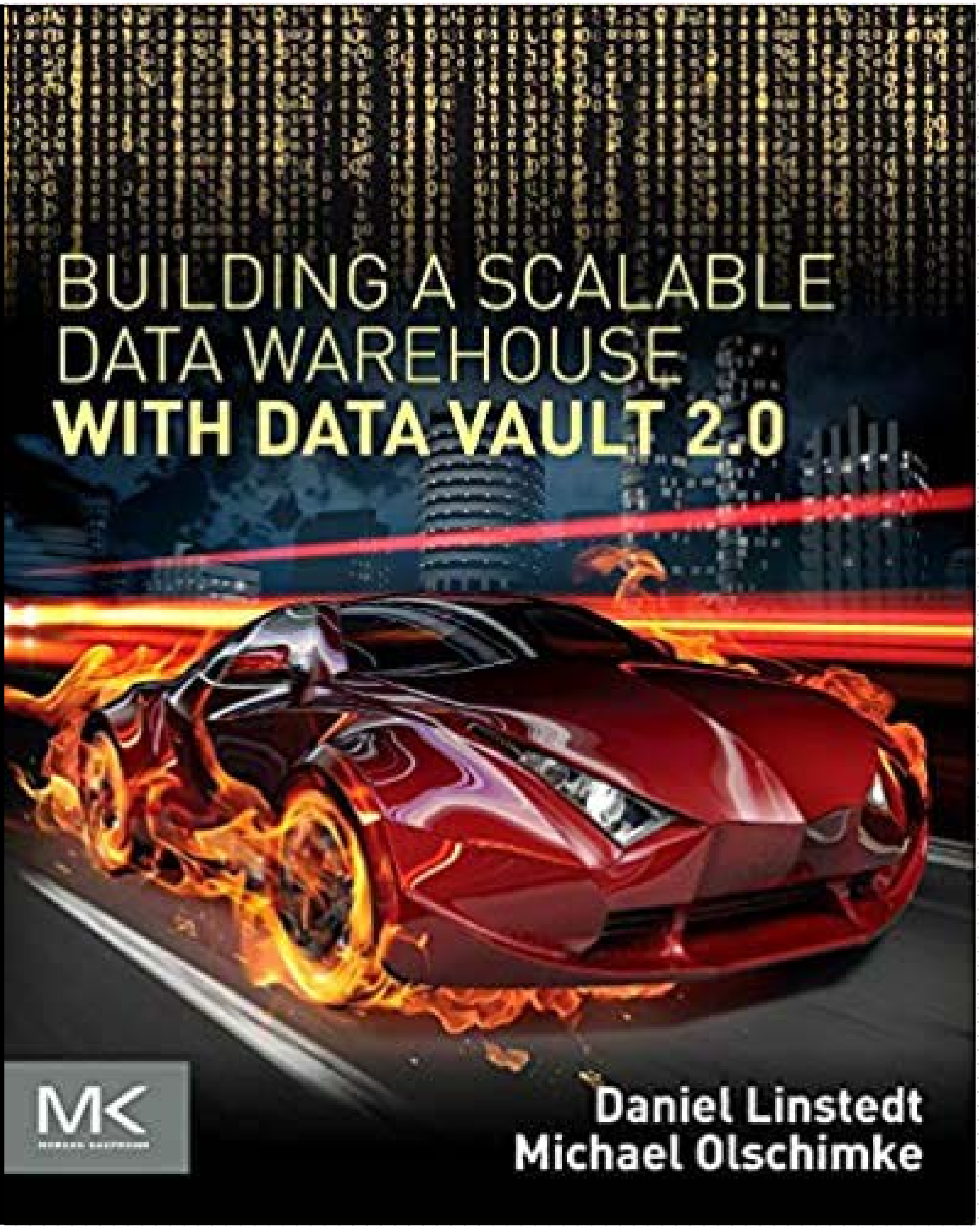
Data Vault 2.0

Adding Value for BI Projects

Ashley Day
Operations Analyst

bakertilly
now, for tomorrow.

Baker Tilly Virchow Krause, LLP trading as Baker Tilly is a member of the global network of Baker Tilly International Ltd., the members of which are separate and independent legal entities. © 2018 Baker Tilly Virchow Krause, LLP



bakertilly
now, for tomorrow.

Ashley Day

Operations Analyst with Baker Tilly

7+ years in data management, warehousing & analysis

Speaking Experience:

- SQL Saturday: Madison, Minneapolis
- PASS: Madison, Milwaukee, Fox Valley
- Internal company opportunities

Primary focus has been data engineering and analysis, working with both on-prem and cloud based platforms

Certifications: Certified Data Vault 2.0 practitioner, Snowflake SnowPro

In Scope

- Overview of Data Vault
    - Methodology
    - Architecture
    - Model
- Overview of main entities
    - Hubs
    - Links
    - Satellites
- Advantages
- Potential disadvantages
- Identifying good candidates for a Data Vault

Out of Scope

- Load patterns & ETL processes
- Advanced Data Vault modelling
- Demo or hands-on lab

Overview

Data Vault 2.0 Defined

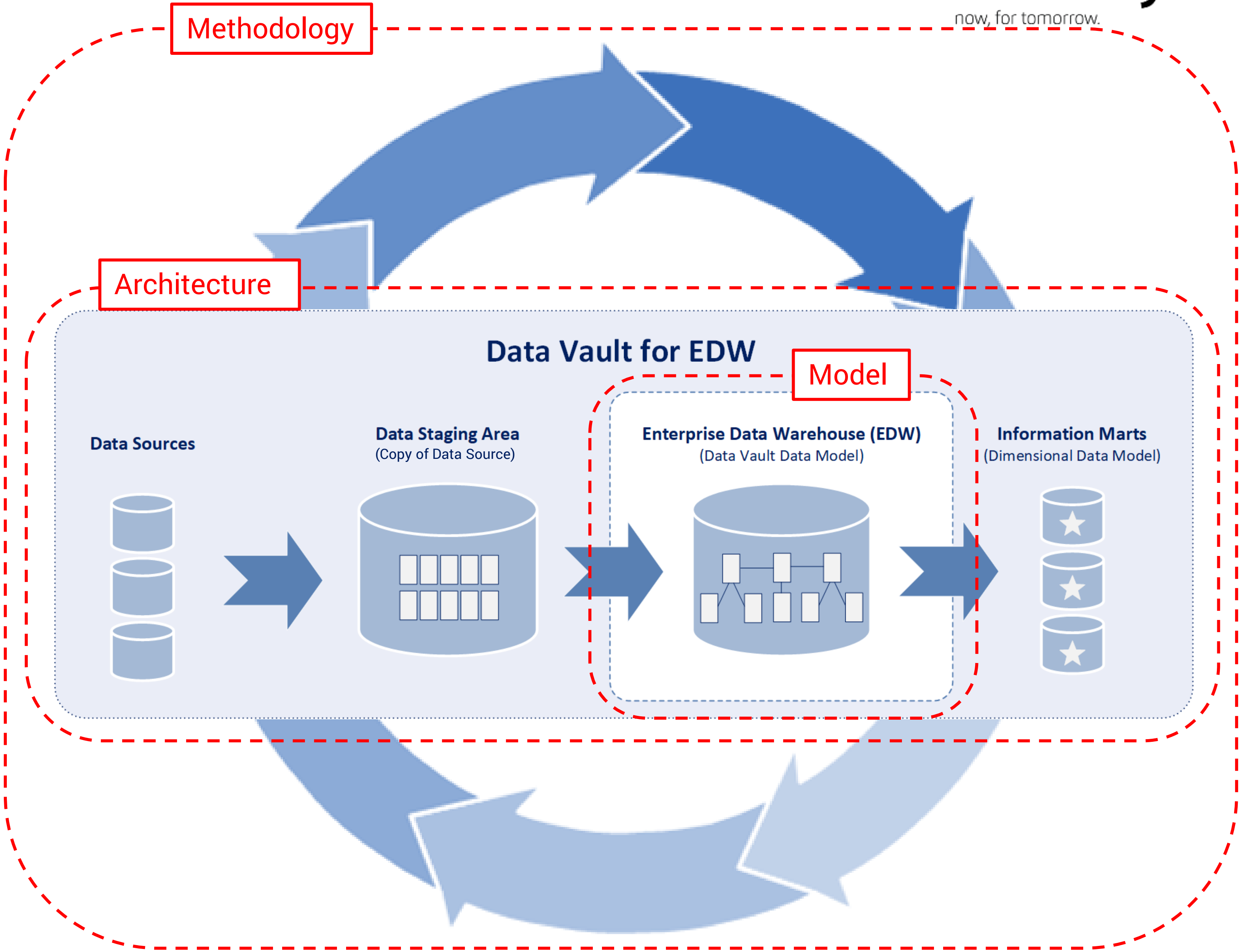
Created by Dan Linstedt, Data Vault 2.0 is a **System of Business Intelligence** containing the necessary components needed to accomplish enterprise vision in **Data Warehousing** and Information Delivery.

Data Vault 2.0 differs from 1.0 in that it is an entire system, and not just a model.

Data Vault 2.0

Data Vault 2.0 is a methodology, an architecture and a model. The most significant advantages of utilizing Data Vault are seen within the model, which is:

- Scalable
- Repeatable
- Auditable
- Adaptable
- Optimized Loading
- Platform Agnostic

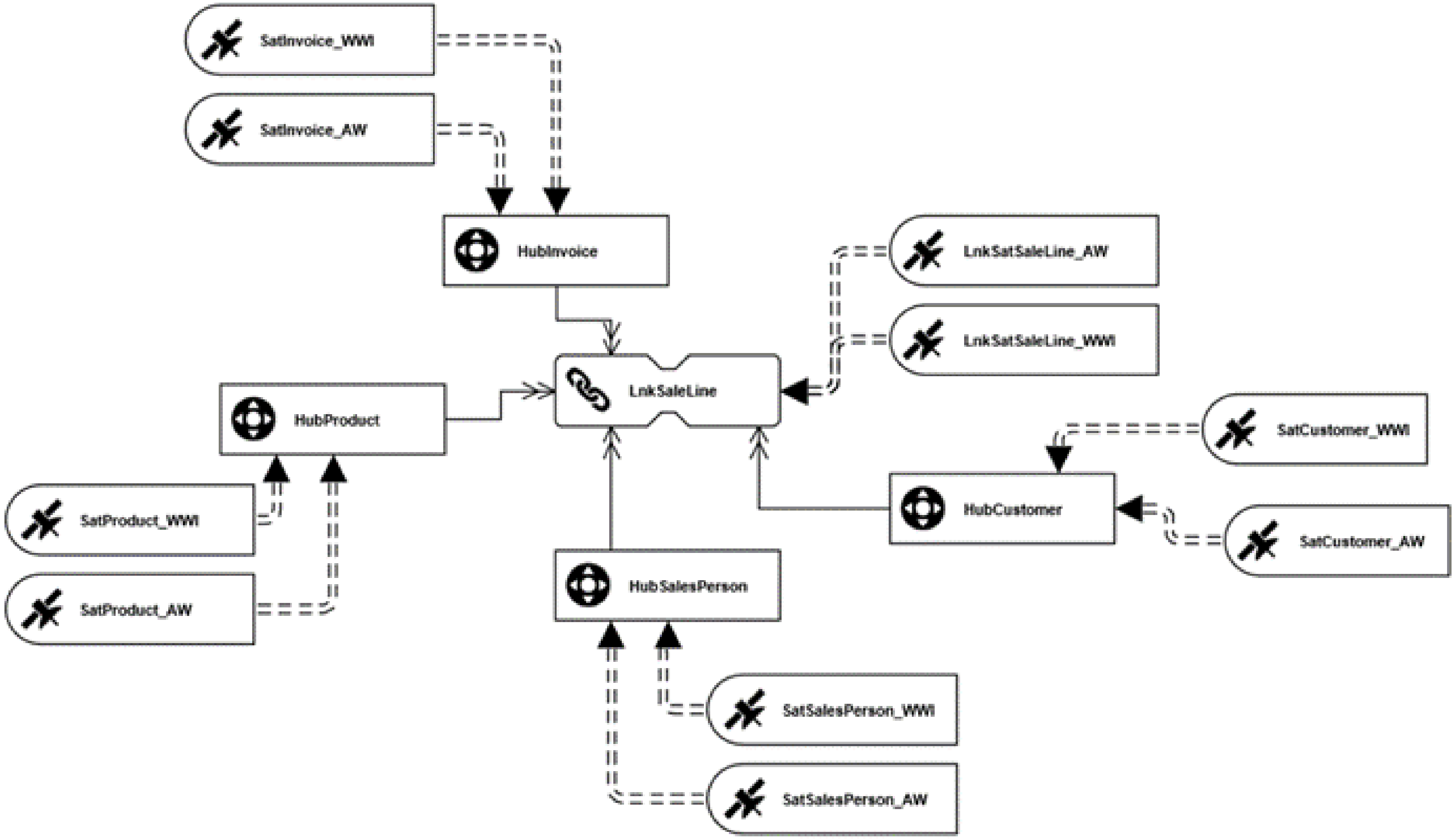


Model

Model Overview

Composed of three main structures:

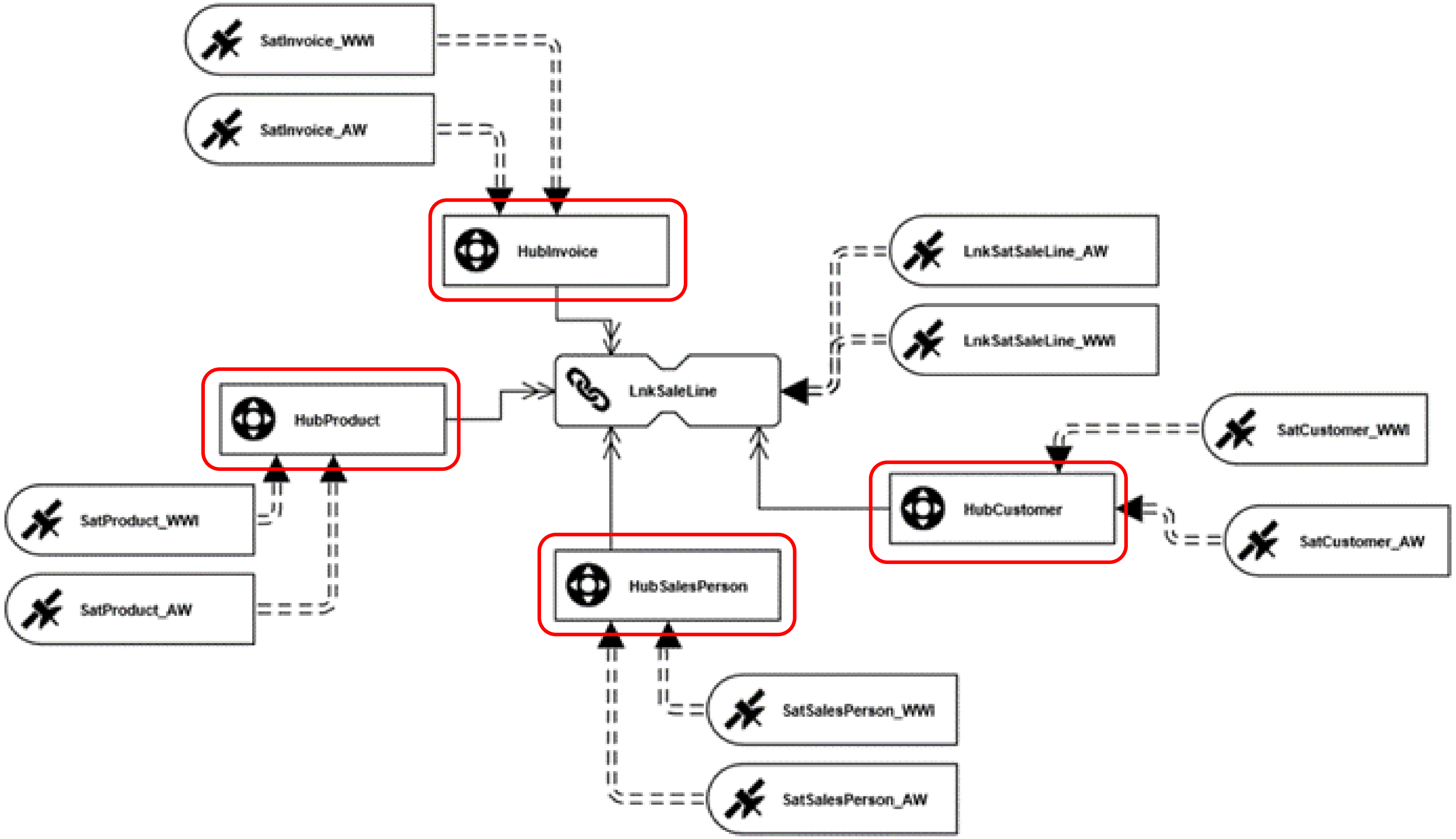
- Hubs
  - *Business Keys*
- Links
  - *Relationships*
- Satellites
  - *Context*



Model Overview

Composed of three main structures:

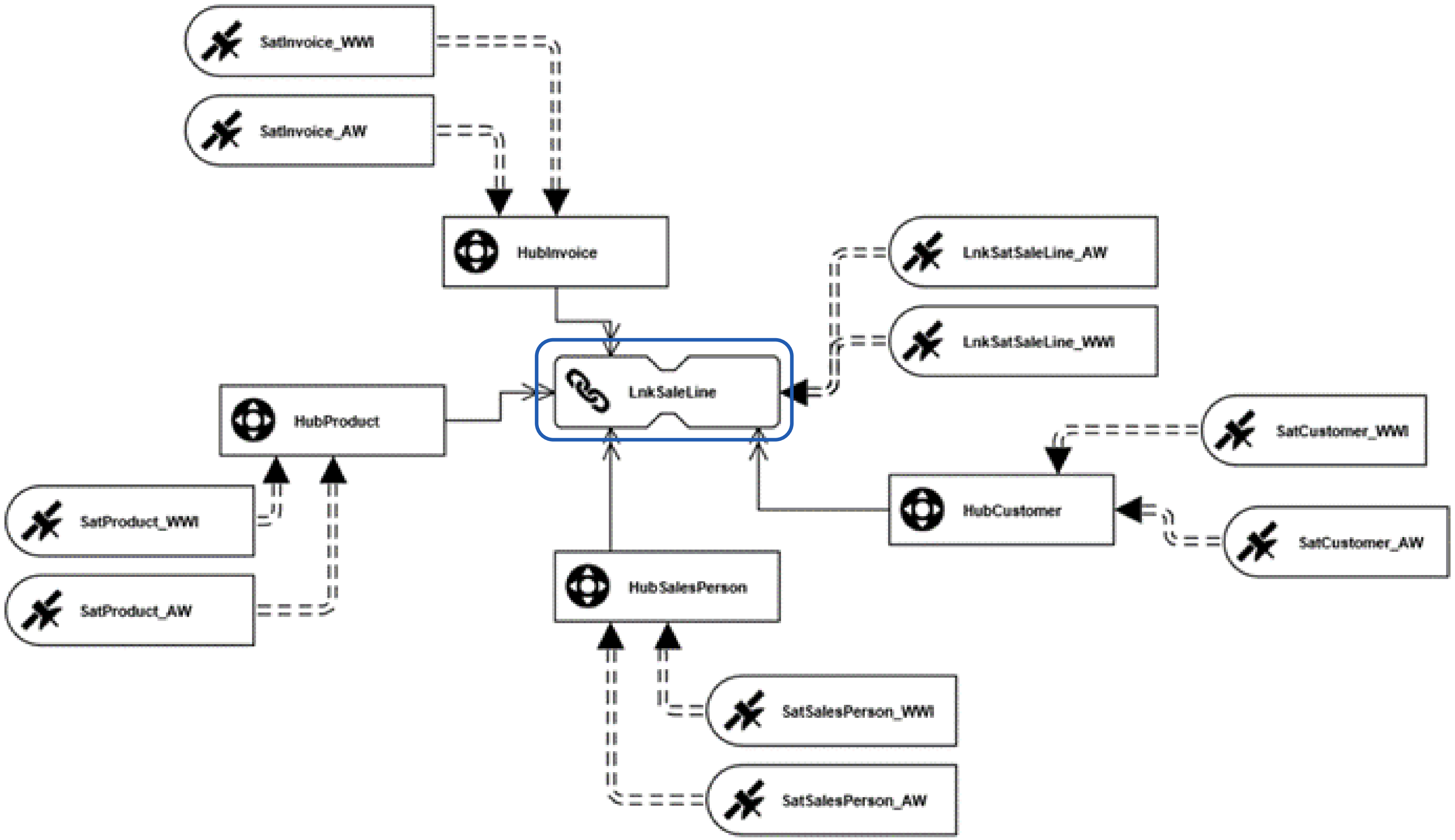
- **Hubs**
  - ***Business Keys***
- Links
  - *Relationships*
- Satellites
  - *Context*



Model Overview

Composed of three main structures:

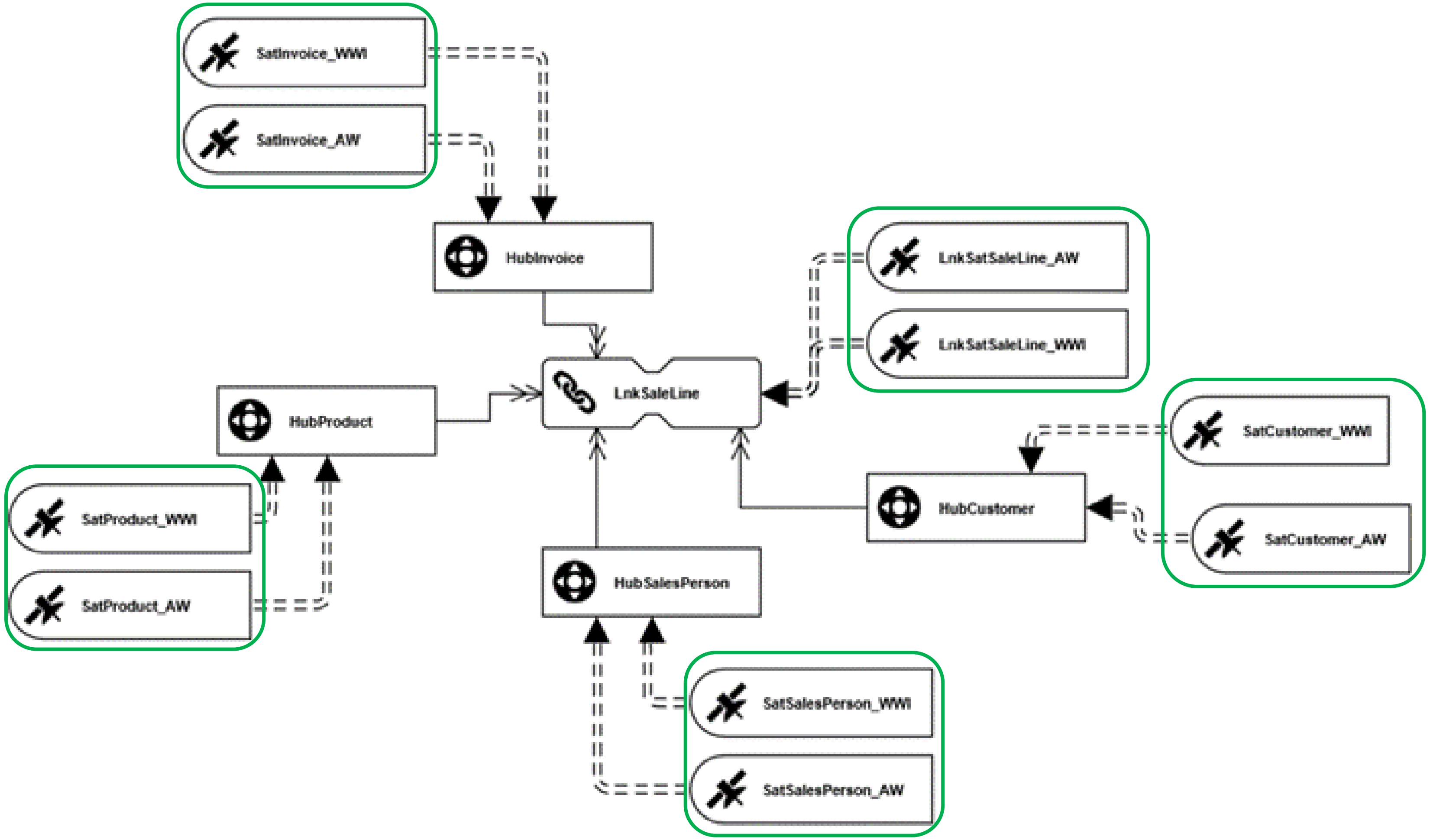
- Hubs
    - *Business Keys*
- Links
    - ***Relationships***
- Satellites
    - *Context*



Model Overview

Composed of three main structures:

- Hubs
  - *Business Keys*
- Links
  - *Relationships*
- **Satellites**
  - ***Context***



Entities

Hubs

What are Hubs?

Tables that consist of a collection of Business Keys

What are business keys

Keys that are supplied by users to identify, track, and locate information, such as a customer number, invoice number, or product number.

Business keys should be…

…unique

…at the same level of granularity

Column Name	Description	Constraints	Inclusion
HashKey	HashKey generated from the Business Key	PK	Required
LoadDatetime	Load Date & Time		Required
RecordSource	Specifies the source system from which the key originated		Required
BusinessKey	Business defined business key	UQ	Required
LastSeenDate	*Optional* Date a record was last included on a data load		Optional

Hubs

What are Hubs?

Tables that consist of a collection of Busin

Hash Values

Hash values are a fundamental component of Data Vault modelling. They are generated using system functions as data is loaded into the data vault. Hashes reduce dependencies, allow for quicker joins between tables, and allow for fast comparisons to detect changes in data.

What are business keys

Keys that are supplied by users to identify, track, and l ation, such as a customer number, invoice number, or product number.

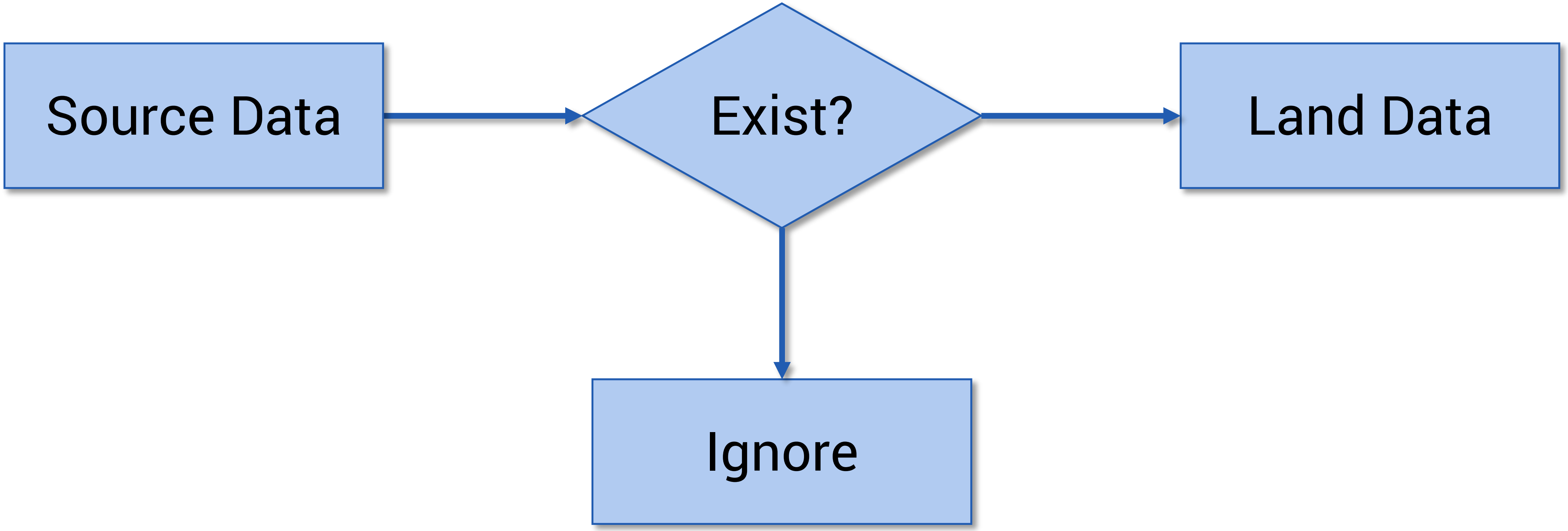
Column	Description	Constraints	Inclusion
HashKey	HashKey generated from the Business Key	PK	Required
LoadDatetime	Load Date & Time		Required
RecordSource	Specifies the source system from which the key originated		Required
BusinessKey	Business defined business key	UQ	Required
LastSeenDate	*Optional* Date a record was last included on a data load		Optional

Business keys should be…

…unique

…at the same level of granularity

Hub Loading Pattern


Source Data
Exist?
Land Data
Ignore

Links 🔗

What are Links?

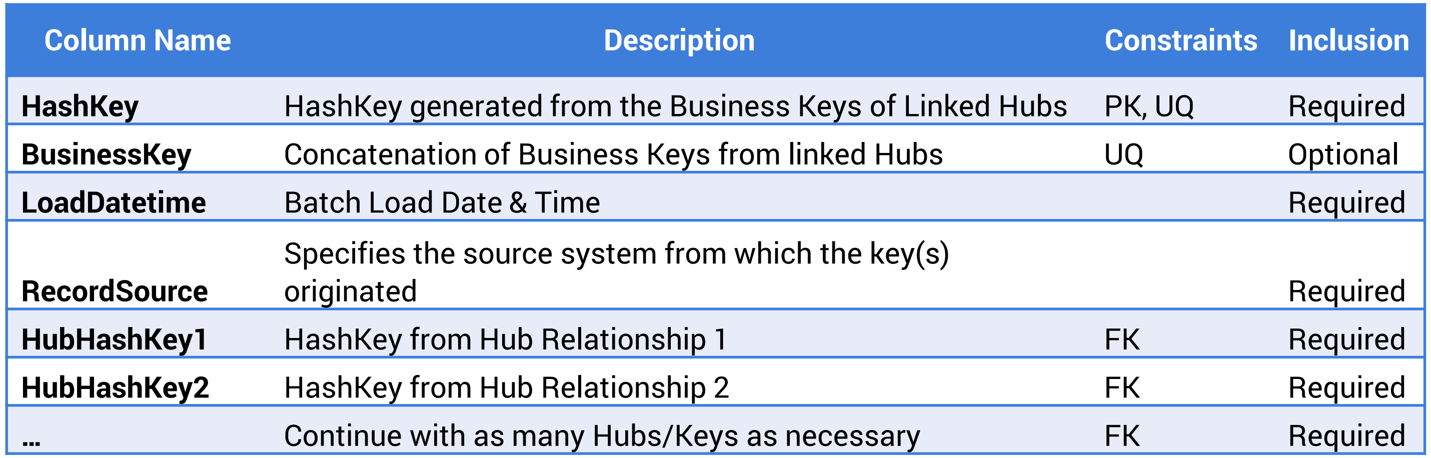
Tables that show the relationships between business keys (aka Hubs). Their level of granularity is determined by the number of hubs they connect and they are non-temporal. When thinking of a traditional star schema, links are often associated with fact tables.

Links…

- Do not show effectivity
- Only support inserts

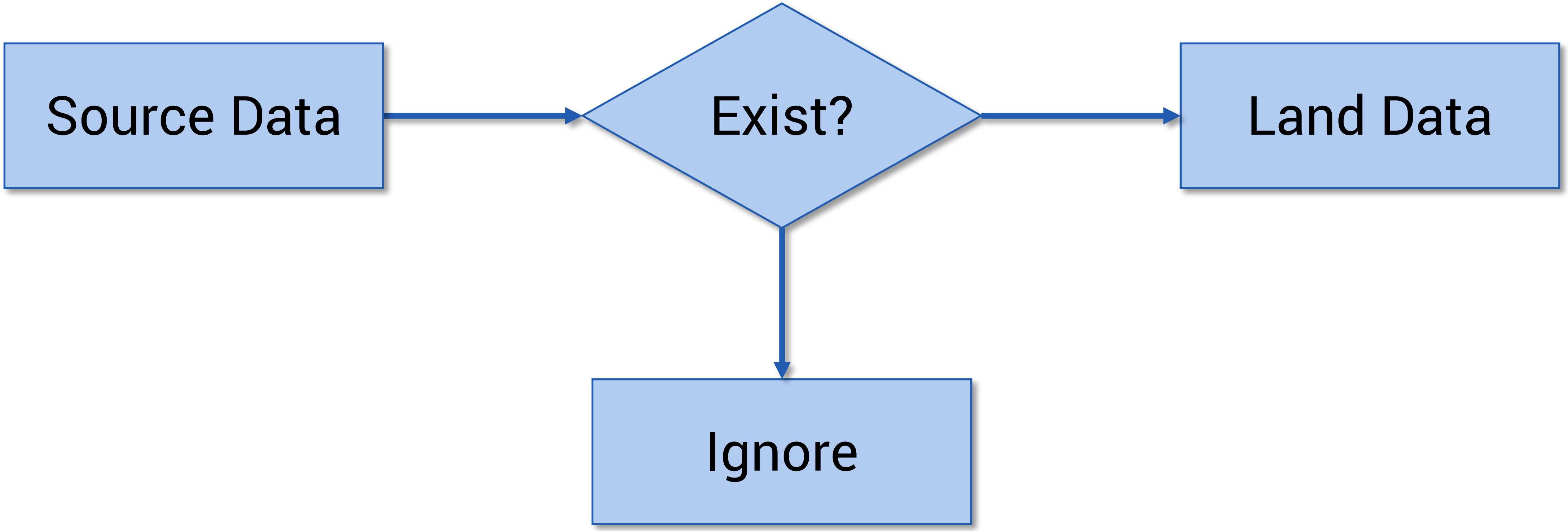
Different types of links can be used to make models more flexible

- Standard Link
- Same As Link
- Hierarchical Link
- Non-Historized Link (aka transactional)



Column Name	Description	Constraints	Inclusion
HashKey	HashKey generated from the Business Keys of Linked Hubs	PK, UQ	Required
BusinessKey	Concatenation of Business Keys from linked Hubs	UQ	Optional
LoadDatetime	Batch Load Date & Time		Required
RecordSource	Specifies the source system from which the key(s) originated		Required
HubHashKey1	HashKey from Hub Relationship 1	FK	Required
HubHashKey2	HashKey from Hub Relationship 2	FK	Required
…	Continue with as many Hubs/Keys as necessary	FK	Required

Link Loading Pattern 🔗



Satellites 🛰

What are Satellites?

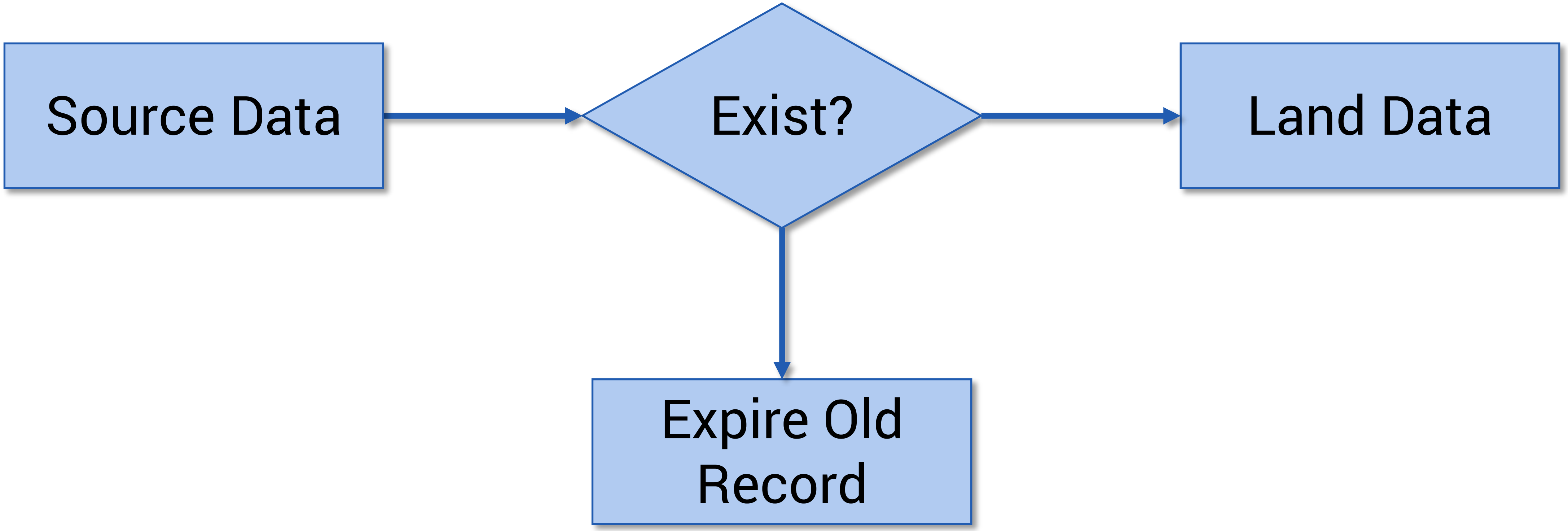
Tables that provide context to the business objects and relationships described in Hubs and Links. Each satellite is connected to only one Hub or Link, but a Hub or Link can have multiple satellites.

Key notes on Satellites

- One per source system
- Stores all context
- Stores all history
- Delta driven, similar to slowly changing dimension
- EndDate is **only** attribute that is updated
- Most flexible construct

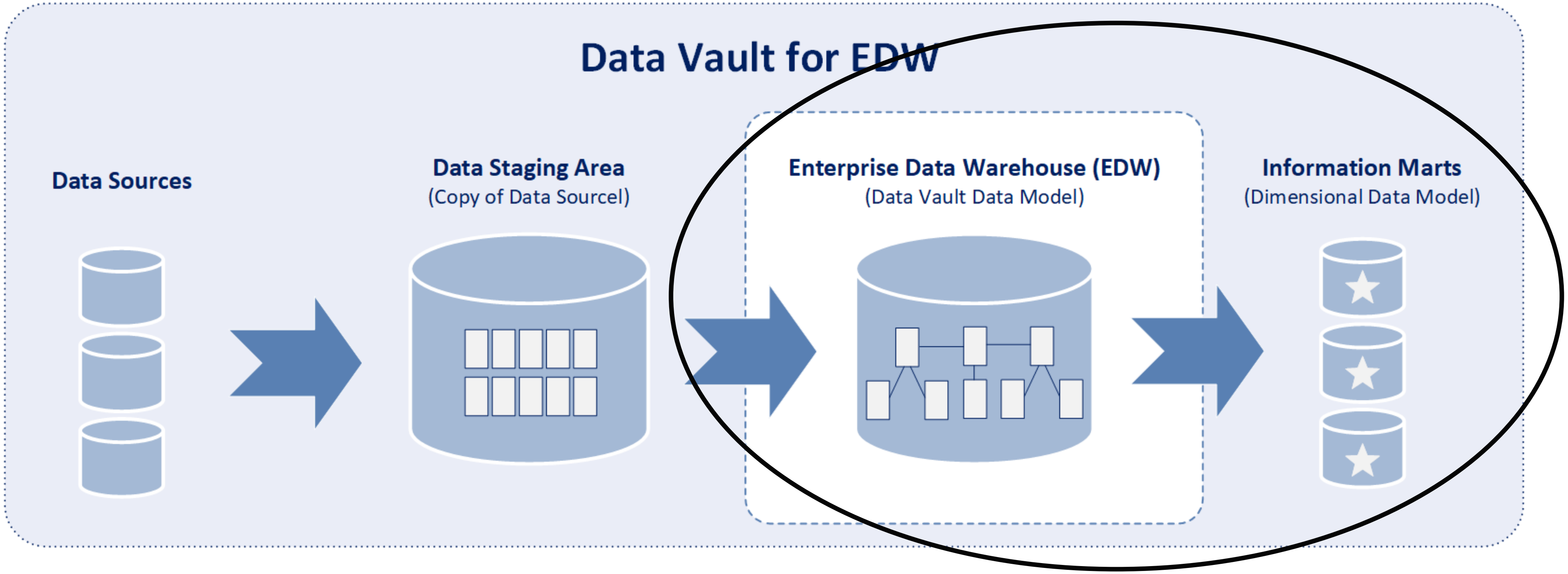
Column Name	Description	Constraints	Inclusion
HashKey	HashKey from the parent Hub or Link	PK, FK	Required
LoadDatetime	Batch Load Date & Time	PK	Required
EndDatetime	Load Date & Time the record became inactive		Required
RecordSource	Specifies the source system from which the key(s) originated		Required
HashDiff	Hashed value of all attribute data		Optional
ExtractDate	Date data was extracted from source system		Optional
…Attributes	Attribute columns. Number and type will vary.		Optional

Satellite Loading Pattern 🛰



Information Marts

Virtualized Information Marts

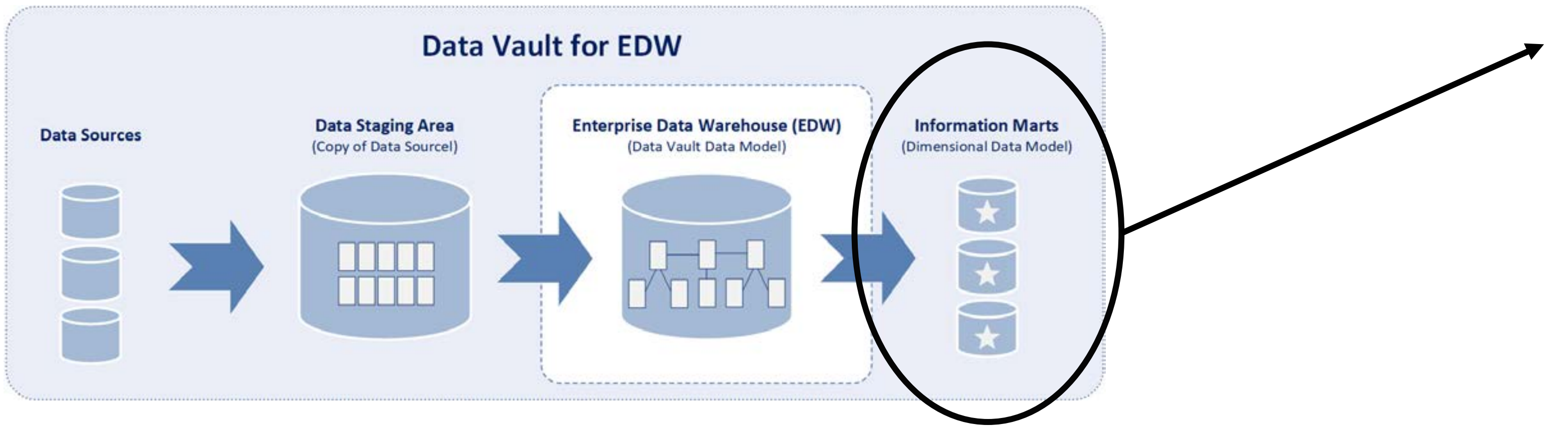


Virtualized Information Marts

‘Kimball’ Style Star Schema

- Information marts should be virtualized using views until such time as performance dictates otherwise

- Application of business rules (aka “Soft Rules” happens here)
  - aggregations, calculations, etc.
- Layer presented to business users & BI tools



Dimension 1	
PK	Id 1
	Attribute 1 Attribute 2 ....

Dimension 4	
PK	Id 3
	Attribute 1 Attribute 2 ....

Fact Table	
PK PK PK PK	Id d1 Id d2 Id d3 Id d4
	Measure 1 Measure 2 ....

Dimension 2	
PK	Id 1
	Attribute 1 Attribute 2 ....

Dimension 3	
PK	Id 3
	Attribute 1 Attribute 2 ....

Advantages

bakertilly
now, for tomorrow.

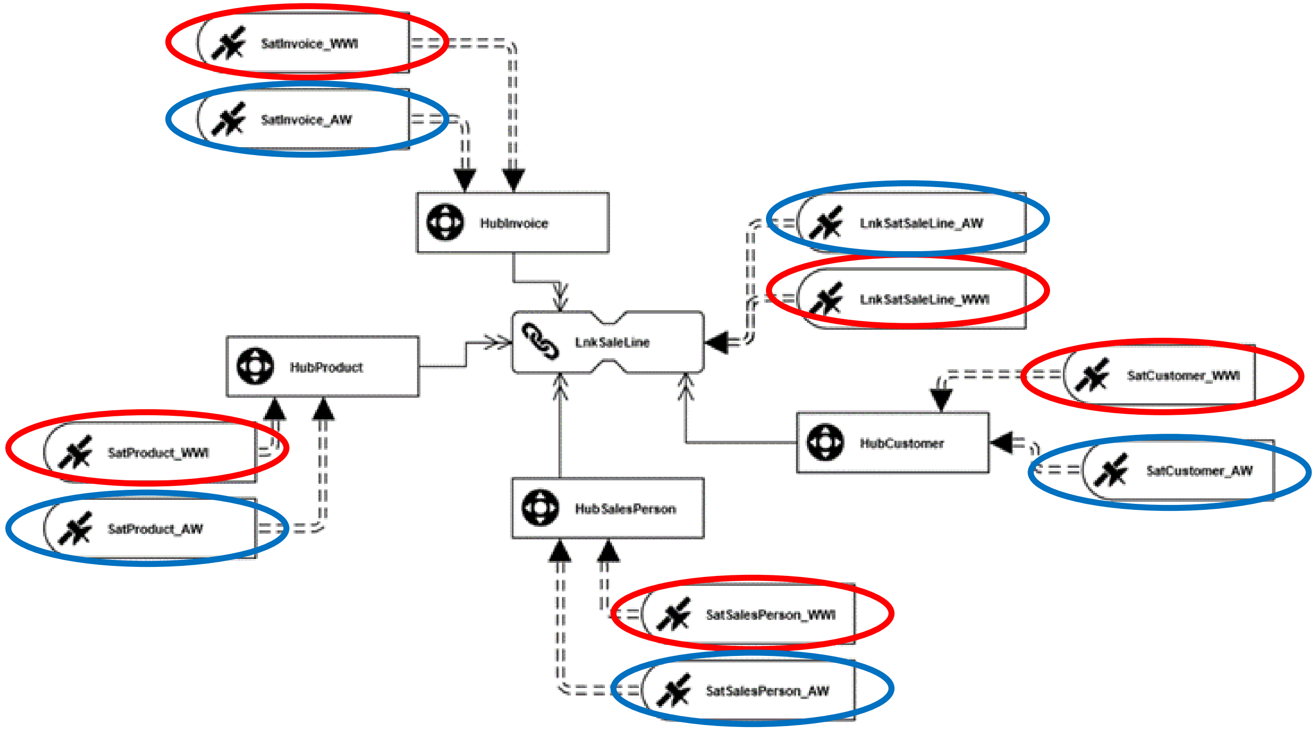
Advantages

Scalable

One of the biggest advantages of a Data Vault model is the ability to scale up or down quickly – a huge asset for companies going through growth or transition periods.

Because satellites are source system specific, adding sources is as easy as adding satellites. No updates to existing satellite tables are required.

Note: Updates may be necessary for views in the Information Mart



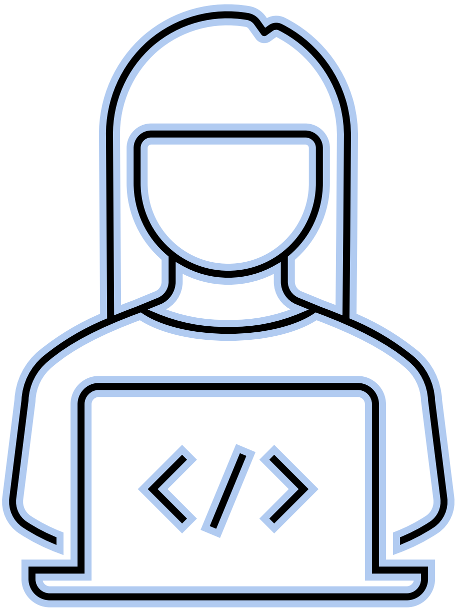
Advantages

Repeatable

Three main entities – Hubs, Links, Satellites – all follow the same pattern.

Scripts to build tables or run ETL processes can be automated based on these patterns and metadata.

A number of services and programs exist to quickly automate these processes



Advantages

Auditable

Built on core principle that data is never deleted, and all data is loaded in its original format.

Record source column in entities allows for tracking back to source system.

Tracking of business keys and separation of business keys (hubs) from context (satellites) allows for easier compliance with GDPR & similar data protection regulations.

Advantages

Adaptable

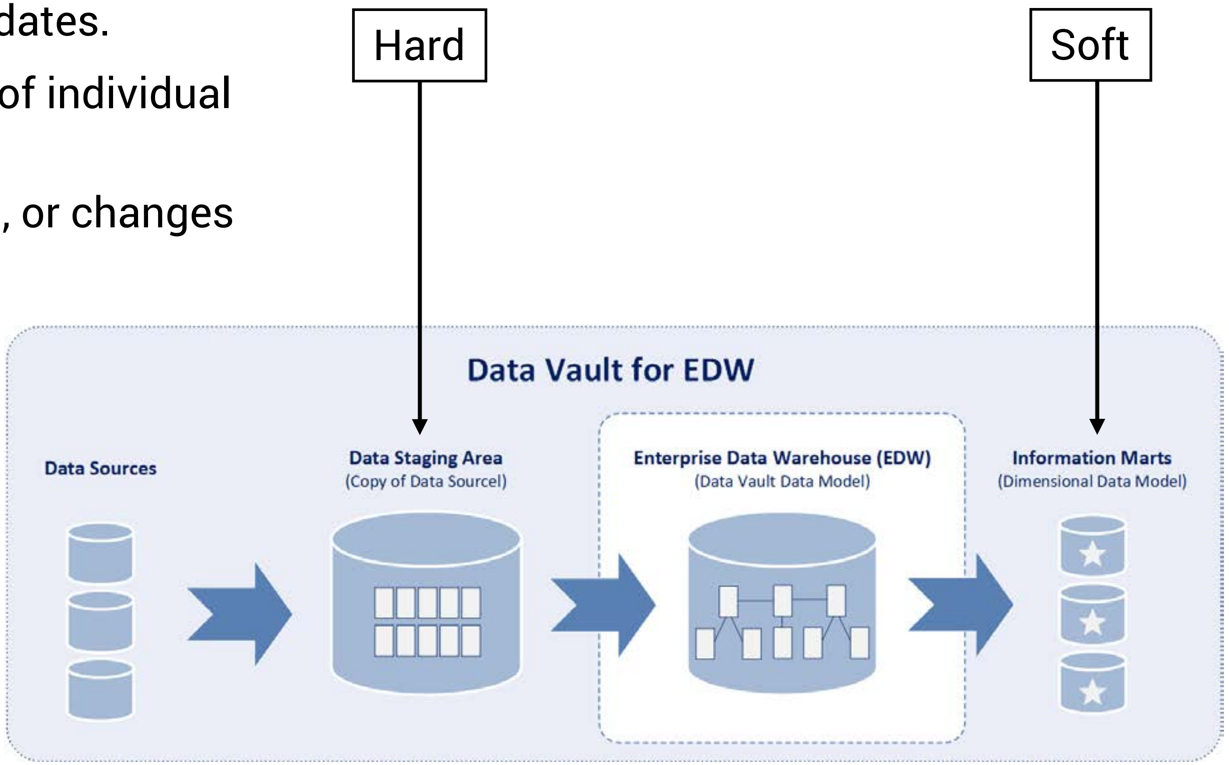
Separation of hard and soft rules allows for quicker updates.

Hard: Any rule that does not change content of individual fields or grain

Soft: Any rule that changes or interprets data, or changes the grain (turning data into information)

Changes in business logic requires no change to ETL processes – only updates to virtualized Information Mart layer.

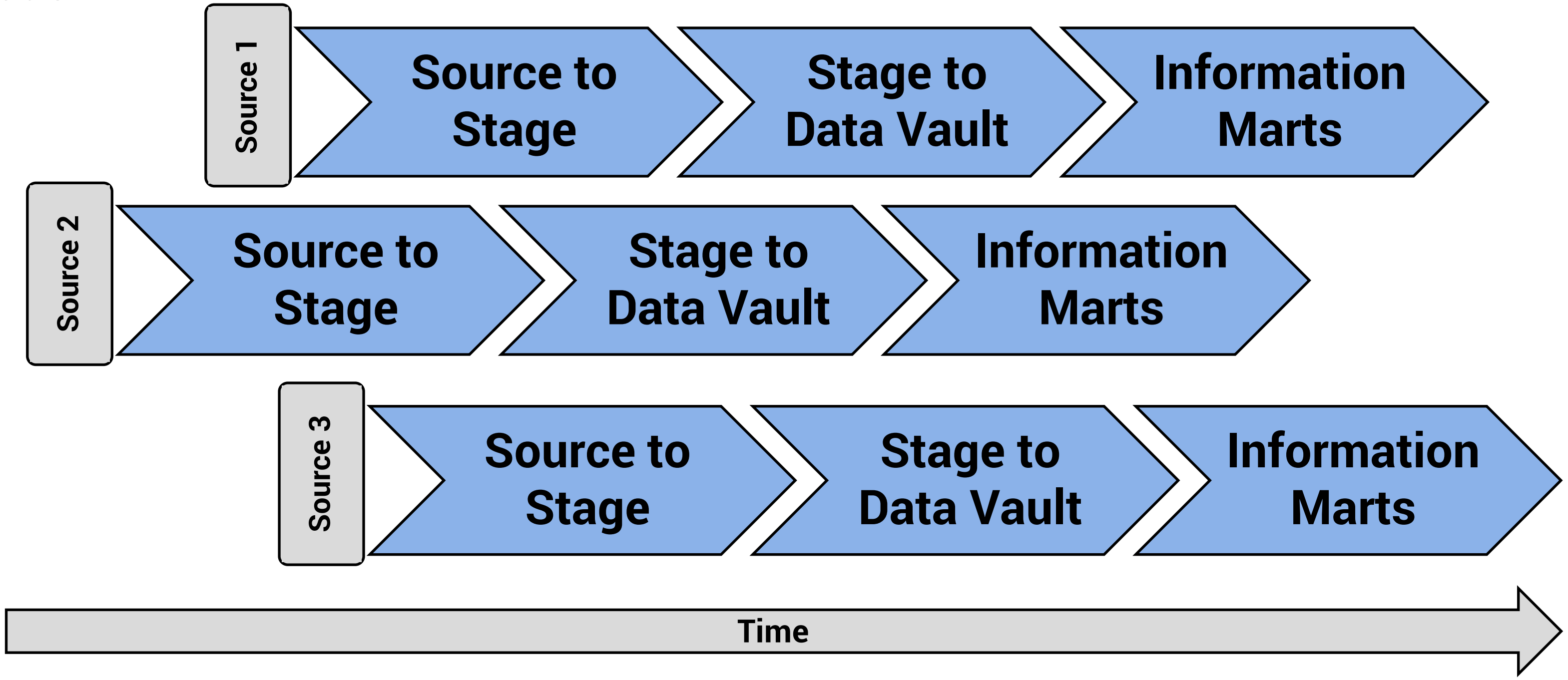
Fits within an Agile framework.



Advantages

Optimized Loading

1. Decreased process complexity
2. Decreased amount of data being processed
3. Increased opportunities for parallelization



Advantages

Platform Agnostic

A Data Vault architecture and model can be built on many different platforms – both on prem and in the cloud

Initial design for Data Vaults were to handle batch processing, but patterns also now exist for handling streaming data

Microsoft® SQL Server®

Azure

Google Cloud

aws

snowflake

Disadvantages

Disadvantages

Training & Complexity

Because the Data Vault is not a well known modeling technique, hiring or training staff may be an issue or expense

Data vault models have a tendency to become very large and complex, which may be a daunting process for those new to the technique

Wrap Up

Identifying Candidates for a Data Vault

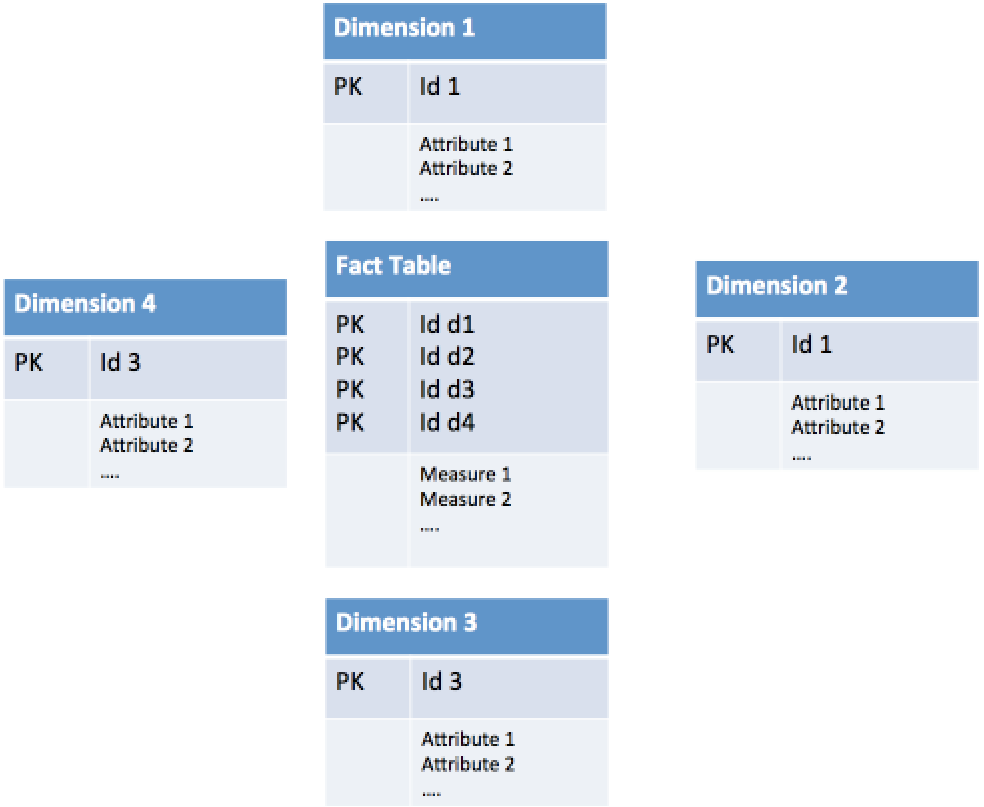
Although the Data Vault provides a tremendous number of advantages, this model isn’t necessarily the right approach for all clients.

Things to look out for:

- Desire to integrate multiple source systems
- Looking to scale
    - Mergers & Acquisitions or organic growth
- Using (or wanting to use) and agile approach to development
- Staffing, or ability to hire, to support a Data Vault

Questions?

Star Schema



Dimension 1	
PK	Id 1
	Attribute 1 Attribute 2 ....

Fact Table	
PK	Id d1
PK	Id d2
PK	Id d3
PK	Id d4
	Measure 1 Measure 2 ....

Dimension 2	
PK	Id 1
	Attribute 1 Attribute 2 ....

Dimension 3	
PK	Id 3
	Attribute 1 Attribute 2 ....

Dimension 4	
PK	Id 3
	Attribute 1 Attribute 2 ....