Data Vault

An in-depth look

Juraj Pivovarov

Background

- Owner and Consultant at Metadata Innovations Inc.
- Clients: AEP, OAS, Divestco, CWD, Nexen, AER
- Data Architect, Data Analyst, Data Modeler
- Software Developer, Combinatorial Optimization, Image Processing
- M. Sc. Computer Science
- B. Sc. Pure Mathematics

Hobbies

- Chess, Scrabble, and Speedcubing

Background and History

Q. What is Data Vault?

A. Data Vault is a **modeling methodology** for the enterprise data warehouse.

- It is not something you **buy,** but something you **implement**

Data Vault encompasses

- Data Warehouse Architecture
- Data Vault Modeling
- Data Vault Methodology
  - Project planning
  - Project execution
  - Review and improvement

Why learn Data Vault?

- You may want to USE it!
- You may come across it
- Adopt some of the ideas

History

- Data Vault was invented by Dan Linstedt
  while at US Department of Defense
- First published 2001
- Very popular in Netherlands in insurance and banking.
- Catching on in North America

Pedestrian Google Search comparison on 2016-09-11

- Data Vault Data Modeling: 296,000 hits
- Star Schema Data Modeling: 462,000 hits
- Data Modeling: 9,140,000 hits

Q. What’s wrong with 3NF data models?

Traditional Data Modeling and Warehousing has difficulties with

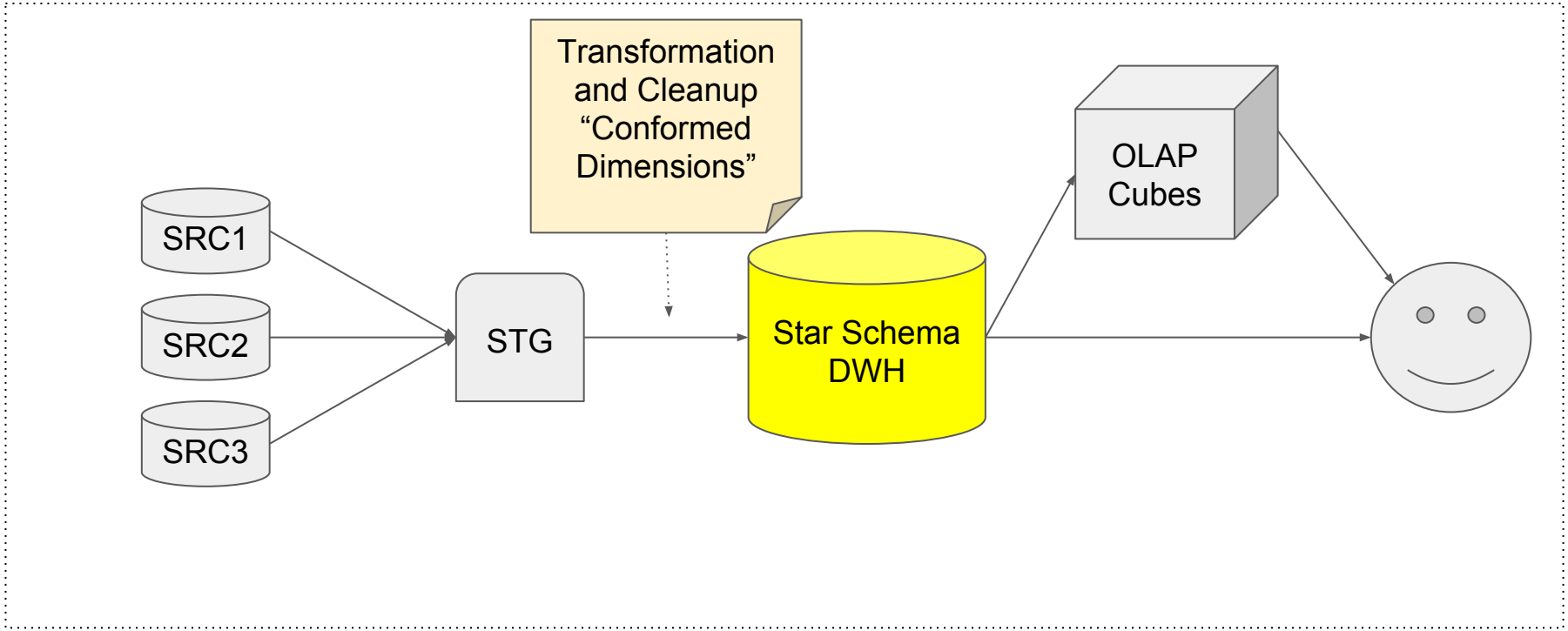
- Changing key structure
- Changing relationships from 1:M to M:M
- Performance at scale
- Complex loading dependencies
- Inconsistent History tracking

Data Vault Promises

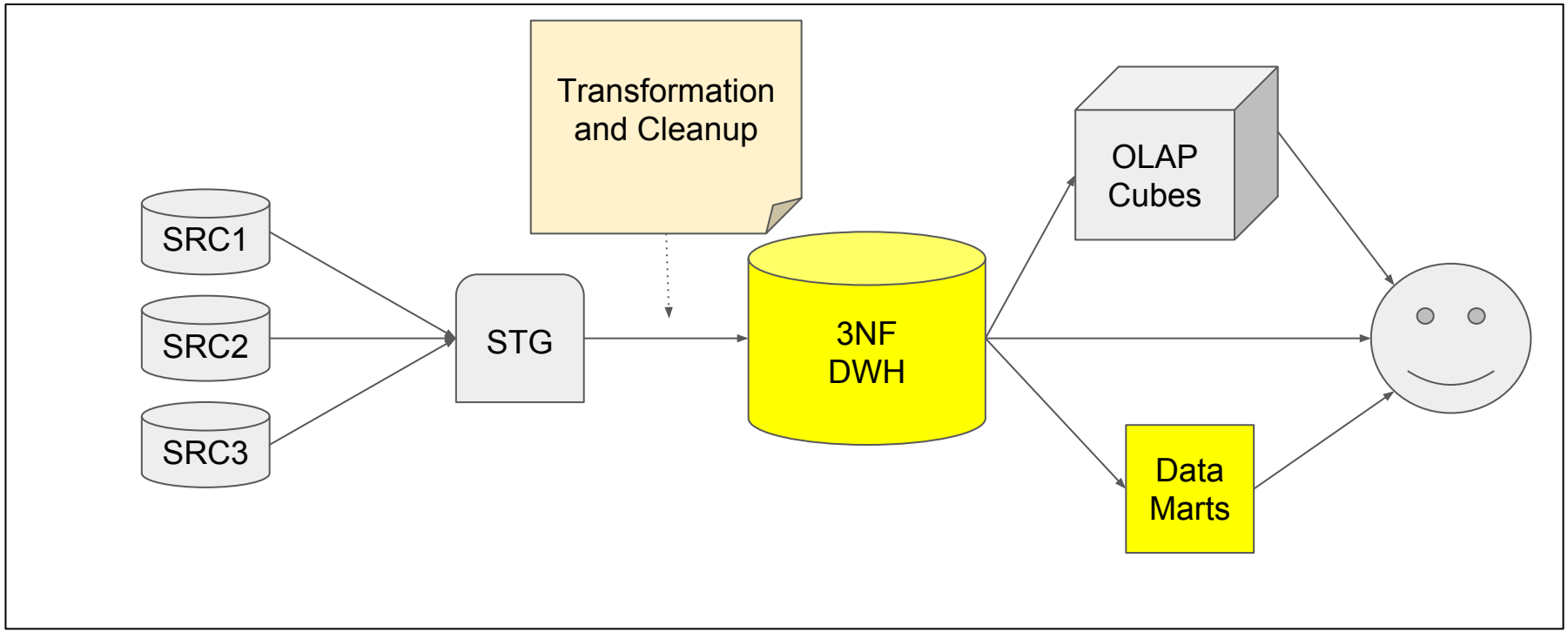
- “All the data, all the time”
- Agile and flexible
- Scalable

Data Vault Architecture

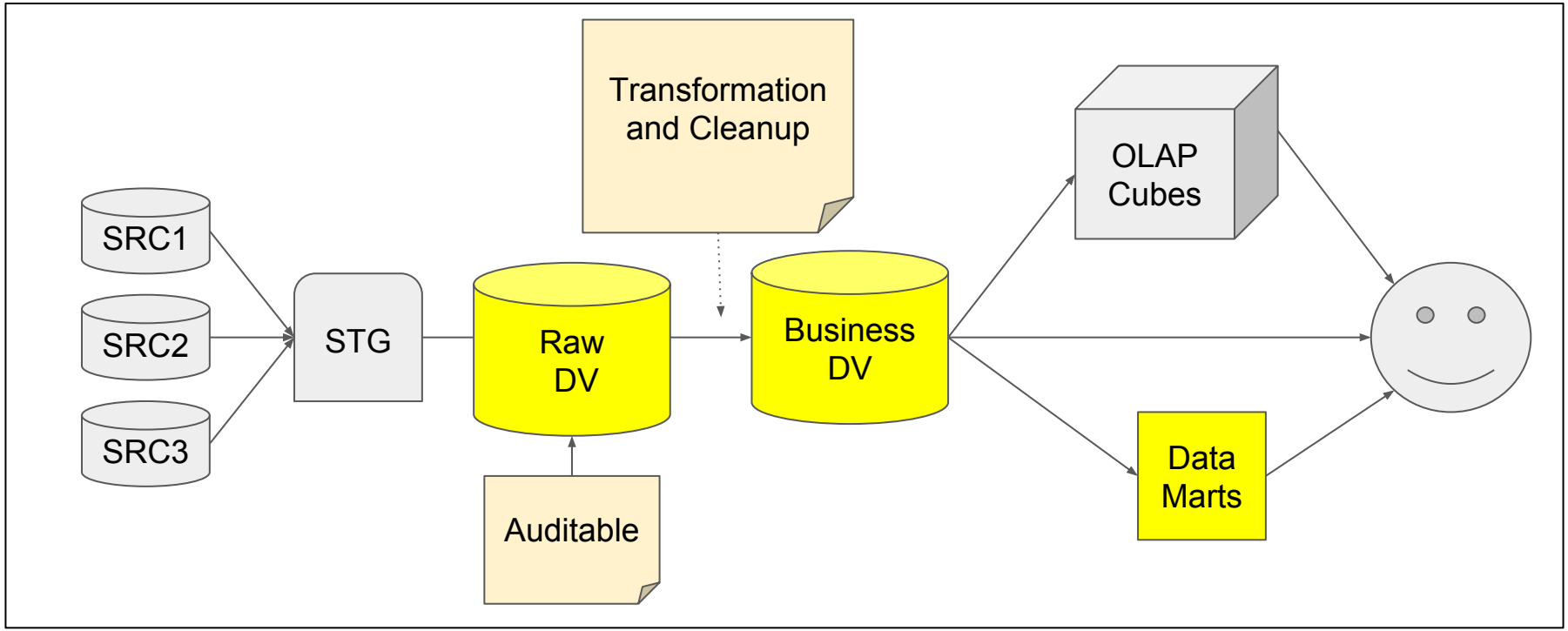
Kimball Warehouse



Inmon Warehouse



Data Vault Warehouse



Data Vault Core Constructs

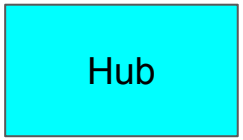
Data Vault Types

- Hubs
  - Unique list of business keys
- Links
  - Unique associations of business keys
- Satellites
  - Descriptive data, time variant
- Reference Tables (optional)
  - For capturing meanings of codes used

Hub - definition

Represents Business Keys

- Hub Surrogate Key (PK)
- ----------
- Business Key (simple or composite)
- ----------
- Source
- Load Date
- Last Seen Date (optional)



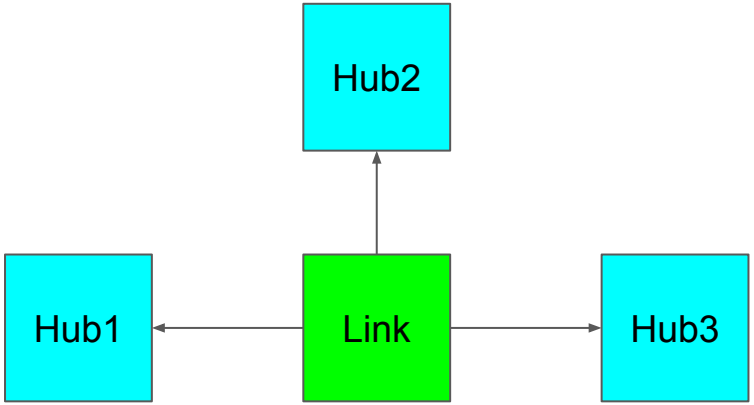
Hub

- Single point of definition for business key
  - Not duplicated across other tables
- Represents the first time the DWH sees the business key
- Never deleted
- Business keys should be able to stand on their own.
- Business keys are what allow you to integrate data across business functions.

Link - definition

Represents relationships between keys

- Link Surrogate Key (PK)
- ----------
- Hub Keys
- Dependent child key (optional)
- ----------
- Source
- Load Date
- Last Seen Date (optional)



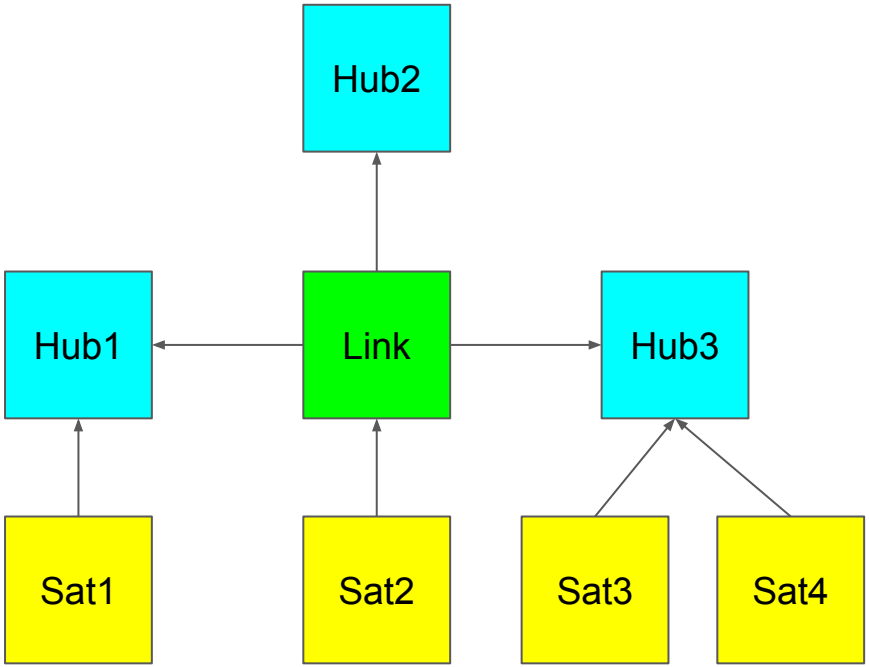
Link

- Links are always M:M
- Links may be between multiple hubs
- “A link **must** have more than one parent table.” [DV1]
- Links provide primary flexibility touted by Data Vault

Satellite - definition

Represents Context, over time.

- Hub or Link Surrogate Key (PK)
- Load Date (PK)
- ----------
- Attributes {1,...,n}
- ----------
- Source (optional)
- Load End Date



Satellites

- Represent context over a fixed time interval
- Split by
  - Source - avoid flip-flop effect
  - Rate of Change
  - Data Types
- Design Decisions
  - Continuum of 1 Satellite per attribute vs 1 Giant Satellite

Reference Tables - definition

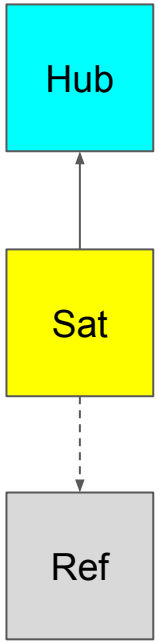
They describe meanings of codes used in Satellites, if applicable.

Many options on how to design them

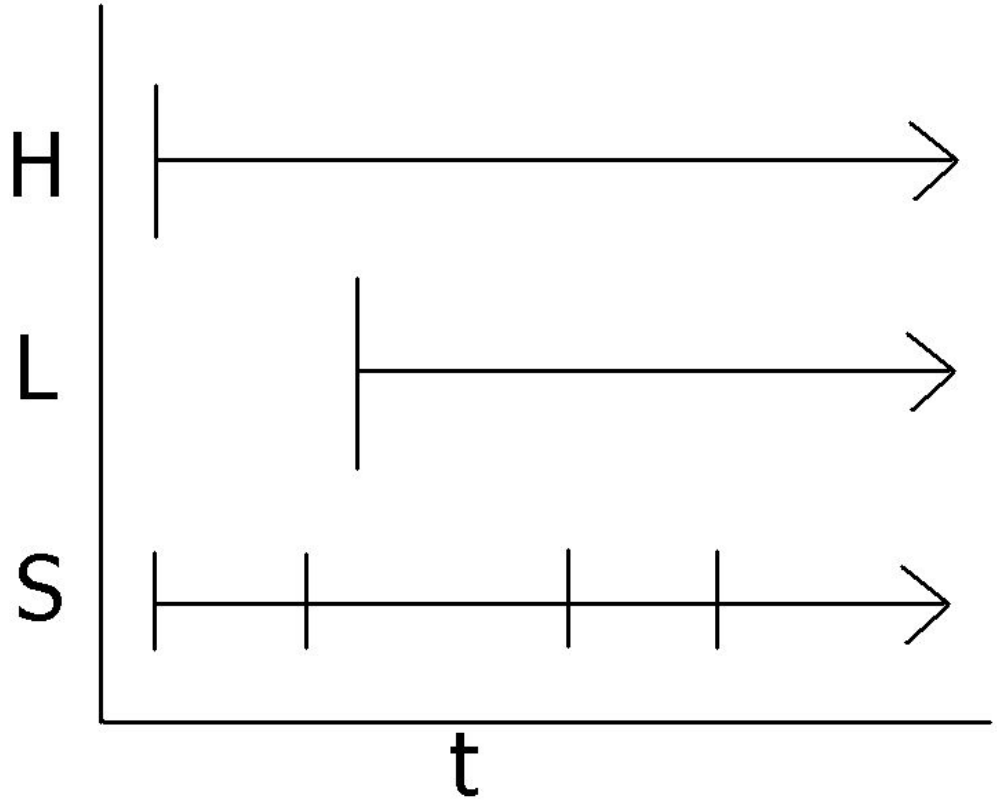
- Directly, with or without history
- or as full blown Hubs and Satellites

Foreign Keys

- Logical foreign keys from Satellites to Reference Tables
- Never physically implemented



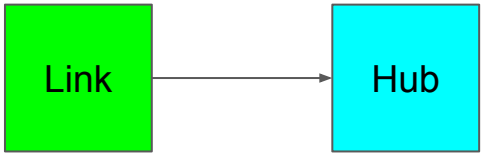
Lifetimes



Links and Satellites - in depth

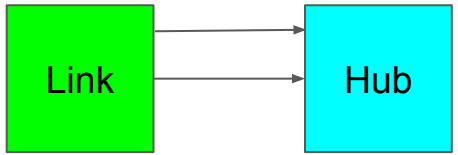
Peg-Leg Links

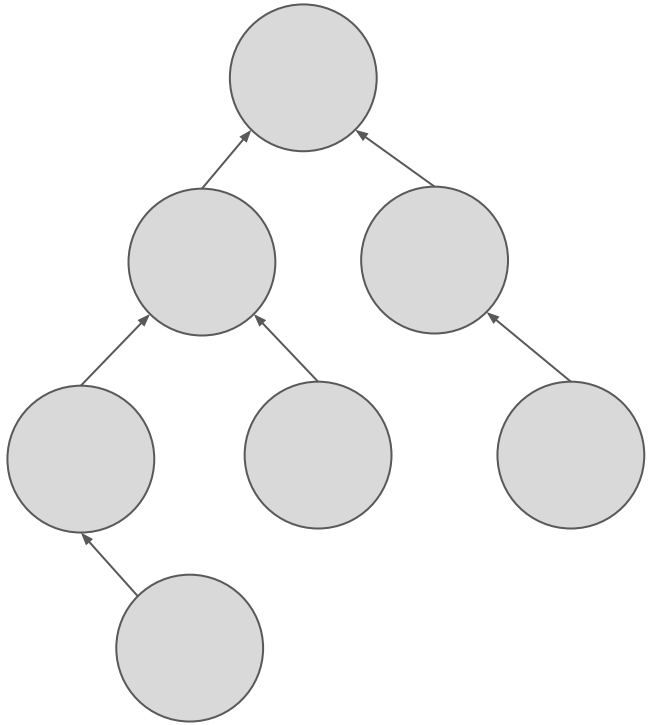
- Degenerate Links, these have only one Hub reference
- “A link **must** have more than one parent table.” [DV1]
- “They connect two or more hubs, (or same hub twice)” [DV2]
- Can be produced as byproduct of some DWH automation tools.



Hierarchical Links

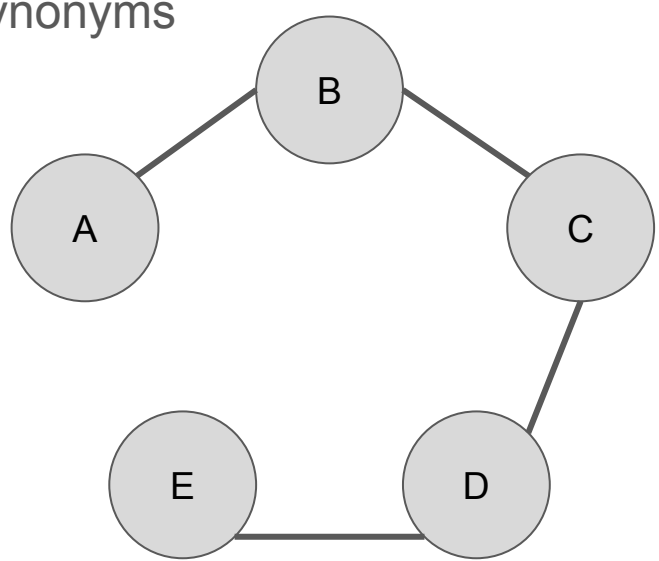
- Parent and child references to same Hub





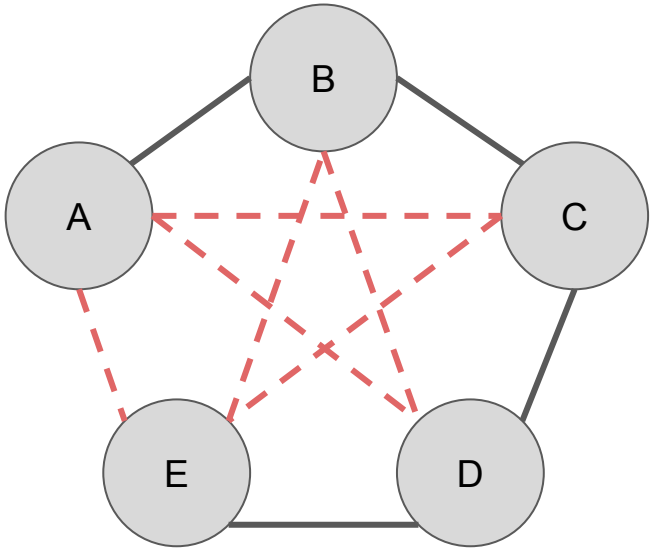
Same-as Links

- Same-as links are used to record hub synonyms
- Here, we have FOUR links
- They mean A=B, B=C, C=D and D=E
- Is A the same-as E?
- Problem: it is not obvious



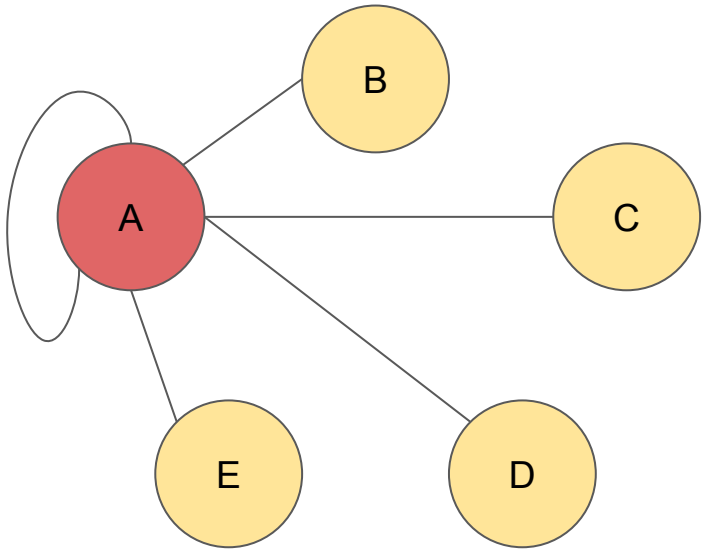
Same-as Links

- We have an equivalence class
  - X=X
  - X=Y means Y=X
  - X=Y and Y=Z means X=Z
- Need Transitive Closure
  - Determine what vertices are reachable
- Four explicit links, Six implicit ones
- In general O(n^2) total logical links

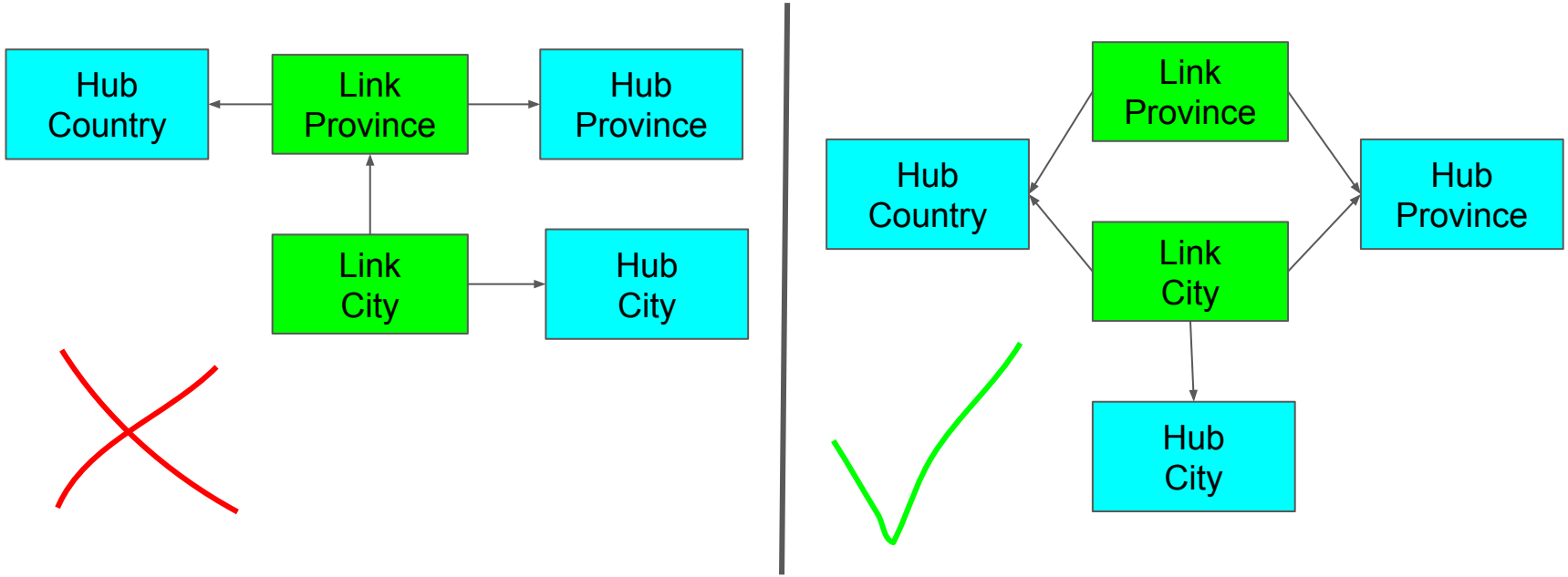


Same-as Links

- Representing Equivalence Classes
- Elect ‘leader’ in each class
- One edge from each to leader
- Don’t forget reflexive edge, A=A



Avoid Links to Links



Multi Active Satellites

Ex: Modeling phone numbers

- Different MOBILE, HOME, and WORK numbers
  - Can have two MOBILE phones etc.
- MOBILE1, MOBILE2, …, MOBILEn is a limited solution.
- Multiple rows are active in the satellite, for given Hub/Link parent.

Implementation:

- Add a SEQ number to the primary key of the Satellite

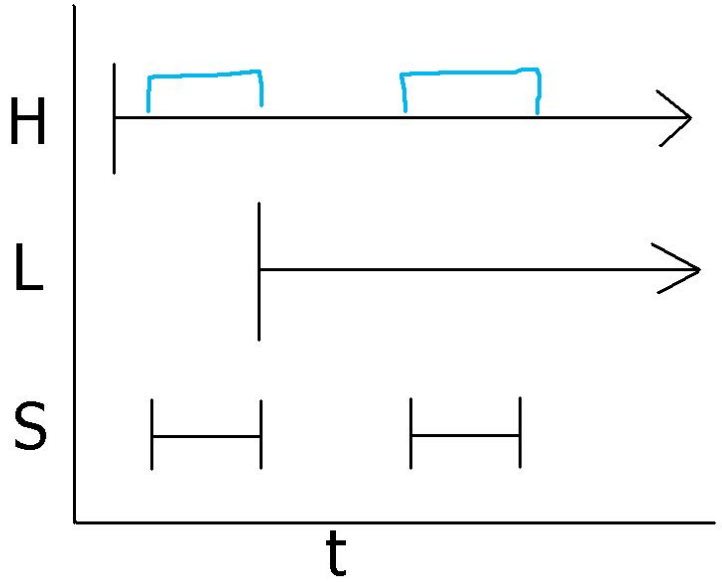
Effectivity Satellites

Ex: Employee leaves, then comes back for short contract.

- Effectivity Satellite models discrete time intervals for which the Hub or Link is valid.

Implementation

- Add Begin Date, End Date to rows of Satellite
- (Do not overload meanings of Load Date, Load End Date)



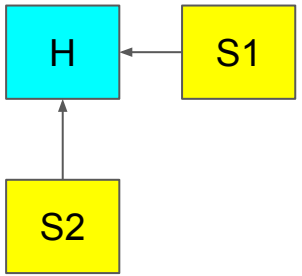
Query-Assist Tables

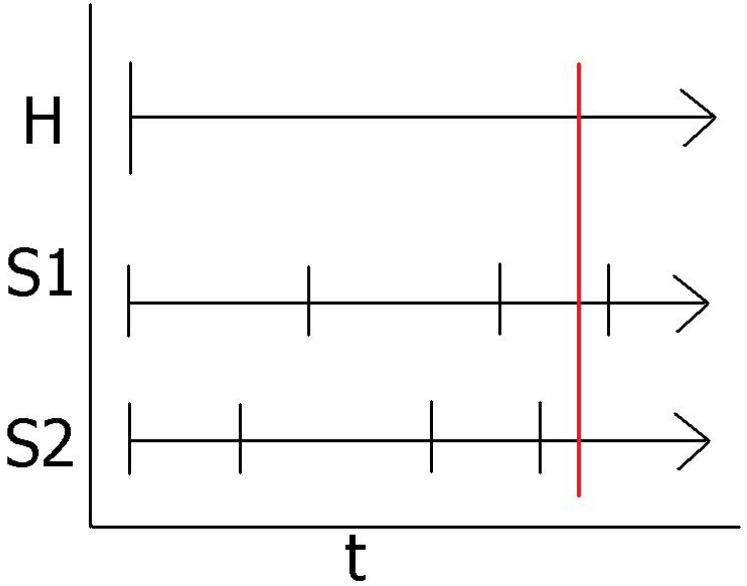
Point-in-Time (PIT) Tables

- Query assist tables (optional)
- Tie together Hub/Link + exact Sat Rows for point-in-time

Example

- Hub Key
- Sat 1 Load Date
- Sat 2 Load Date



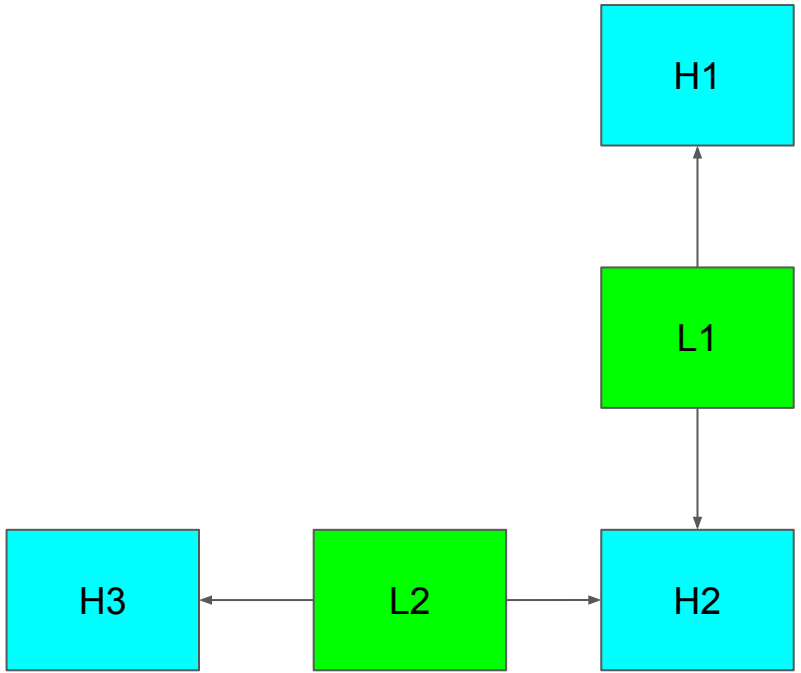


Bridge Tables

- Query assist tables (optional)
- Tie together Links and Hubs

Example

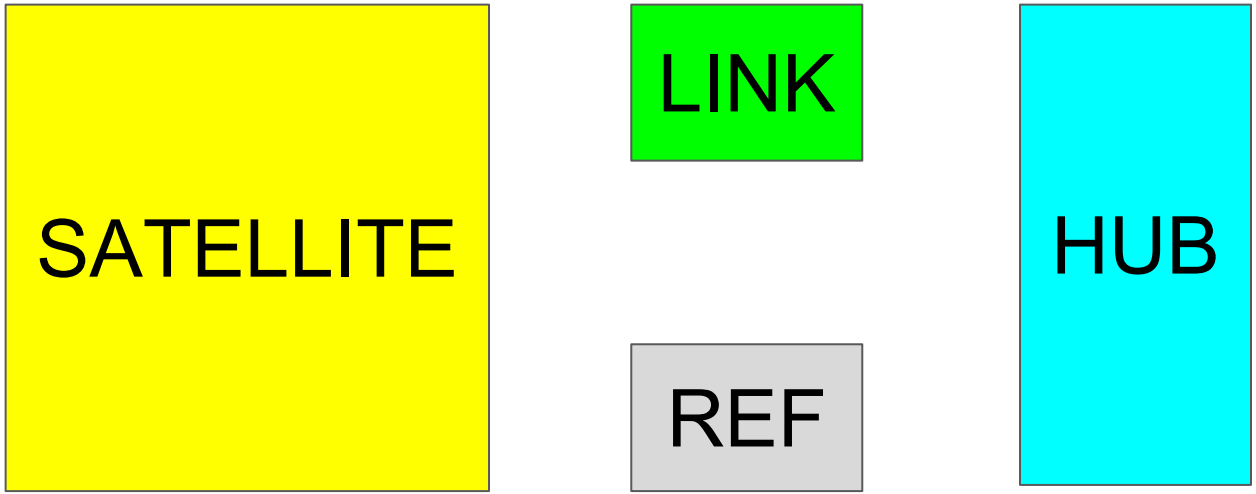
- H1 Key + H1 business key
- H2 Key
- H3 Key
- L1 Key
- L2 Key



Data Vault Tradeoffs

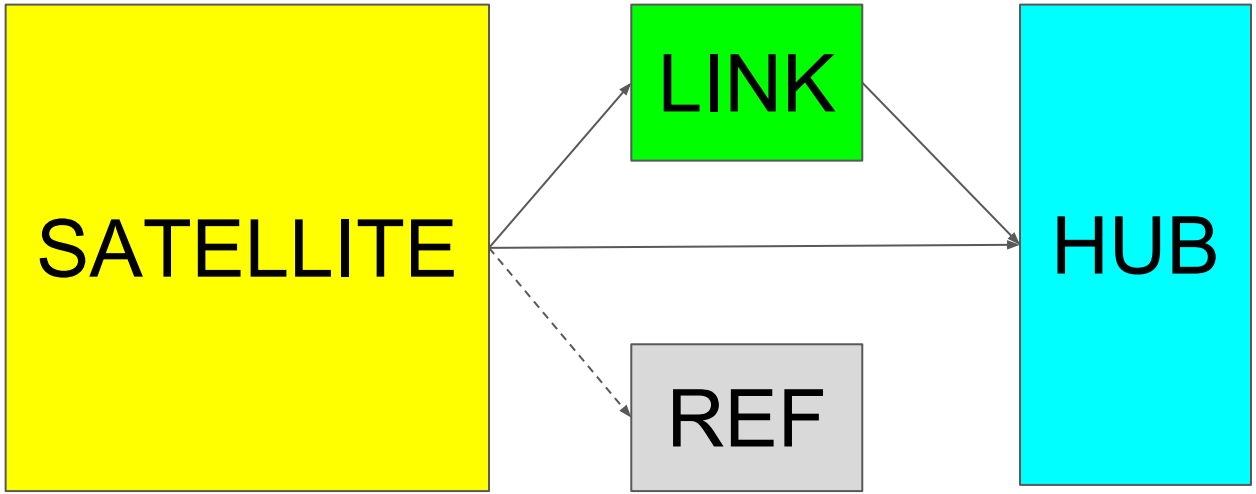
+ Simple Structures

- Consistency: Easy to understand and extend.
- Template-based SQL.



+ Few Dependencies

- Simplicity: Easy to determine “Load of the Rings.”
- Scalability: Very easy to parallelize



+ Other Benefits

- Auditability “All the data, all the time”
  - Nothing is changed on the way into the Raw Vault
  - Business rules and data cleanup happens downstream
- Flexibility: M:M Links
  - If the cardinality of a relationship changes for some reason, no change required in DV
- Extendable
  - Add new hubs, links and satellites without reengineering any existing structures

- Downsides

- Labelled “Hypernormalized”
  - The data model is much more abstract
- Not easily queryable
  - M:M links mean every query needs to handle this case
  - Satellites have multiple rows, must find appropriate one
  - Must respect effectivity satellites on Hubs and Links.
- To automate well, requires some metadata mgmt capabilities
  - Benefits of consistency only come with rigorous adherence to standards
  - Model driven development goes a long way
- Not all constraints can be enforced
  - see Country/Province/State example
  - No R/I with Satellites and Reference Tables.

When to use Data Vault?

- Many source systems, with sometimes contradictory facts
- Auditability is required
- Some team members have familiarity with Data Vault
- Big Data requirements, at least volume and variability
- Anticipated changes deal with cardinality of relationships

- To simplify data warehousing efforts, make use of repeatable patterns

For more information

Q. What's new in Data Vault 2.0 (2016)?

- Hashes vs Integer Keys
  - Hashes are more parallelizable
  - Avoid lookups of surrogate keys
  - Compute them instead!
- Expanding on Data Vault Methodology
  - Many examples with SSIS and SqlServer
  - Ties in to Master Data Services
  - Producing Star Schema Data Marts, etc.

Data Vault Automation Tools

Data Vaults can be auto generated, to some extent, by examining source schemas.

- WhereScape
- Quippu (open source)
- BI Ready
- AnalytiX DS
- Rapid Ace (Dan’s original toolset)

Disclaimer: YMMV.

Bibliography

[DV2] Building a Scalable Data Warehouse with Data Vault 2.0 ~Dan Linstedt, Michael Olschimke. 2015

[DV1] Super Charge Your Data Warehouse ~Dan Linstedt. 2011

[HH] Modeling the Agile Data Warehouse with Data Vault ~Hans Hultgren. 2012

[PENT] Pentaho Kettle Solutions ~Kasper de Graaf (p465-495). 2010

[LinkedIn] LinkedIn Data Vault Group discussions