VAULTSPEED

How to choose your automation solution

Finding the right solution to automate your data warehouse, lakehouse or mesh, for faster analysis and data-driven decisions

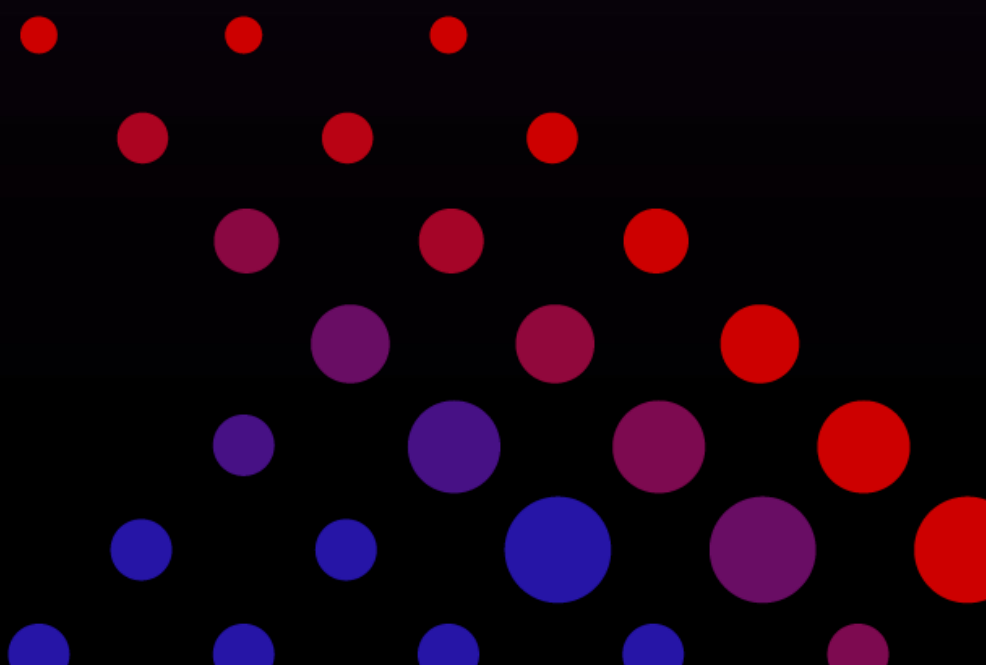

Table of Contents

Why you need to automate data transformation

Dealing with more diverse and dispersed data than ever before

Organizations deal with more varied and dispersed data than ever before.

The most crucial data division driver is the rise of microservices and SaaS applications, each serving specific functions within modern organizations. All of them are prone to change: new releases, new instances, new services, migrations to other platforms, discontinuation of service, and so on.

It is a significant challenge to secure the continuous integration of these data from different sources and vendors into one centralized repository, even for companies with ETL mappings.

A data transformation automation solution is designed to automate the repeatable manual tasks for ingesting data, managing data with different schema, and creating views for extraction throughout the lifecycle of a data warehouse, lakehouse or mesh.

Going beyond data integration

Automated data transformation, done the right way, should automate as many stages as possible – from loading data to presentation - and free up time so data engineers can focus on value-creating tasks.

This guide will help you scope your evaluation so you can choose the right solution for your business now — and three years from now. You will evaluate seven significant areas: scope, standardization, productivity, agility, output quality, extensibility, and operational costs.

Scope

Automating as many stages as possible

Automation needs to cover breadth and depth, automating multi-source data integration, data modeling and ETL/DDL code generation but also integrating and modeling different data from the various microservices and SaaS applications that each serve particular functions.

Think of a CRM tool, ERP system, or IoT stream: they all produce data about specific business processes.

Metadata ingestion is a prerequisite for integrating the tables, primary keys, unique keys, foreign keys, attributes and their data types (text, data, numbers, boolean) specific to each source.

Tech stack parametrization is what makes automation work in any company. No-code automation solutions enable data teams to select and specify the technologies and versions of their sources, target platform, CDC or ETL tools without coding any integration logic.

Parametrization applies to set the data integration parameters such as case-sensitive data comparison, loading logic, data quality settings, schema name options and many more.

Companies that want to set up sources as streaming sources should check if that feature is included.

Everything is a repeatable pattern

Objects such as computed satellites, dimensions, facts, flat tables, etc., require more customization in code. This is where data engineers can excel. The automation paradox, however, indicates a trade-off between logic's repeatability and its customization level. That paradox is no longer valid.

Modern tools – that adhere to Data Vault standards - offer advanced users the possibility to apply automation to company-specific logic designed for presentation layers. They are allowed to build their own templates to duplicate the logic coded for objects such as computed satellites, links, custom PITs, facts, and dimensions.

Standardization

Follow Data Vault 2.0 standards

The Data Vault 2.0 methodology has rapidly become the automation standard because it provides standardized structures.

Data Vault 2.0 modeling breaks down and consolidates source data into three core components: hubs (unique entities crucial to the business such as 'customer,' 'product,' 'store'), links (relationships between entities), and satellites (all the properties and history of an entity).

All hubs look alike, as do all links and satellites. This makes it possible to use metadata-driven approaches to automate the Enterprise Data Warehouse, Lakehouse or Mesh. The metadata describes the structure of the incoming data set and the organization's business conceptual data models. It identifies business keys, their relationships, and descriptive data to generate the CREATE TABLE and INSERT statements for all entities using templates that provide a customizable target structure.

Using templates makes it possible to generate the Data Vault model based on the same metadata to various target databases and adjust the templates to the organization's needs.

Good templates are, therefore, the foundation of a high-quality EDW. The better the templates and metadata quality, the better the generated solution's quality. The best automation tools provide built-in templates to integrate and model different data from the various sources and technologies that organizations have today.

Why you need a Data Vault 2.0 architecture

Data Vault is more than just a methodology or model. It also provides the architecture to help you implement a central data repository that is both efficient and focused on business. Keep in mind that although there are many ways to combine and implement the methodology, model, and architecture of the Data Vault system, it's essential to abide by the standard recommendations when building a Data Vault.

If Data Vault standards are not met, problems such as rework, a drop in performance, and mix-ups within the data teams will quickly arise. So it might be a good idea to check which automation vendor has been certified by the Data Vault Alliance, the official Data Vault body.

Productivity

Accelerate development

Any central data repository requires a great deal of upfront design and development. A lot of manual work goes into data modeling, data mapping, metadata management, etc.

All automation solutions promise to collect the right data in the right way. But we strongly suggest identifying in detail which automation solution will allow your data teams to do the work faster and more efficiently. Apply reverse engineering by looking at how many lines of code and how much data modeling time each automation stage or feature will exempt your teams from.

Choose no-code automation

Look in particular for no-code automation solutions that are designed to free teams from having to code templates themselves to make automation work for their specific technology stack. That task alone can set your company back months before automation kicks in.

And it is always a true statement that the more stages a tool can automate, the shorter your development stages will be.

Agility

Technology & data changes will happen

Technical environments tend to change. When sources are added or technologies get upgraded, data teams ideally only have to adapt parameters and incorporate the changes in their existing target model without starting anew.

The level of abstraction that Data Vault modeling brings is specifically designed to deal with source changes without the almost mandatory rework that goes with it. On top of that, Data Vault captures the history of all source changes, which lets companies go back in time.

Do look at out-of-the-box source version management as well to support agility. Tracking different versions is not just a matter of tracking the model changes. It would help if you also accounted for the effect these changes may have on data loss.

Data migrations scripts are helpful if a change in the source data model also requires a change in the target model. Examples are migration from single to multi-master hubs, data type changes, satellite split, etc.

People and opinions change too

The human factor plays a key role as well. Automation and standardization done right guarantees that every new profile that joins your data team will build on what has been done before.

A guided setup is just essential because it forces everybody to follow the same methodology and apply the same modeling language.

What has business value today might not deliver any value tomorrow. Data Vault supports multiple versions of the truth: this allows catering to changing business definitions in the consumption area. Data Vault captures the complete change of history in the sources, giving business decision-makers the right to change their minds about business concepts and definitions or change use cases.

That is why we believe that Data Vault is a stable layer, a fortress that buffers against all kinds of changes over time.

Extensibility

Easy to integrate by design

An automation solution is crucial to any modern data infrastructure. It, therefore, has to be easy to integrate with the ETL, CDC, governance, source, and target technologies your organization uses today.

The best automation solutions are designed for extensibility through native integration or REST APIs. The automation engine and corresponding template language must be robust in creating abstractions between integration logic and code. The smaller the percentage of the template code is target-specific, the better.

This design principle allows for frictionless integration and swift migrations in case of changes or disruptions in the data technology stack.

When a new technology gets introduced, it entails nothing more than describing the exceptions for the template interpreter. At the same time, exceptions are stored in metadata, specific SQL functions, for example, that deviate per database engine technology.

Integration requirements

Integrations	Requirements
Sources	Look at tools that can accommodate the way you work, extracting metadata straight from the source or from separate schemas holding specific source schema versions.
Target	Ensure that your target platform's best practices comply with the data definition language (DDL) and SQL procedures the automation tool generates. Data teams should be able to set specific options for storage, MPP (Massively Parallel Processing) processing, or query optimization.
ETL	Automated data transformation shouldn't replace your ETL tool but integrate with it.
CDC	The automation code has to knit together seamlessly with your CDC tool fields (journal date, flag, log position, etc.). It must adapt to different settings of your CDC tool (remote journalling tables, all or only changed records, reliable CDC, pre-image availability, etc.)
Other	Advanced DWA tools integrate seamlessly with best-of-breed orchestration, CI/CD, data governance, data virtualization and modeling tools.

Quality of output

Eliminate technical debt

A significant challenge in establishing a central data repository is technical debt. Technical debt can be caused by taking shortcuts or deviating from standards. Different developers write SQL code differently, which also contributes to technical debt. The constant is that it always results in costly rework.

Several DWA features combined almost eradicate technical debt:

- standardization
- extensive code generation
- no-code automation, which does away with coding errors or any inconsistencies
- Data Vault 2.0 methodology (& certification)
- automated migration scripts that solve issues with the data model that occurred in the past (wrong business key, satellite splitting, single- to multi-master conversion)

The rule of transparency

Engineers think in data, and analysts think in models. Data and models are two sides of the same reality. The best solutions unify (meta)data-driven and model-driven automation. Models representing the organization and its processes are blended with facts, aka the data from the sources.

And talking about transparency, data engineers must also involve business users in data governance processes. Metadata and lineage need to be automatically fed into best-of-breed data governance solutions using REST API endpoints to convey the technical lineage and, most importantly, the matching vital business mapping.

Operational costs

Manual process can get expensive

Any data warehouse, lakehouse or mesh requires a great deal of manual upfront design and development. This includes developing the data's integration, modeling, and quality assurance.

Hand-coded solutions and classical ETL development lose a lot of time in the design and build phases. The entire design needs to be constructed manually; source metadata is manually incorporated into the target data model based on lengthy analysis documents. The physical data model's full details need to be manually built in classical data modeling tools. Besides being slow, these approaches are also prone to errors and costly rework.

Moving the problem doesn't help

Templates are being used to speed up the build. The problem is that the focus might shift from ETL development to template development. Good templates are hard to build and even harder to test and maintain. There is always another exception to consider. So instead of having quick results, developers build, test, share, copy, and adapt templates for a great deal of time.

The best way to mitigate this problem is to check if and how many built-in templates are provided to integrate and model different data from various sources and technologies without having to write them first. Or to put it differently, you don't want your highly paid professionals to end up solving elementary data problems instead of spending their time delivering real value.

The right way, right away

Finally, the success of the solution strongly depends on the completeness of the support, and the ability to adapt code to your environment. Not complying with standards or incomplete support will result in technical debt or, even worse, having to build part of the solution manually.

Save up on runtime costs

Choosing cloud-native automation means reducing infrastructure costs. But make sure that you do not need to pay runtime costs either. The most advanced solutions don't charge for loading data, only for jobs related to building and adapting the data model.

Planning for now and three years from now

Evaluating the seven areas discussed in this guide will help you and your buying team make the choice that suits your needs. Looking for the following differentiators could be the difference between a DWA solution that creates more problems than it solves and a solution that will genuinely help you accelerate analysis and data-driven decisions.

Criteria to choose the best solution

Based on the criteria from Patrick Cuba mentioned in 'the Data Vault guru.'

#	Attribute	Score
1	Ease of use • Developer interface • No download or install required on a local machine	/10
2	Version control • Model lifecycle management • CI/CD integration	/10
3	Extensible • Plugins • APIs • Language support: JavaScript, Python, C#	/10
4	Cloud native • Autoscaling • Does the tool integrate with AWS, Azure or GCP? • Can the tool be used for multi-cloud deployment? • Can the tool be used for Hybrid cloud deployment?	/10
5	Administration • Easy to manage tasks • Schedule (or 3rd party integration with scheduler support) • Access, roles, and users are easy to set up and assign	/10
6	Operation • Tool patching and upgrades • Is there downtime?	/10
7	Documentation • Easily searchable • Online community	/10
8	Service support • Responsiveness (hour, 1-3 business days, etc.) • Paid for when subscribed, or is there an additional cost?	/10

9	Change requests • Responsiveness to change requests • Ability to extend templates	/10
10	Vendor lock in • How generic is the code? • Can the pipelines function without the tool if I stop paying for the service?	/10
11	Interoperability with other tools • What can the tool also plug into?	/10
12	Cost • Compute • Storage (in addition to customer data) • 3rd party software	/10
13	Training • Developer to use the tool • Administration	/10
14	Flexibility • Tool does not lock you in the way they model a Data Vault • Tool allows for setting any standard metadata column names • Tool supports batch and streaming	/10
15	Data vault support • Tool supports the full range of Data Vault artifacts with extensible templates	/10

Need more information?

To recap, automated data transfo does four things: it solves complexity, accelerates delivery, increases agility and reduces the chance of human error. Please reach out to our team if you'd like help figuring out how to easily pull that off in a transparent manner.

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