the data vault guru

a pragmatic guide on building a data vault

The Data Vault Guru: a pragmatic guide on building a data vault

Why write a book?

There exists plenty of literature on the web about Data Vault and often the message and methodology is conflicting and designed with old school principles to data delivery or even focussed on a particular understanding of agile delivery. As a data professional with experience in ETL/ELT and data modelling I felt that there needed to be something to bring both viewpoints of data delivery together, one from data automation and one from data modelling. They are not the same but if a data vault is properly delivered it must be cognisant of both.

Data Vault continues to evolve and communities of data vault practitioners exist around the world. Forums and meetups focus on how data vault has been applied to specific situations and tools (which is great!) but for attendees they often do not have access to the tools or may never encounter the unique situation. Data Vault models are delivered with a partial view of the standards; although Data Vault implementations have evolved the standards remain consistent.

Although the book is a written account of what I consider to be an almost complete implementation of data vault it has to be said that reading about data vault and getting trained in it does not necessarily mean you know how to deliver a data vault. The book takes a platform-agnostic view and describes the intended delivery and modelling patterns for Data Vault. Think about this, a Data Vault is not designed to model your data, it is designed to model and integrate ***all*** of your data, but don't boil the ocean! The beauty of data vault is that once you have established the standards and understanding building onto the data does not change any of the existing data, it simply adds to it; vertically by adding data sources to existing structures and horizontally to new structures.

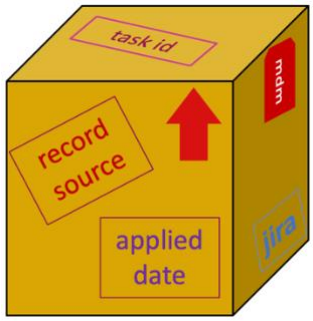
Often once a Data Vault is committed to, some considerations are not thought of until you get there, for example what the categories of business rules are and how to build a Business Vault and how to get the data out of data vault! And falsely comparing dimensional modelling to data vault!

What is in it for the reader?

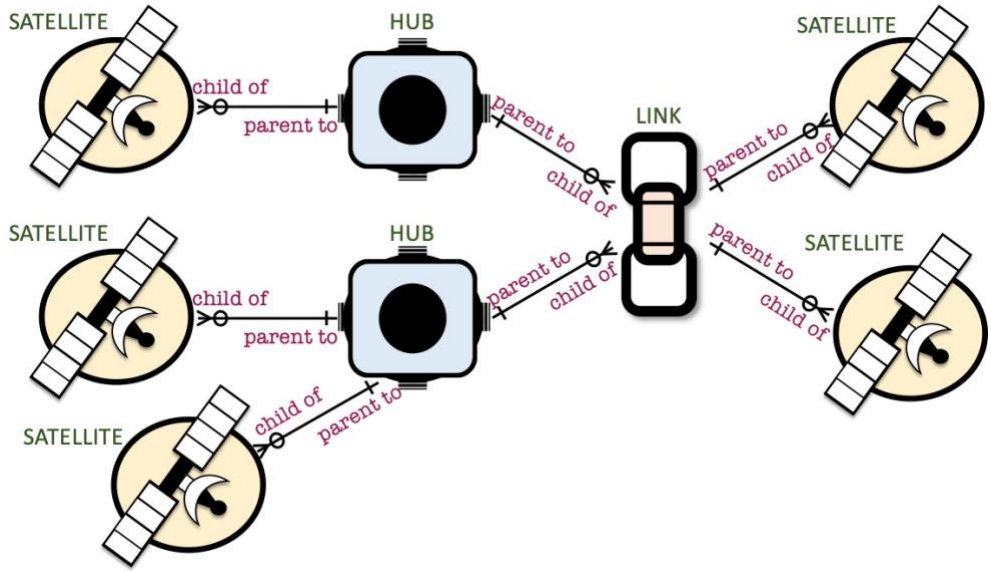
- We discuss the levels of architectures present in any data-centric enterprise; how to view business processes and how they are modelled into raw and business vault.
- We get into data latency and how that pertains to persistent vs transient staging, whether or not creating views for Business Vault is a good idea.

Relevant types of business and non-business keys and their relationship to data vault, themes like surrogate hash or sequence, smart, natural, driver and zero keys.

- How to think about time in the data vault context through an insurance example by comparing business dates, data packaging dates and load dates as well as the three forms of time: discrete, evolving and recurring



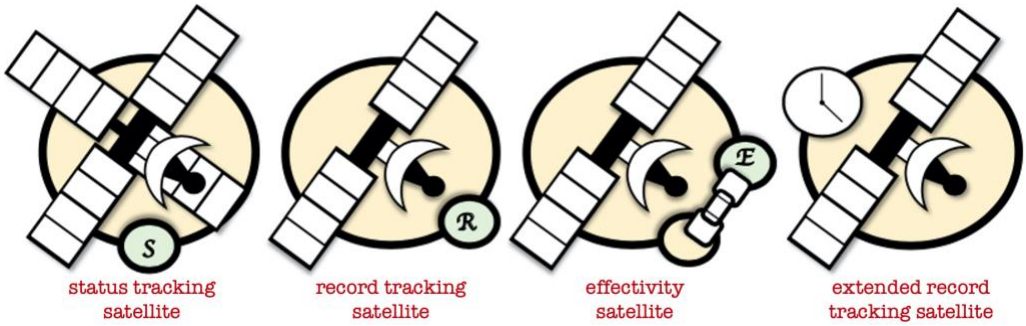
- We get into the classic Raw Vault structures and also discuss naming standards, the DV-tags (metadata) expected in each data vault structure, recommended indexing of each. As for the three model structures themselves…
    - Hubs - we present what they are, the standard SQL to populate a hub table, sample structure, defining what the standard business key treatments are and passive integration. How to apply variable business key treatments and business key collision avoidance strategies. How to load composite key business entities into a simple key entity hub table
    - Links – we discuss what the unit of work is, how to build link tables either as regular links, same-as links, hierarchy links or a combination of those with an example on why you would or would not include a dependent-child key in a link table. All with the standard SQL to populate a link table.
    - Satellites – the grain of change/snapshot-data we get from source systems. From single record grain to multi-grain with dependent-child keys, how to include multiple updates to a business entity in a single batch, how to implement a multi-active satellite and satellites with advanced data types like arrays and structs. Importantly how to split satellites and why, which content will be loaded to link-satellites and which to hub satellites and why you would split those. Not to mention how to handle personally identifiable content (PII). All satellite grains and types are presented with sample SQL on how to load them.



- Change data capture structures are also discussed, how you would load CDC and near-realtime content. As well as how to model reference data into data vault.

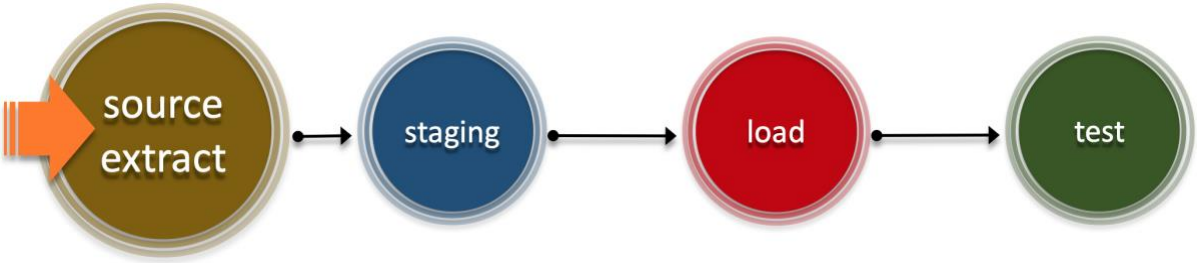
- The content derived in the analytics platform is stored in your Business Vault because it was sourced from Raw Vault but populated using the same automation templates as Raw Vault. The book shows you how to build a BV-link and BV-satellite and how to ensure data and business rule lineage while tying that content to a data governance tool
- We cap off the data vault modelling section with a decision tree on how to decide what artefacts to build and at what grain.
- The modelling section is completed with examples, SQL, structure and considerations for
    - Record Tracking Satellites
    - Status Tracking Satellites and
    - Effectivity Satellites

  Comparing where and why you would use one peripheral satellite over another.



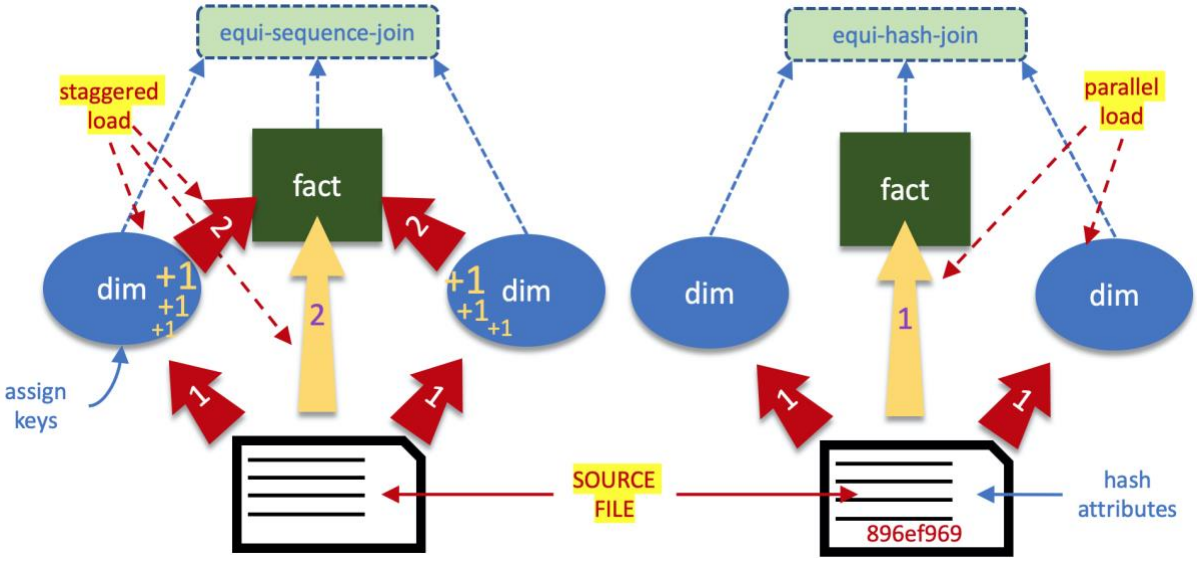
- From the automation perspective we discuss
    - the three loading patterns for hubs, links and satellites
    - loading patterns for PITs and Bridges
    - how to load to shared hubs
    - orchestrating the load from staging content to raw vault,
    - orchestrating business vault,

  And we include code for a test automation framework!

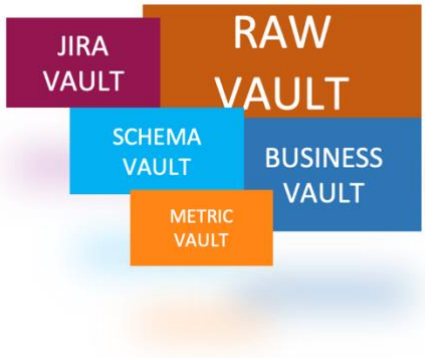


- The book dedicates a chapter to data-driven timeline correction using a variation of the record tracking satellite that is decoupled from RTS but is capable of delivering timeline correction to all data vault artefacts, EFS and NH excepted. The chapter includes the SQL to populate and correct, scenarios and caveats to using the pattern.
- Getting the data out, we discuss the SQL needed, patterns of data vault queries, automating views over the standard data vault tables and automating views over a hub/link and all its satellites together for all data vault satellite table grains (multi-active/dependent-child/intra-day/regular)

- Query assistance structures and patterns are described and discussed for point-in-time (PIT) and bridge tables and where ghost records come into play.
- We follow that up by discussing data-driven dimensional modelling that is backed by data vault as the data warehouse



- Towards the backend of the book we get into data warehouse migration towards a data vault with a decision tree on how and what to migrate plus how to deal with content overlap. This is followed by a discussion on integrating the various levels of data models into a single data vault model (enterprise, industry and application models).
- We also present variances of data vault, a metric vault, JIRA vault and a schema vault designed to support data vault automation.
- As far as tools goes the book includes a checklist on what every data vault artefact must include, how to automate loads to them. A model scorecard for data vault inspired by Data Model Scorecard by Steve Hoberman and a guidance on how to Mob Model and how to setup a model register for collaboration and governance. And lastly a data vault automation tool scorecard.



The examples and samples in the book have a master data management flavour on how to use MDM as a source.

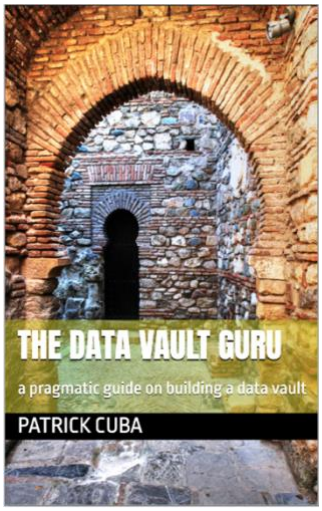
Who would benefit?

- Data Architects considering data vault,
- Data Vault Modellers for a reference guide,
- DevOps engineers tasked with building a data vault,
- Solution Architects designing where data vault fits,
- Data Analysts on how to query and use data vault,
- Data Migration specialists on what to consider migrating to data vault.

Where can you find the book?

If you want to know every detail on how to build a data vault, this book is your answer.

US: https://amzn.to/3d7LsJV
UK: https://amzn.to/3nsqTfR
AU: https://amzn.to/30IxOYF
DE: https://amzn.to/2TiAsAb
FR: https://amzn.to/37yfnKl
ES: https://amzn.to/3jl5tOr
IT: https://amzn.to/37Awag6
NL: https://amzn.to/35sCpjc
JP: https://amzn.to/3dNJgYq
BR: https://amzn.to/3dRvIek
CA: https://amzn.to/3jl5LVx
MX: https://amzn.to/35pkslI
IN: https://amzn.to/3jl65DJ



#datavault #datawarehouse #analytics #datamodelling #thedatamustflow

Where to get more Data Vault?

A Data Vault Alliance exists that brings together Data Vault professionals from around the globe with decades worth of experience from delivering the classic batch-oriented data warehouse to real-time delivery. Discussions are both business and technical oriented and often the topics cover even the not-yet popular technical platforms as Data Vault is being used to deliver the modelling advantages you would expect!

https://datavaultalliance.com/