eBook

A Confluence of Metadata

Snowflake is a single, fully managed platform that powers the AI Data Cloud. Snowflake securely connects businesses globally across any type or scale of data to productize AI, applications and more in the enterprise. www.snowflake.com

VaultSpeed is the only solution that lets you automate every step of your cloud data warehouse, lakehouse or mesh. Setup, maintenance and beyond. vaultspeed.com

Data Vault is a system of Business Intelligence containing the necessary components needed to accomplish enterprise vision in Data Warehousing and Information Delivery. The method is designed to be non-destructive to change and provide long-term historical storage of data coming in from multiple operational systems.

Written by Patrick Cuba
Senior Solutions Architect, Snowflake

About the author

Patrick Cuba joined Snowflake as a Senior Solutions Architect with over 20 years of experience in Data and Solution Architecture. Patrick is also a published author and a frequent contributor to Data Vault 2.0 thought leadership worldwide.

At Snowflake Patrick guides customers through their Snowflake journey to get the best out of their Cloud Data platform with best in breed methodologies and practices, especially on building Data Vaults on Snowflake itself! Patrick earned a Bachelor's degree in Information Technology at the University of Johannesburg and is a certified Data Vault 2.0 practitioner.

Table of Contents

Introduction

Through years of experience in the IT and Data Analytics field we know how difficult it is to gain **trust** in your enterprise data; we also know that if you have achieved that trust a sustainable data-driven organisation must also maintain that trust because as the adage goes, “Trust takes years to build, seconds to break, and forever to repair.”

> *"Trust takes years to build, seconds to break, and forever to repair."*

As your enterprise scales, your data footprint will scale too, as well as the amount of compliance and regulation you are potentially exposed to. You can no longer rely on handcrafted data pipelines but must now rely on delivering repeatable data patterns based on automation. The scrutiny your data is exposed to is both internal from your business stakeholders and external from your valued customers.

You need a dedicated team of business and data professionals working in tandem to ensure accurate information delivery while sticking to industry-based compliance. But as every business knows, industries evolve and so does the need to keep your data accurate and compliant.

It takes metadata to manage data

Gartner believes the future of data-driven organisations is the adoption of **Data Fabric** as a style of information delivery.

In their own words, “data fabric is an emerging data management design for attaining flexible, reusable and augmented data management (i.e., better semantics, integration and organisation of data) through **metadata**. Data fabric augments data analysis by simplifying data understanding for business users and enabling them to consume data with confidence.”

Metadata is *easily* interpreted as the ‘data about data’, the kind of information that can be classified as metadata includes information about technical and business processes, data rules and constraints, logical and physical data structures.

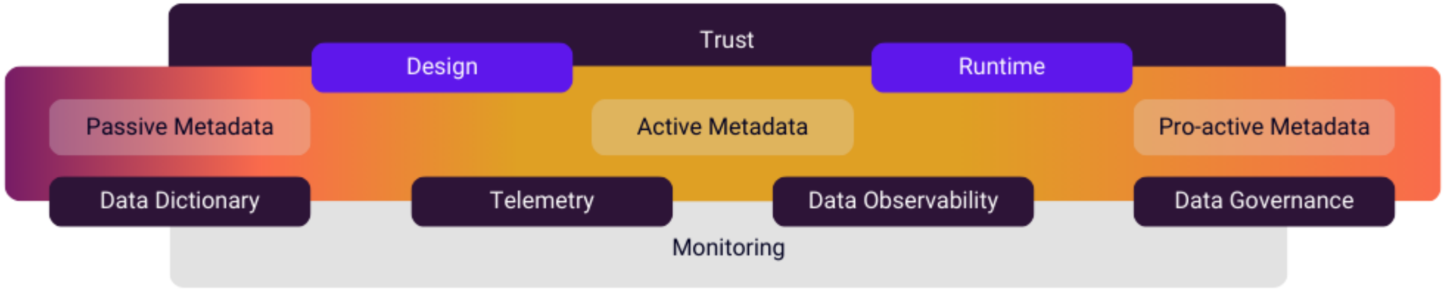
According to **DAMA**, reliable metadata helps

- Increase confidence in data providing context and enabling measurement of data quality and improves impact analysis.
- Increases the value of **strategic information** and improves operational efficiency as well as ensuring regulatory compliance.
- Prevents the use of out-of-date data, improves communication between data consumers and those data professionals delivering that data.

> *"Metadata is crucial to Data Governance."*

DAMA segregates metadata into business, technical and operational metadata; Gartner further categorises metadata into **Passive** and **Active** metadata describing how the spectrum from passive to active inherently increases trust in that data. A parallel concept to plot on this spectrum is from **Design** to **Runtime** metadata.

Let's illustrate this spectrum below:



APIs are metadata and require their own well-described metadata to ensure the data it supports is used *correctly*. In the context of a **Data Mesh** a **data product** needs to be discoverable, addressable, understandable, trustworthy, interoperable, accessible, secure and **valuable on its own**. Data products require **data contracts** to establish a level of guarantees that the data product will support downstream data requirements to a tolerable standard, at least for those **critical data elements** (CDEs).

Source application data products live in a world based on support for individual business applications and with access to other APIs in a service mesh. What they lack is historical context across a business' application landscape.

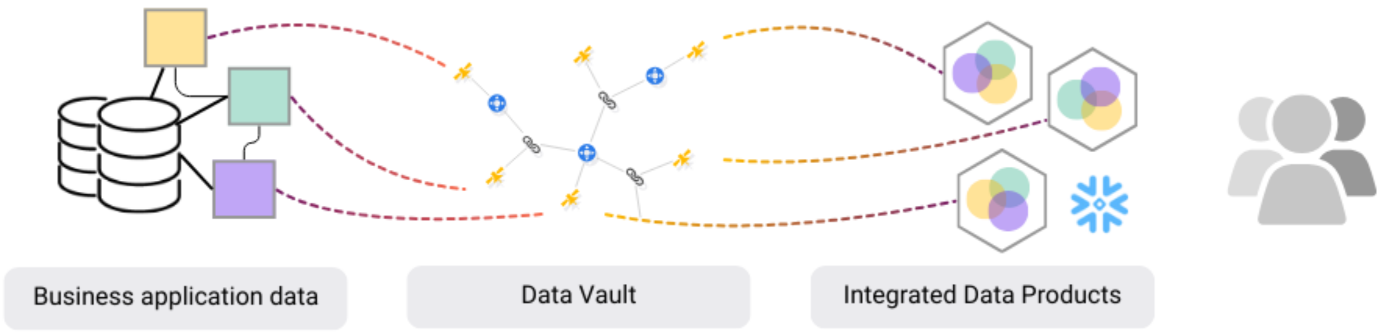
Integrated Data = New Data Products

Whether you're building a data mart, data warehouse, data lake or even a data lakehouse, one outcome remains desirable all data methodologies: predictability. This is in essence why we think Data Vault and its repeatable patterns in architecture, modelling and agility is exciting. Data Vault prescribes only three table types,

- **Hub** tables that house the **business keys** that are representative of the business entities an organisation care about.
- **Link** tables capture the business process **interactions** between business entities as relationships, transactions and business events.
- **Satellite** tables capturing the **true-change state** of those business entities and business entity interactions.

Three table types mean three data loading patterns, three testing patterns and repeatable patterns for **information delivery**. Each data vault table structure is equipped with standard *passive* metadata columns to track the lineage of how that record got there. We prescribe that each record in a data vault includes,

- The load timestamp of when the record made it into the data platform.
- The applied timestamp denoting when we *copied* the record from a source.
- The IDs denoting who and/or *what* process loaded that record.
- Any other metadata column data you need to store full record lineage.



As desirable as these outcomes are, the challenge for data engineering and enterprise data modellers the world over is that none of the application data (business process outcomes) comes in the form of hub, link and satellite tables! That is by design, every application data model is built to serve the architectural challenges for delivering that data in the form and scale required by that business need.

For business analytics, data is not easily understood when shaped in that form. Business application data must be *curated, cleansed* and *transformed* into a data model that is designed to be flexible to change, reusing the same table patterns we described above. In essence, to provide the business data in the form data analysts need, it must be *copied* onto an OLAP platform *like* Snowflake.

Business application data models may come in the form of *structured* relational tables, semi-structured content like JSON or XML as event or time-series data or even *unstructured* data as PDFs and emails! With such a variance in form and volume the act of copying this data into data vault structures requires **automation**.

The Key to Automation is Metadata

Bridging that gap between business applications and **an enterprise data model** is the need for metadata-driven data integration. Yes business applications and their data are diverse but the method to harness that data into data vault table types is repeatable and thus their metadata will be too.

> *"One person's data is another's Metadata"*

Vaultspeed as a data integration automation tool recognises this pattern and offers out of the box tooling for modelling application data models into targeted data vault artefacts without writing a single line of code. Vaultspeed does this with templates that cover a wide range of application data and change data capture patterns. The process itself is repeatable, template based and accomplished through 4 simple steps.

Step 1: Source metadata harvesting

Gain an immediate and thorough view of your business application's data model, enabling you to efficiently prepare it to align with your data vault model.

Step 2: Map metadata to Data Vault model

Tag the harvested metadata to customise Vaultspeed's suggested data vault model

Step 3: Generate template-based data pipelines.

Repeatable logic as SQL templates is generated ready to be orchestrated

Step 4: Deploy and orchestrate on Snowflake

Choose from Apache Airflow, Azure Data Factory, Matillion Native Scheduling or even Snowflake's own tasks to orchestrate and execute our generated SQL logic.

For the next business case, repeat from step one and integrate your new business problems into the same enterprise data model and in parallel. Let's not forget that your business applications are expected to evolve too, VaultSpeed will detect and manage those changes on your behalf; you simply authorize and merge those changes into that enterprise data model.

Passive, Active and Proactive Metadata

The core theme of this paper is metadata, and we need to leverage metadata for just one more task – guaranteeing the peace of mind for the business to function without worrying about falling foul to regulators and other reputational risks and any other data-related business risks. Yes, we must automate **data governance** purposely into information delivery.

Every query run on Snowflake has a query-id attached to its execution, this includes the execution patterns generated by Vaultspeed we run on the platform to load data vault artefacts, and the queries used to consume that loaded data. You will find a wide of array of aggregated and fine-grain telemetry and metadata of all actions within your Snowflake account, and by extension your Snowflake organisation.

We will broadly categorise this metadata into,

1. Passive Metadata

- Snowflake Information Schema a set of defined views and table functions based on the SQL-92 ANSI information_schema. The state change of all your information schemas is historized and aggregated into an account_usage set of views with a retention of 365 days.
- Access History and Object Dependencies to trace object lineage and impact analysis, Query_History giving a full trace of what was executed on the platform.
- Vaultspeed takes this a step further by storing the *history* of metadata changes deployed on Snowflake; this makes the deployed assets “replayable” at any point in time, insofar as the data this metadata is based on is still available.

2. Active Metadata

- Data Quality functions to periodically analyse your data assets and the ability to generate alerts when thresholds are breached.

Snowflake ensures your SQL query is
***Metadata-Rich** and **Secure** by default*

3. Pro-active Metadata

Now to ensure the peace of mind we alluded to earlier, we encourage the application of pro-active metadata, i.e. data governance functions supported by:

- Column-level security (CLS) in Masking Policies, Conditional Masking Policies, External Tokenization and Tag-based Masking Policies.
- Row-level security (RLS) in Row Access Policies.
- Query and Object Tagging beyond assigning policies but can be used to trace domain ownership of data assets too.
- Classification to detect privacy and semantic categories and assign Snowflake or your own tags to.

Snowflake pioneered other metadata-driven automation features its competitors cannot *easily* emulate. These features have been a mainstay of Snowflake since its inception,

- Zero-copy cloning – a metadata operation that copies the metadata *pointers* to the micro-partitions of the table you're cloning. It's effectively like taking an instant photo of a table and you have the option to take **time-travel** snapshots of individual tables, schema or databases. This operation effectively provides a live backup of your data in Snowflake.
- Secure Data Sharing – you can securely share your data in your Snowflake account by granting another Snowflake account live access to the data you choose to share within your organization or your partners. Another metadata operation which in reality is the only true implementation of real-time data.

VaultSpeed has taken a unique approach to metadata-driven automation too. Being a tool designed to leverage metadata to essentially mould your data into your desired artefacts, VaultSpeed recognizes that there are signature artefacts you will define and use repeatedly. Out of the box these signature artefacts are your typical hub, link and satellite tables variations as well as the same signature columns that each of these signature objects use.

Now, let's say you have a pattern or calculation that Vaultspeed does not include in its standard offering, you can reuse the aforementioned signature objects to develop your own templates through the new Vaultspeed Studio no-code interface.

That's right, the studio offers your modellers the ability to develop and reuse their own metadata-driven templates over and above the standard metadata-driven templates offered; and that extends to building information delivery patterns such as your traditional Kimball-styled facts and dimensions.

Your imagination is your only limitation!

A Confluence of Automation

Snowflake offers best in class analytical data capabilities securely on the cloud; and as a service. Automating your data integration and data modelling patterns into Snowflake with Vaultspeed accelerates your ability to deliver data-driven business value at scale. Why develop the ingestion and consumption patterns hundreds of other data professionals have done before if you can simply utilise that automation out of the box and without writing a single line of code?

Analytic Data is built on repeatable SQL patterns

Tabular data is universally easier to rationalise between business users and data professionals. Boyce-Codd cemented the relational semantics needed to efficiently support business logic in the form of tabular data through constraints. Data Vault established just three repeatable tabular patterns and therefore just three loading and integration patterns for all your analytical data. Vaultspeed bridges that gap between the various forms of business application data into those three data vault table types. With just three table types as your base for information delivery, Vaultspeed also supports the templated SQL to consume that data downstream.

Analytic Data lives on repeatable Data Management patterns

As custodians of customer data, we as data professionals must provide securable data-driven solutions to guarantee the data we work with is trustworthy and secure for our internal stakeholders and external customers. Snowflake is secure by design and by using the supported data governance features Snowflake provides customers can ensure no data is inadvertently exposed to unauthorised users or processes. Our automation patterns deployed by Vaultspeed and built as data vault structures ensures your analytical data can adapt as quickly as your business adapts.

"There is no AI without a Data Strategy"

- Sridhar Ramaswary, Snowflake CEO

Enterprise Data Models are your **corporate memory** and what better data model is there than the data vault to ensure your enterprise data model reflects your **Business Architecture** and in the parallel scale your business demands. The more you can automate data ingestion into your corporate memory the less costly it is because you are utilising years of experience in delivering the SQL patterns Vaultspeed architects have already included in their suite of templates.

Yes, the industry today has turned its data budgets to Generative AI; but why are most data-driven organisations still struggling with data integration to this day? Surely your data professionals would rather spend less time figuring out the integration problems Vaultspeed have solved hundreds of times before?

References

- How Data Fabric Can Optimise Data Delivery, https://www.gartner.com/en/data-analytics/topics/data-fabric
- Data Management Body of Knowledge, 2nd edition
- The Meta-Data Fiasco, https://tdan.com/the-meta-data-fiasco/18502
- 'We Kill People Based on Metadata', https://www.nybooks.com/online/2014/05/10/we-kill-people-based-metadata/

Visit our site

vaultspeed.com

Contact sales

sales@vaultspeed.com

Book a demo

vaultspeed.com/book-a-demo

Join our community

community.vaultspeed.com

VaultSpeed is the Automated Data Transformation solution of choice for leading companies. By combining automation templates, a data modeling GUI and a metadata repository in one platform, VaultSpeed helps businesses improve delivery and maintenance of theirycloud data warehouse or lakehouse. With offices in London, Seattle, Leuven and Vilnius, VaultSpeed works with the likes of HDI, Olympus, Eurocontrol or Bleckmann.

VAULTSPEED