with the unique and specific business defined association between that set of keys. Just like the Hub, it contains no descriptive information. The Link consists of the sequence ids from the Hubs and Links that it is relating only, with a warehouse machine sequence id, a load date/time stamp and a record source.

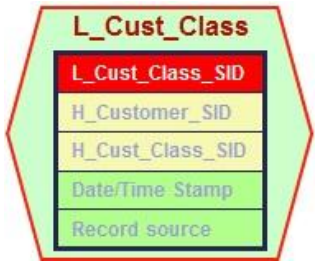

Fig. 2 ***Link***

Notice the similarity between the Hub and the Link. Both represent the first time a core business concept (Hub) or natural business relationship (Link) is introduced to the DW.

The **Satellite** contains the descriptive information (context) for a business key. There can be several Satellites used to describe a single business key (or association of keys) however a Satellite can only describe one key (Hub or a Link). There is a good amount of flexibility afforded the modelers in how they design and build Satellites. Common approaches include using the subject area, rate of change, source system, or type of data to split out context and design the Satellites. The Satellite is keyed by the sequence id from the Hub or Link to which it is attached plus the date/time stamp to form a two part key.

Note that the Satellite then is the only construct that manages time slice data (data warehouse historical tracking of values over time).

Fig. 3 ***Satellite***

A Satellite does not have a Sequence ID of its own and in fact cannot have a different key than the Hub or Link sequence to which it is attached. Further, a Satellite does not have any foreign key constraints (no snow-flaking, branching or bridging).

These three constructs are the building blocks for the DV EDW. Together they can be used to represent all integrated data from the organization. The Hubs are the business keys, the Links represent all relationships and the Satellites provide all the context and changes over time.

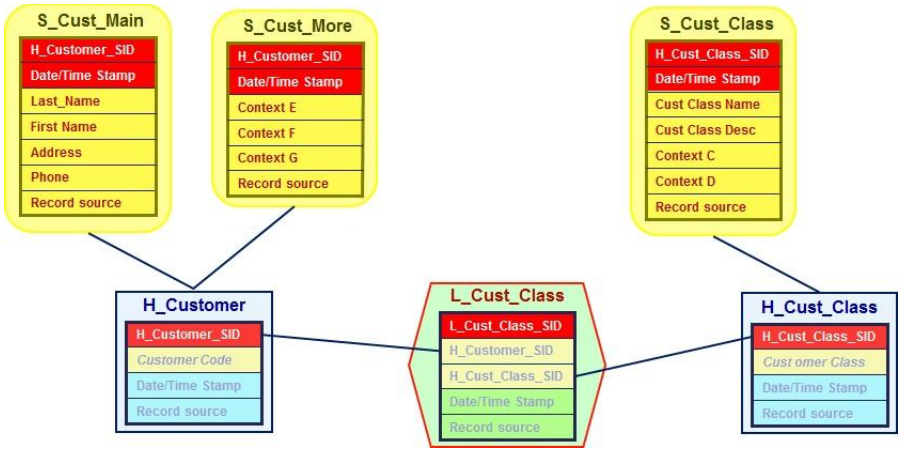

Fig. 4 ***Data Vault Model***

When we look at the Hub and Link together, they form the ***backbone*** or "Skeletal Structure" of the model. This backbone model represents a 1:1 relationship with core Business Concepts and their natural business relationships.

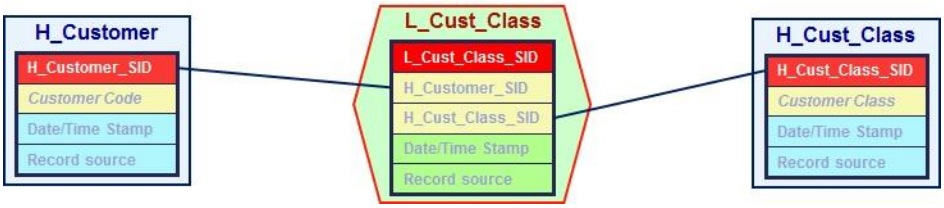

Fig. 5 ***Backbone or Skeletal Structure***

Note that ***all context*** (descriptive information) and ***all history*** are found in the Satellites.

Modeling with the Data Vault

The process of modeling with the Data Vault is closely aligned with business analysis. The first step is to identify the Hubs for the given subject area. Once the Hubs are defined we next model the natural business relationships between these Hubs. Then we design and add the Satellites to provide context to these constructs.

STEP	TASK
1.1	Identify Business Concepts
1.2	Establish EWBK for Hubs
1.3	**Model Hubs**
2.1	Identify Natural Business Relationships
2.2	Analyze Relationships Unit of Work
2.3	**Model Links**
3.1	Gather Context Attributes to Define Keys
3.2	Establish Criteria & Design Satellites
3.3	**Model Satellites**

Fig. 6 ***Steps to modeling with Data Vault***

This process is not concerned with separating facts from dimensions, or from separating master entities from events or transactions. The focus is squarely on core business

concepts – and their unique business keys. In that regard, all of the above are candidates for Hubs. For example events including transactions are modeled as Hubs.

Think Differently

Modeling with Data Vault requires us to *think differently*. Most of us first learned 3NF modeling for operational databases. To manage third normal form, all attributes in an Entity must depend directly on the key of that Entity. So the context attributes that describe a customer (last name, first name, address, city, state, postal code, home_phone, mobile_phone, etc.) must be placed in the Customer Entity where the key uniquely identifies an instance of a customer. If we included attributes that do not depend on the key of that entity then we would not be in 3rd normal form. Likewise if we placed some of the attributes that depend on that key into another entity then again we would no longer be in 3rd normal form.

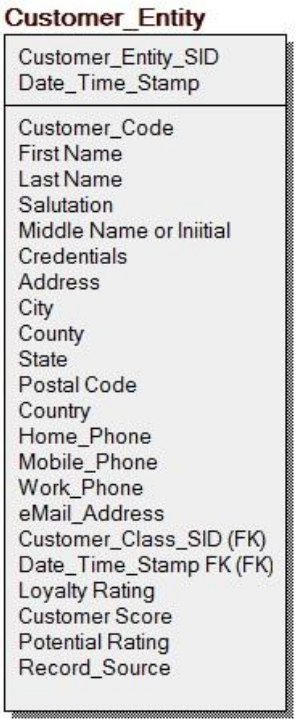

Fig. 7 ***3NF Customer***

At some point we may have also learned how to model using dimensional modeling techniques. Though different modeling constructs and other rules for modeling, the concept of including context attributes inside a table with a key for those attributes remains the same. A Conformed Dimension requires that context attributes depend on the key of that Dimension. Again if we move out attributes depending on a dimension key to some other construct then we no longer have a conformed dimension.

Shown here is a Customer_Entity in 3NF where we can see the Business Key (Customer_Code), the relationship (Customer_Class_SID) and all the context in the form of all remaining attributes in the table. Notice that this is one table including all of this these components.

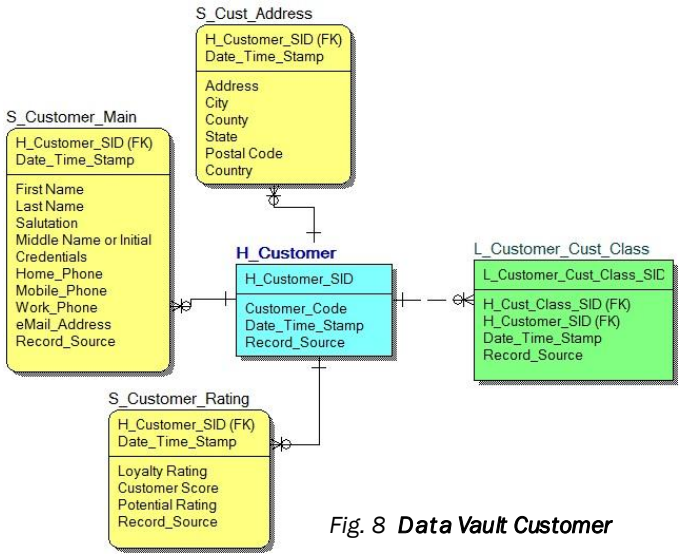

Fig. 8 ***Data Vault Customer***

With Data Vault modeling we separate the business keys from the relationships from the context. All of the business keys are modeled as Hubs, all relationships and associations are modeled as Links, and all context and history is provided for through the Satellites. Shown here we can see that the Business Key (Customer_Code) is in a Hub (H_Customer), the relationship (Customer_Class_ID) is in a Link (L_Customer_Cust_Class), and the context is modeled in several Satellites.

Look back to the 3NF model and now consider that all of the same information (the same components of data) about "Customer" are represented fully in both models. Interestingly both models represent a dependency on a single business key. Actually if we draw circles around each of these models we can see that what is inside each circle is a representation of the same single business key, the same set of attributes and the same relationship.

Notice here that the reaching out from the "Customer" to the Customer Class is modeled through a relationship with a FK inside the 3NF circle.

The same is true for the DV circle in that reaching out from "Customer" to Customer Class is modeled though a relationship (Link) with a FK in that Link and on the perimeter of the circle.

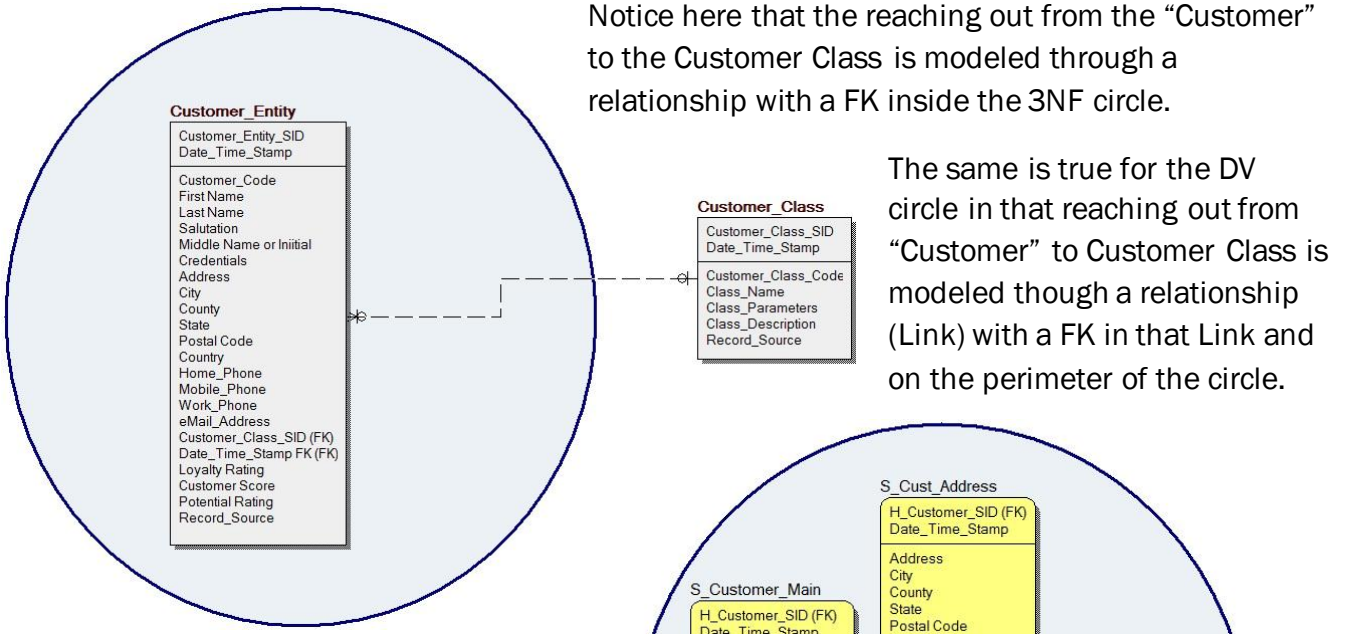

Fig. 9 ***3NF Model***

It is important to think of the DV circle in the same way as the 3NF circle.

This means that a) all things in either circle are dependent on a single business key, b) relationships pass through the circle directly from table with the BK, c) the only grain shift in either circle is based on Date/Time stamp for the purpose of tracking history.

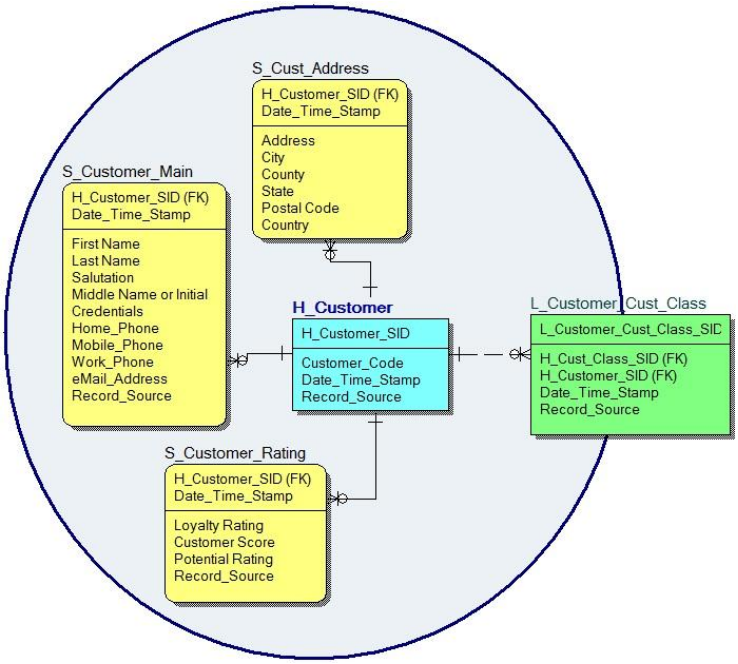

Fig. 10 ***Data Vault Model***

HINT: *As you progress with Data Vault Modeling, this view of thinking differently will become more and more important. We tend to see tables the same way we have always see them. For this reason, we tend to re-combine keys with relationships with context. But as soon as you do, you actually stop vaulting and return to other forms of modeling. So before you change the grain of a satellite, include a relationship FK in a Satellite or Hub, please consider the above circles analysis and reconsider.*

The Business Key

At the core of the Data Vault is the **Hub** which we refer to as the business key. Perhaps the most important initial step in modeling a DV EDW is to identify and thoughtfully design these business keys. To begin with, a Business Key is representative of the core business entity like "customer" or "product" for example. In addition, the BK also represents event based keys such as "sale" or "transfer". In this way, the design process for the Data Vault does not concern itself with the differences between the person/place/thing type entities and the event type entities. To put this another way, we are not concerned with differentiating Dimensions from Facts but rather are focusing on identifying Business Keys which can represent either.

This approach is then different from traditional approaches for modeling operational systems or data marts. The closest comparison would be to consider our efforts in defining Master Data elements for an MDM initiative. In this case as well, the focus is on the core terms used in managing the business.

Since the DV Program is organizational in scope, the business keys should also strive to be meaningful across the enterprise. So our quest for these keys should result in Enterprise Wide Business Keys (EWBKs). Note also that the keys arriving from specific source systems are typically not fully aligned with these EWBKs. For this reason, we do not place too much emphasis on the keys represented in any particular source system.

> *NOTE: Since we are typically dealing with hundreds of sources, each commonly subject to updates and changes, we should not plan to model our EDW using keys driven by a subset of these source systems.*

The process of identifying and modeling these EWBKs is then closer to a business requirements gathering process than a source system analysis. Balancing the various input factors, with an emphasis on the business versus the technical, effectively summarizes the best practices for this process.

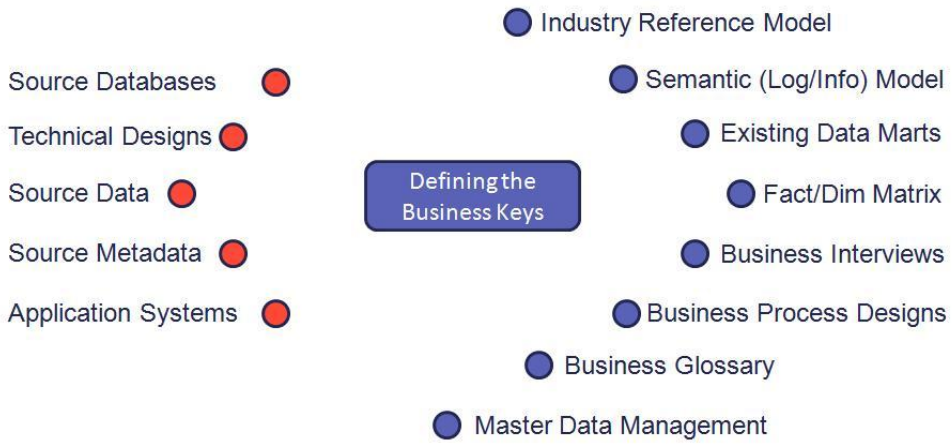

Fig. 11 ***Input factors for designing the DV EDW Business Keys***

The primary inputs for this process include:

First the business process designs, business interviews, existing data marts, business metadata, process metadata, a business glossary if is exists, semantic models, logical models, information models and the master data management artifacts (if existing) and the industry reference model (to the extent certain components are aligned) and then

Second the technical designs, source databases, source metadata, application system (guides, manuals, designs) and actual source system data.

Note that the EWBK should be a key that transcends time and withstands the replacement of any specific source system. The source system keys will then require some form of alignment to match up with their related target EWBKs. This alignment will often be at odds with the fully raw and auditable characteristics of persisted source system loads. In the past we have either resolved this alignment on the way to the marts, or more commonly, created a cleansed "gold" record within the four walls of the data warehouse itself. The former solution leads to silos and anomalies while the latter can compromise auditability and user acceptance.

Business Key Alignment

Because the DV EDW absorbs all data all of the time and maintains full traceability back to source system feeds, the data warehouse must not lose resolution on these auditable systems of record. At the same time integration around the business key – the EWBK – is a core function of the DV EDW. So the EDW today has a built-in challenge related to data integration – the alignment of the Business Keys (enterprise-wide) with the Raw & Auditable components of the Data Vault.

For the one side we know that we cannot rely on leaving the raw details only in the staging area or in our archives. We do need to have all data loaded into the EDW to be a true auditable "mirror" of the sources. See the bottom left "Raw Keys" in figure 5 below.

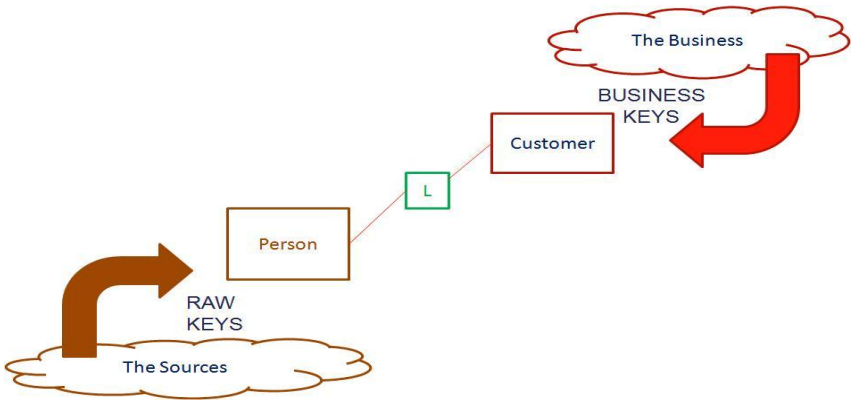

Fig. 12 ***DV EDW Key Alignment***

On the other side we know that the enterprise data warehouse (EDW) must be aligned with the organizational view of the business keys/terms (EWBKs). See the top right "Business Keys" in figure 5.

The integration of these Raw keys with the EWBKs represents a core function of the EDW today. In effect, we have been boxed in by the upstream requirements (build a DW that includes all data at the atomic level and with full traceability) and the downstream requirements (align the Business Keys with the organization at the enterprise level using business terminology).

> *NOTE: We cannot rely on having these transformations happen in the Mart Staging or Data Mart layers as a) the Mart Staging is not intended to be persisted and b) the Data Marts are departmental in scope (not Enterprise Wide). Not sharing this business key alignment through the EDW will result in a failure to integrate around the true business keys and will result in the downstream inconsistencies common to data silos.*

The extent to which the sources are already aligned with the EWBKs will determine the scope of integration and alignment that must occur in the DV EDW. See figure 6 below.

Fig. 13 ***DV EDW Scope of Key Alignment***

The alignment of these keys is facilitated through **Links**. As can be seen in figure 5 on the previous page, the Raw Key of "Person" is aligned with the Business Key of "Customer" using a link structure.

Note that the naming conventions are not adequate in and of themselves to warrant the separation of Raw and Business key Hubs. If in fact Person and Customer meant the exact same thing to the business and were in fact true business term synonyms, then the raw system load of Person records could populate the Customer Hub directly. However, in this case the Person Raw key does in fact mean something different than the Customer Business key. In this case we assume that there are business rules at play – for example a Person record is determined to be a Customer if they were involved in a Sale transaction, there was a non-zero purchase price, the transaction was successfully completed, and the Sale was not cancelled. As you can see here, the raw auditable load is to the Person and the business aligned load is to the Customer.

Those tables that are loaded using this type of business processing must be identified as "***sysgen***" record source records (generated by us through a business rule driven process).

These components of the DV EDW are often referred to as business data warehouse (BDW) or business data vault (BDV) components.

> *NOTE: Business logic in the BDW or BDV can take many forms and can relate to many types of transformations. The logic specifically targeting the alignment of raw and business keys is a subset of this area and often referred to as the BAR component (Business key Alignment Rules). Why is this important? Because the primary objective of the DV EDW is to integrate and historize data from various sources. The BAR rules are the heart of this activity.*

Both the BDW and the BAR layers are part of the business aligned data warehouse.

Architecture

The high level DV EDW architecture includes an EDW that aligns the RAW with the BAR levels. The high level architecture can be represented as seen here.

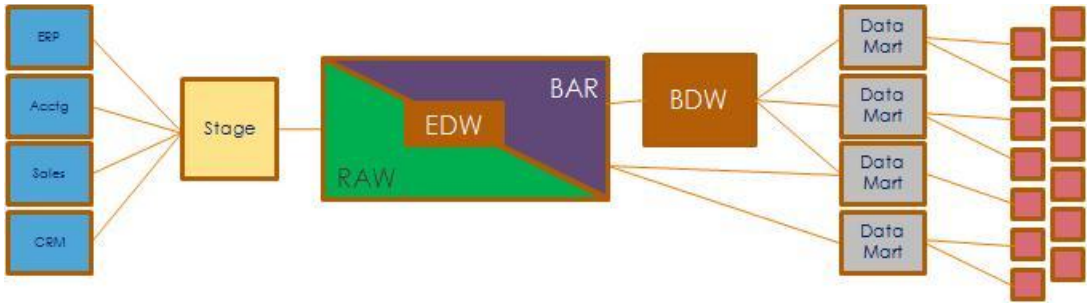

Fig. 14 ***DV EDW Architecture***

Note that the business rules represented in the BDW layer can be applied either in the same area as the EDW or as a separate layer. Data Marts can then source the BDW and the EDW layers as appropriate. The Stage and Data Mart areas are not persisted.

In a typical scenario, the Stage area is not persisted (data is not kept but rather overwritten). The EDW components are persisted, the BDW layer is commonly persisted however it is not required, and the Data Marts are not persisted. This latter point represents an important distinction between dimensional (federated) data warehouses and the DV data warehouse. Dimensional data warehouses are based on the concept of persisting the dimensions and associated facts.

The RAW and BAR layers represent a logical designation. While this separation is often also physical, this factor is not required. In fact one common approach includes a logical separation with Links managing the alignment within the physical layer.

Sample Data Vault Model

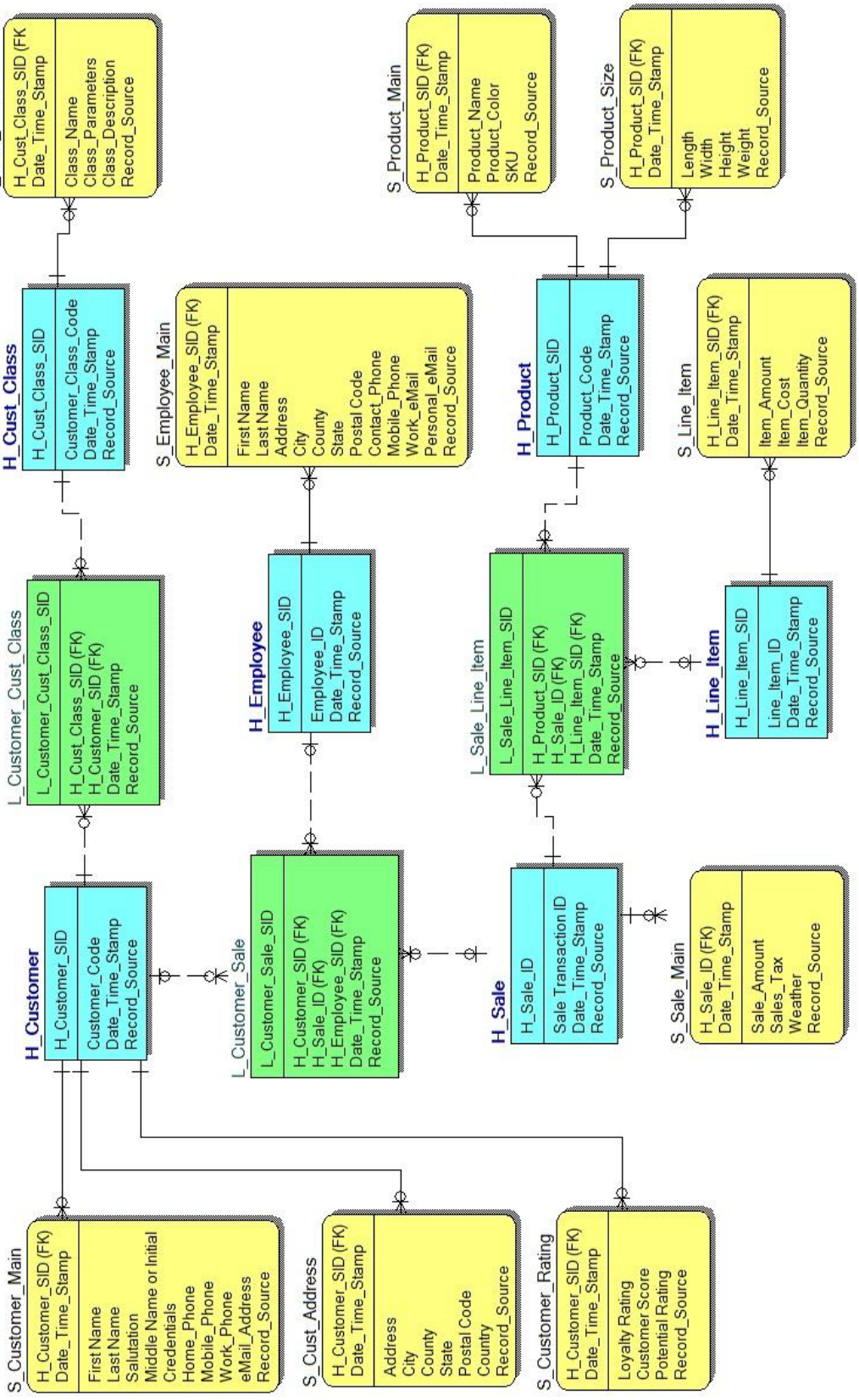

Fig. 15 ***Full Data Vault Model***

Hybrid Tables

The data vault approach has defined a set of Hybrid tables that can be used to make the overall deployment more efficient. These are applied on a case-by-case basis as appropriate for the specific circumstances.

The **Point In Time table (PIT)** is a modified Satellite table that tracks the valid time slices of the satellites surrounding a particular Hub. This is populated to make the process of associating relative context/descriptive data together for reporting purposes.

The **Bridge table** is a modified Link stable that flattens the relationship between Hubs including important related context/descriptive data (potentially also the business keys) into a single table for ease of access and performance.

In all cases, these and other constructs can coexist in the DV EDW provided however that they are always noted as "sysgen" tables and utilized only for performance reasons. The related historical and auditable data that is used to load these constructs must remain the sole source of the EDW data over time.

Applying the Data Vault

Data Vault modeling is uniquely useful when modeling a data warehouse. An Enterprise Data Warehouse (EDW) project is specifically well aligned with the features of data vault modeling. One primary benefit is the ability to adapt easily to changes in both upstream sources and downstream data mart requirements. This provides us the ability to build incrementally and to run a truly agile data warehouse program. The data vault data warehouse also easily integrates data and inherently manages history providing for a true enterprise data warehouse.

Data Vault modeling has also proven to be the preferred modeling pattern for special data warehouse situations including truly operational data warehousing, Big Data integration, Information model based DW models, meta-data driven data warehouse deployments and even data-driven generic data warehouse models.

Understanding the full benefits of the data vault modeling approach starts with getting your certification. This process is facilitated by Genesee Academy and includes materials, online lectures, exercises, two days in a classroom with lectures, labs and group modeling exercises. On the last day there is an exam which results in the certified data vault data modeler (CDVDM) designation.

Please visit GeneseeAcademy.com for more information on course schedules and registration.

Final Note

The Data Vault approach is growing and adapting from year to year. Incremental changes to the modeling approach, rules and best practices can be expected with some frequency. Please note that this guide should be applied in concert with current updates found online on data vault forums, LinkedIn and from certified practitioners.

Online training including updates and current topics is also available online at DataVaultAcademy.com