VAULTSPEED

The Essential Buyer's Guide to Data Vault Automation

Automate to Accelerate

Table of Contents

Objectives

This guide outlines how analytics solutions have evolved to help organizations deal with vastly increased volumes of data.

These solutions comprise an ecosystem of components, where the data warehouse is crucial and acts as a central repository for all relevant organization data.

A key requirement of the data warehouse is to facilitate reporting and general analysis. Data Vault 2.0 is emerging as the leading methodology for warehouse design.

Here we show how we combine automation with the Data Vault 2.0 methodology to expedite the warehouse design and subsequent build.

The Data Explosion

The Data Explosion

The world has seen fundamental shifts from analog to digital while tremendous advances in computing power and data storage have enabled new applications and a consequent explosion in the amount of data generated.

The market intelligence company IDC has forecast that the total amount of data will grow to 175 zettabytes in 2025, some ten times the figure for as recently as 2018.

A zettabyte (you might be wondering) is equivalent to a trillion gigabytes. It sounds a mind-boggling amount though, as the futurist and influencer Bernand Marr suggests, we consider how much we are generating in an average day.

“Every interaction with your computer or phone creates data. Every interaction on social media creates data. Every time you walk down the street with a phone in your pocket, it’s tracking your location through GPS sensors – more data. Every time you buy something with your contactless debit card? Data. Every time you read an article online? Data. Every time you stream a song, movie, or podcast? Data, data, data.”

It’s a similar tale with organizations. With systems to cater for a wide range of particular functions such as sales or marketing, the amount of data has increased dramatically.

To complicate matters further, many of these systems are not interconnected, making it difficult and time-consuming to answer questions of a cross-organization nature.

Major Data Challenges

Major Data Challenges

No Control over Data Delivery

The objective of any data warehouse is to deliver readily available quality data without having to search for it. However, central data repositories are hard to implement. They require a lot of upfront design and development. This includes developing the structure, meaning, and quality assurance of the data and integrating data from disparate sources.

Lots of manual work, rework and a lack of common standards and understanding result in unmet expectations. Data warehouse projects tend to go over time and budget, and a general lack of trust in data projects settles in. At the same time, maintenance costs are soaring.

High Impact of Change

Markets are constantly changing. Change encourages organizations to respond faster and wiser. And they, in turn, are placing increasing demands on business intelligence, analytics, and data warehousing solutions.

What delivers value today won't necessarily have value tomorrow. Vendors' schedules have gone up from two or three releases per year to fortnightly release cycles. New types of data sources, such as event tracking – website and mobile applications tracking and recording user actions – pop up regularly. Data warehouse development can never be finished because new sources need to be added by the time deployment has been done.

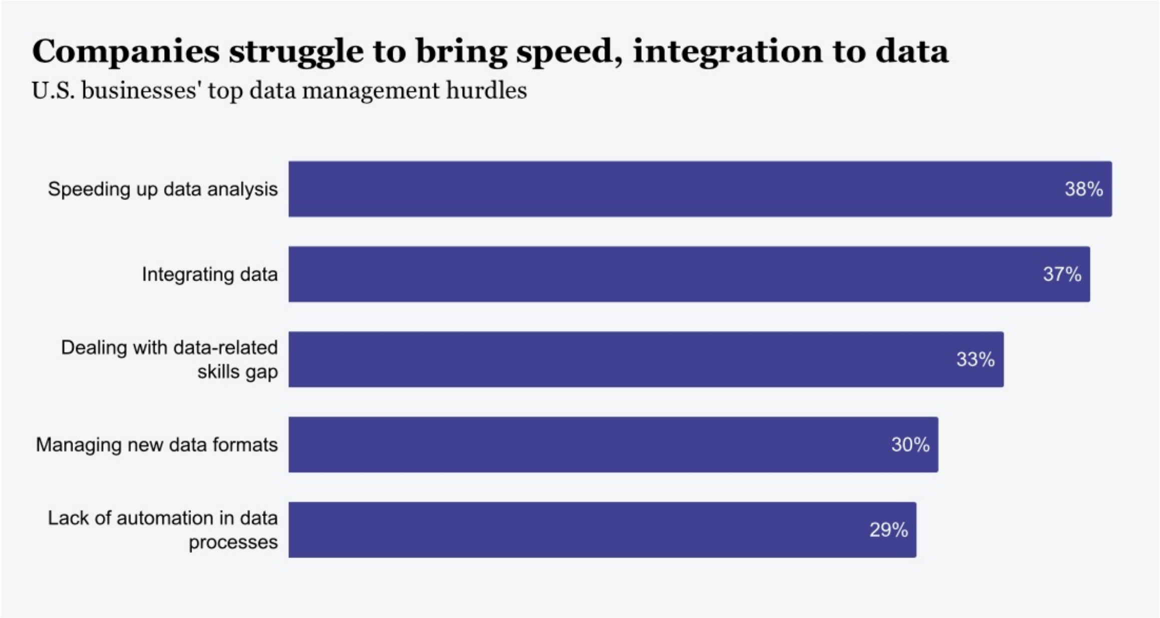
Regulations change too. Market and government regulators push organizations to add and alter reporting levels to other actors with varying data definitions in demand for more transparency.

Agile has to become the default way of working. Continuous change has a high impact on data warehouse delivery, adding additional complexity, delays, and costs to an already precarious process.

And at the same time, the job market is becoming increasingly flexible, forcing organizations to standardize to safeguard business continuity.

Lack of Speed

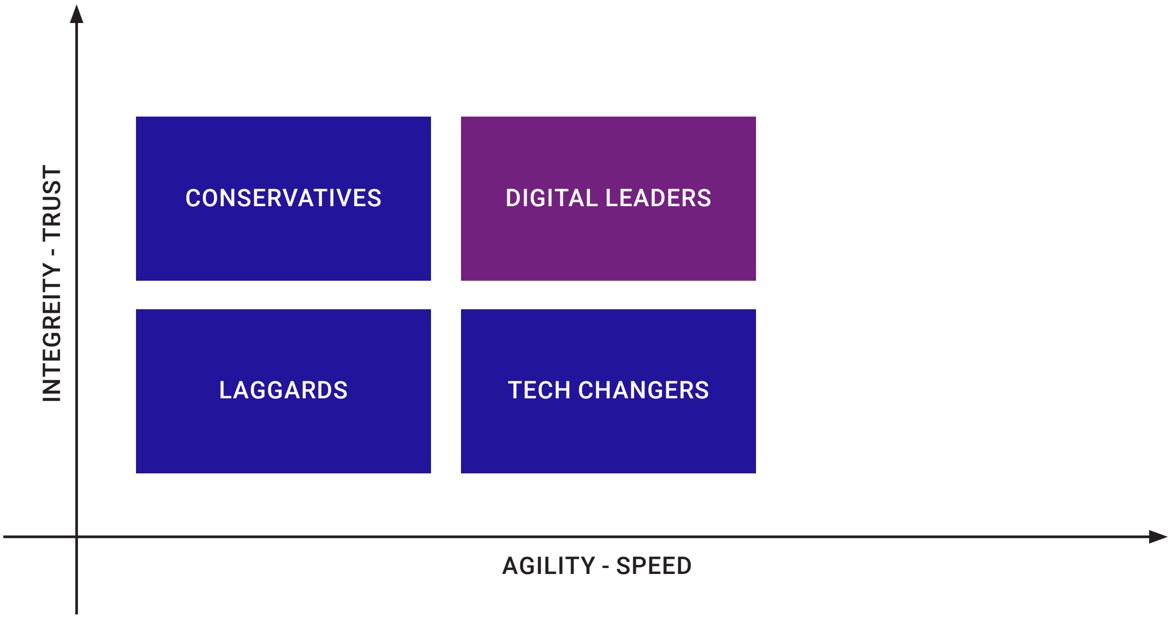
Companies find it hard to bring speed to their data integration efforts. Data integration is perceived as a slow and costly process. End users complain about a lack of efficiency and effectiveness, and data engineers, data scientists, and analysts become overloaded with work. Data engineers struggle to keep pace due to changes, while highly paid professionals end up solving elementary data problems instead of spending their time delivering real value.



Roberto Torres / CIO Dive, with data from CompTIA.

No Data Integrity

Data Integrity versus Agility



The data warehouse has been around for a long time, and it’s here to stay. Over the years, we have witnessed data architectures evolve from data warehouse to data lake to data lakehouse. And it doesn’t stop there. New architecture solutions appear almost every day, like the data mesh, data hub, and data fabric, all proposed as the silver bullet for solving data integration challenges. In general, data stakeholders have two basic needs. They want data integrity, and they want speedy delivery.

The traditional data warehouse scored high on data integrity but lacked speed. The data lake proved to be a fast way of getting data in, but companies paid the price of lacking data integrity.

	Integrity	Speed
Data warehouse	√	
Data lake		√

Technical Debt

Changes in the number and complexity of data sources affect data integration and significantly impact the work required to create and maintain the data model.

> **Gartner**
> The cost of poor-quality software in the U.S. in 2018 was approximately $2.84 trillion. Approximately $517 billion of that is the cost of technical debt (18.22%).

The data model is more than just a simple superset of data source models, so designing and maintaining the data model is a complex and time-consuming task.

New data sources need to be reflected not just in changes to data integration but also in changes to the design of the data model.

In this regard, taking shortcuts can be costly in the long run as they may make models harder to maintain and decrease quality. The cost of poor-quality software has been estimated to run into trillions of dollars, with technical debt estimated to contribute to almost 20% of the sum.

Technical debt can also be caused by deviating from standards. Different developers tend to write SQL code differently.

The constant is that technical debt always results in costly rework.

* Source: Robert Torres / CIO Dive with data from CompTIA.

VaultSpeed Solution

VaultSpeed Solution

VaultSpeed is the fastest automation tool designed to accelerate data warehouse deployment without incurring technical debt.

VaultSpeed is built on a decade of hands-on experience in data integration projects and combines three strategic technology trends to guarantee the best possible, future-proof solution: automation, Data Vault 2.0 modeling, and a cloud native architecture.

Automation

Data sources generally comprise actual data (customer details, product details, orders, etc.) and metadata.

Metadata is, in essence, data about data. It comprises ways of describing the actual data and how it is stored. With basic sources such as spreadsheets, this would mean details such as column headings. In contrast, sources held in databases contain a significant number of details, like data type (e.g., is it text, numeric, or data), and whether any particular significance is attached to any of the data fields (e.g., which fields uniquely identify a record).

In addition, database metadata typically includes details of the relationship between different data entities.

Metadata can be used not only to assist with automating the design of the data model but the integration stage as well. A significant portion of the integration stage can also be automated by retaining details of the mapping from data source to data target.

The data model can also be used as the basis for automatically producing governance and lineage data.

VaultSpeed offers the most advanced automation tool on the market, making it easy to harvest metadata from every source that holds relevant data.

We're the only tool that provides built-in integration templates to handle the complex diversity of technologies and sources without coding. This does not only accelerate the setup of your data warehouse but also solves agility. The same templates accommodate new sources or technology upgrades.

VaultSpeed automation replaces repetitive, error-prone human work and makes deployment secure, agile and scalable.

Data Vault Modeling

Data Vault 2.0 is a data model design methodology that is rapidly becoming the standard for data warehouse automation.

Promulgated by Dan Linstedt, the Data Vault approach has extended the reach of earlier methodologies such as those promoted by design gurus Bill Inmon and Ralph Kimball.

Data Vault enables a designer to start with the single version of the facts – actual data held in the various data source systems and to create a Business Data Model design.

Data Vault decomposes business concepts to their core level and integrates them. Modeling of the data warehouse is automated through the repeatable Data Vault patterns of hubs, links, and satellites.

VaultSpeed is the first and only Data Vault 2.0 automation tool certified by the Data Vault Alliance. We rigorously follow the Data Vault 2.0 standards, which means that we deliver clean code that doesn’t cause technical debt.

On top of that, the VaultSpeed tool goes as far as suggesting a first Data Vault model based on the metadata it collects from each of the selected sources. The only thing the data analyst is invited to do is tweak that model to make sure it reflects the particular business and needs.

A guided setup, tested logic, and error & exception handling ensure data integrity and save data teams from months of headache, confusion, and rework.

Every change to the repository is viewable to the data team which helps to solve any issue promptly.

Data Vault is becoming the industry standard

McKinsey digital

“Data Vault 2.0 techniques, such as data-point modeling, can ensure that data models are extensible so data elements can be added or removed in the future with limited disruption.”

Gartner

“Mitigate the weakness of the structured data platforms by using Data vault modeling or techniques from it. Although this technique needs skill and discipline, it can also substantially reduce the impact of change on a structured data model. The discipline needed can be assisted through the use of DWA tools.”

Eckerson group

“Data Vault modeling techniques are going to hit a tipping point in 2020 where a plurality of projects that involve building or re-factoring the “hub” layer of a 3-tier data warehouse architecture will employ this modeling technique.”

Cloud Native Tool

VaultSpeed matches any workload with the right compute resources. The scalability and elasticity this brings make VaultSpeed the first choice to support any cloud data warehouse implementation.

Cloud native architecture not only reduces infrastructure costs but, in the VaultSpeed case, also runtime costs. We don’t charge for loading actual data, only for jobs related to building and adapting the data model.

Built-for-Purpose Solution

VaultSpeed connects with the best CDC, ETL, source, and target technologies you rely on today.

We aim to be part of a diverse ecosystem of best-of-breed solutions that seamlessly integrate.

Every solution in this system evolves rapidly, enabling the ecosystem to adhere to the agile and ever-changing market. The integration between solutions, as opposed to monolithic single-vendor solution suites, ensures the delivery of a solution at the speed required by modern enterprises.

VaultSpeed Modeling Experience

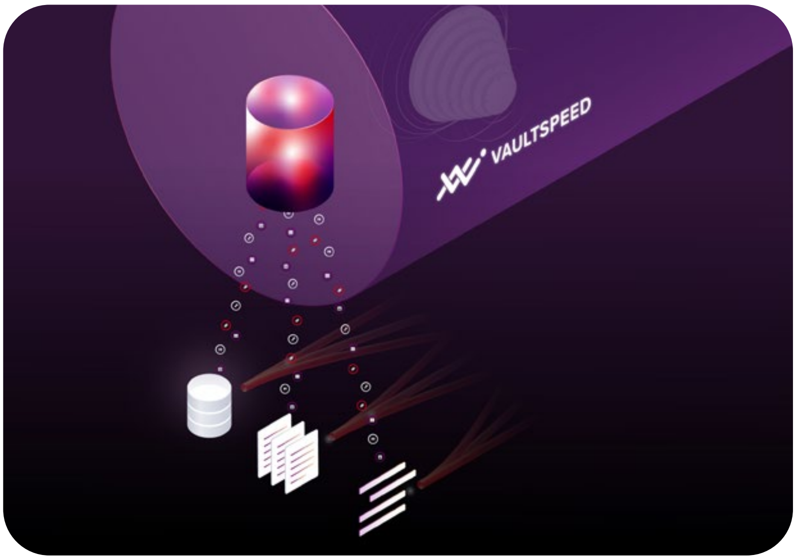
VaultSpeed Modeling Experience

VaultSpeed automates much of the work required to design your data model and build your data warehouse schema.

Not only does it automate work concerning your data model, but it can also automatically produce implementation details that can be used by your data integration tool.

It doesn’t stop there. It can also produce implementation details that can be used by your workflow orchestration and data governance tools.

Harvest the Metadata

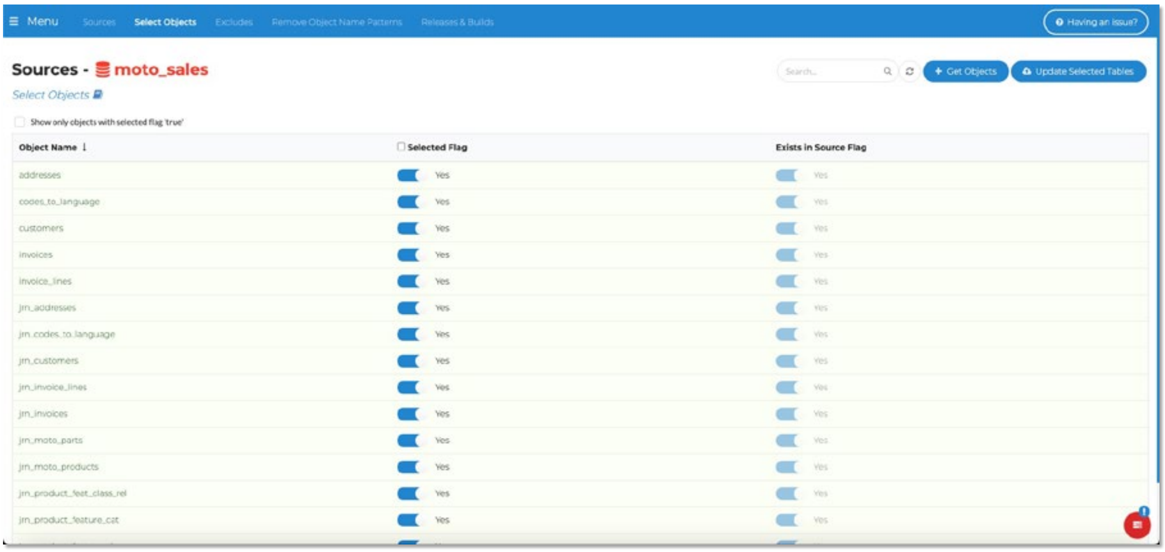


Start building your data model by harvesting metadata about each of your data sources.

Your data is safe. We provide a secure connection to your sources and will only extract metadata. VaultSpeed will never access actual data.

Metadata can be harvested from a variety of data sources, including files, many software-as-a-service solutions, and any JDBC-driver enabled database.

Metadata harvesting gives you a head start in the creation of your data model.



Sources - moto_sales

Select Objects

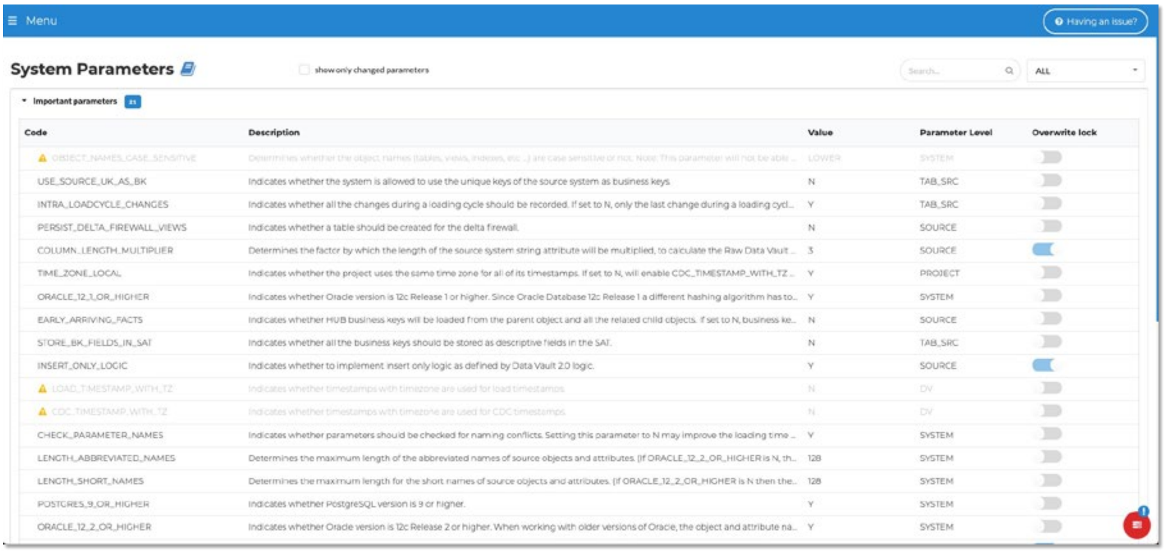
☐ Show only objects with selected flag 'true'

Object Name ↓	☐ Selected Flag	Exists in Source Flag
addresses	Yes	Yes
codes_to_language	Yes	Yes
customers	Yes	Yes
invoices	Yes	Yes
invoice_lines	Yes	Yes
jrn_addresses	Yes	Yes
jrn_codes_to_language	Yes	Yes
jrn_customers	Yes	Yes
jrn_invoice_lines	Yes	Yes
jrn_invoices	Yes	Yes
jrn_moto_parts	Yes	Yes
jrn_moto_products	Yes	Yes
jrn_product_feat_class_rel	Yes	Yes
jrn_product_feature_cat	Yes	Yes

Tech Stack Parametrization

Integrating data can be tricky. Luckily, VaultSpeed's built-in templates provide integration logic for every possible combination of source, target, CDC, and ETL technologies.

No need for coding, all our customers need to do is specify settings such as case sensitivity, loading logic, data quality or others.



System Parameters

□ show only changed parameters

Search...

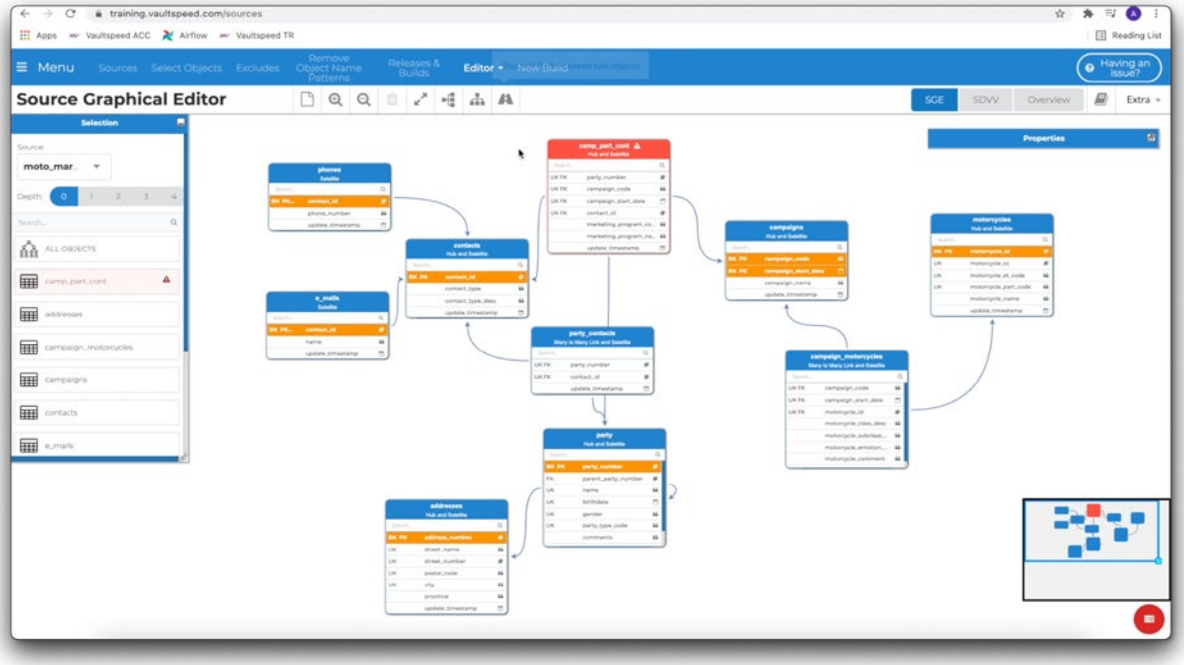
ALL

▾ Important parameters 21

Code	Description	Value	Parameter Level	Overwrite lock
⚠ OBJECT_NAMES_CASE_SENSITIVE	Determines whether the object names (tables, views, indexes, etc ...) are case sensitive or not. Note: This parameter will not be able ...	LOWER	SYSTEM	○
USE_SOURCE_UK_AS_BK	Indicates whether the system is allowed to use the unique keys of the source system as business keys.	N	TAB_SRC	○
INTRA_LOADCYCLE_CHANGES	Indicates whether all the changes during a loading cycle should be recorded. If set to N, only the last change during a loading cycl...	Y	TAB_SRC	○
PERSIST_DELTA_FIREWALL_VIEWS	Indicates whether a table should be created for the delta firewall.	N	SOURCE	○
COLUMN_LENGTH_MULTIPLIER	Determines the factor by which the length of the source system string attribute will be multiplied, to calculate the Raw Data Vault ...	3	SOURCE	●
TIME_ZONE_LOCAL	Indicates whether the project uses the same time zone for all of its timestamps. If set to N, will enable CDC_TIMESTAMP_WITH_TZ ...	Y	PROJECT	○
ORACLE_12_1_OR_HIGHER	Indicates whether Oracle version is 12c Release 1 or higher. Since Oracle Database 12c Release 1 a different hashing algorithm has to...	Y	SYSTEM	○
EARLY_ARRIVING_FACTS	Indicates whether HUB business keys will be loaded from the parent object and all the related child objects. If set to N, business ke...	N	SOURCE	○
STORE_BK_FIELDS_IN_SAT	Indicates whether all the business keys should be stored as descriptive fields in the SAT.	N	TAB_SRC	○
INSERT_ONLY_LOGIC	Indicates whether to implement insert only logic as defined by Data Vault 2.0 logic.	Y	SOURCE	●
⚠ LOAD_TIMESTAMP_WITH_TZ	Indicates whether timestamps with timezone are used for load timestamps	N	DV	○
⚠ CDC_TIMESTAMP_WITH_TZ	Indicates whether timestamps with timezone are used for CDC timestamps.	N	DV	○
CHECK_PARAMETER_NAMES	Indicates whether parameters should be checked for naming conflicts. Setting this parameter to N may improve the loading time ...	Y	SYSTEM	○
LENGTH_ABBREVIATED_NAMES	Determines the maximum length of the abbreviated names of source objects and attributes. (If ORACLE_12_2_OR_HIGHER is N, th...	128	SYSTEM	○
LENGTH_SHORT_NAMES	Determines the maximum length for the short names of source objects and attributes. (If ORACLE_12_2_OR_HIGHER is N then the...	128	SYSTEM	○
POSTGRES_9_OR_HIGHER	Indicates whether PostgreSQL version is 9 or higher.	Y	SYSTEM	○
ORACLE_12_2_OR_HIGHER	Indicates whether Oracle version is 12c Release 2 or higher. When working with older versions of Oracle, the object and attribute na...	Y	SYSTEM	○

Business Model Mapping

No need to build your data model from scratch. VaultSpeed creates a unified Data Vault model for you, based on the collected metadata. The VaultSpeed intuitive interface lets you adjust it to better correspond to your business needs by changing or adding relationships, confirming business keys, grouping hubs, or splitting satellites.

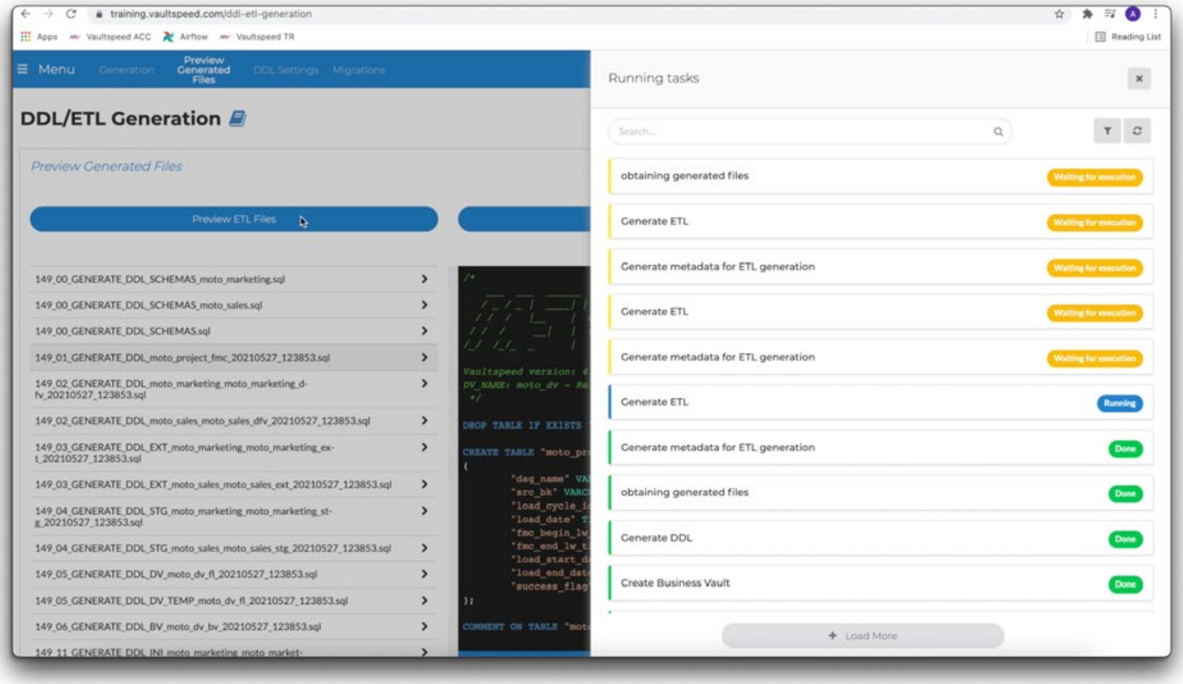


ETL & DDL Code Generation

Generate the code needed to build the data warehouse schema.

Once the data model is built, VaultSpeed can be used to generate relevant code and related detail.

- Schema code
- Data Integration detail



DDL/ETL Generation

Preview Generated Files

- 149_00_GENERATE_DDL_SCHEMAS_moto_marketing.sql
- 149_00_GENERATE_DDL_SCHEMAS_moto_sales.sql
- 149_00_GENERATE_DDL_SCHEMAS.sql
- 149_01_GENERATE_DDL_moto_project_fmc_20210527_123853.sql
- 149_02_GENERATE_DDL_moto_marketing_moto_marketing_dfv_20210527_123853.sql
- 149_02_GENERATE_DDL_moto_sales_moto_sales_dfv_20210527_123853.sql
- 149_03_GENERATE_DDL_EXT_moto_marketing_moto_marketing_ext_20210527_123853.sql
- 149_03_GENERATE_DDL_EXT_moto_sales_moto_sales_ext_20210527_123853.sql
- 149_04_GENERATE_DDL_STG_moto_marketing_moto_marketing_stg_20210527_123853.sql
- 149_04_GENERATE_DDL_STG_moto_sales_moto_sales_stg_20210527_123853.sql
- 149_05_GENERATE_DDL_DV_moto_dv_fl_20210527_123853.sql
- 149_05_GENERATE_DDL_DV_TEMP_moto_dv_fl_20210527_123853.sql
- 149_06_GENERATE_DDL_BV_moto_dv_bv_20210527_123853.sql
- 149_11_GENERATE_DDL_INI_moto_marketing_moto_market-

```
/*
 Vaultspeed version: 4
 DV_NAME: moto_dv - Re
 */

DROP TABLE IF EXISTS

CREATE TABLE "moto_pr
(
        "dag_name" VA
        "src_bk" VARC
        "load_cycle_i
        "load_date" T
        "fmc_begin_lw
        "fmc_end_lw_t
        "load_start_d
        "load_end_dat
        "success_flag
);

COMMENT ON TABLE "mot
```

Running tasks

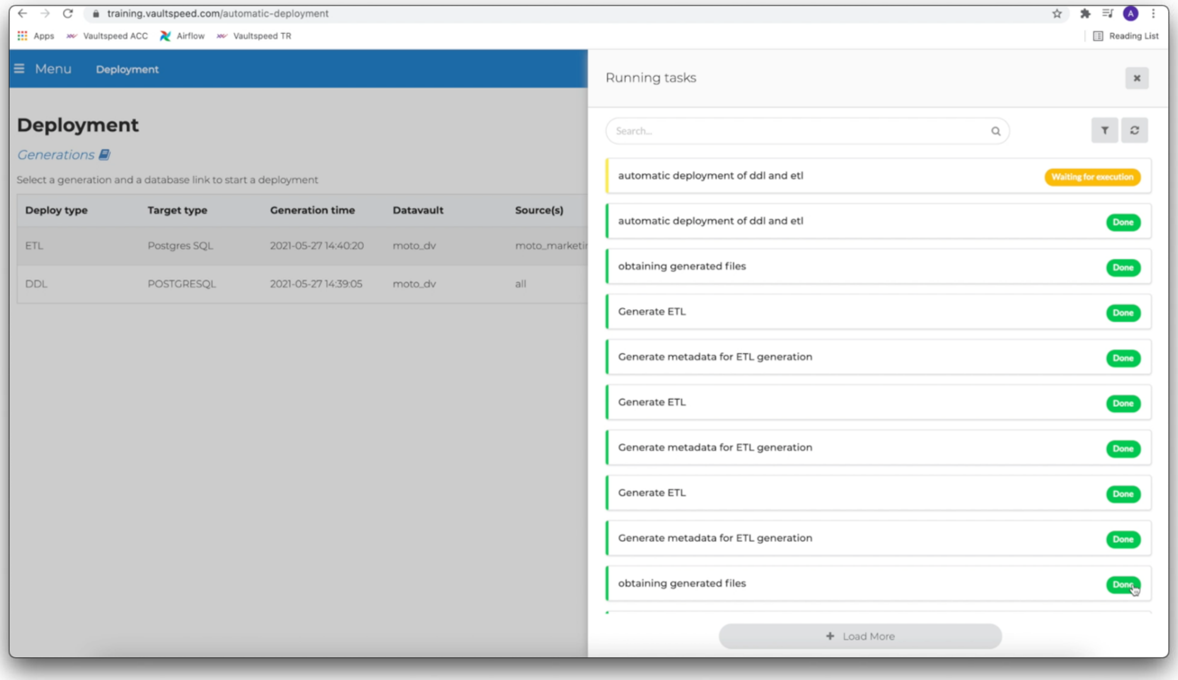
obtaining generated files	Waiting for execution
Generate ETL	Waiting for execution
Generate metadata for ETL generation	Waiting for execution
Generate ETL	Waiting for execution
Generate metadata for ETL generation	Waiting for execution
Generate ETL	Running
Generate metadata for ETL generation	Done
obtaining generated files	Done
Generate DDL	Done
Create Business Vault	Done

Full or Delta?

VaultSpeed supports code generation for a full version of your data model or a delta version to migrate from one version to the next. The delta version includes data migration scripts

Pipeline Deployment

Finally, deploy the schema code.



Deployment

Generations

Select a generation and a database link to start a deployment

Deploy type	Target type	Generation time	Datavault	Source(s)
ETL	Postgres SQL	2021-05-27 14:40:20	moto_dv	moto_marketi
DDL	POSTGRESQL	2021-05-27 14:39:05	moto_dv	all

Running tasks

automatic deployment of ddl and etl	Waiting for execution
automatic deployment of ddl and etl	Done
obtaining generated files	Done
Generate ETL	Done
Generate metadata for ETL generation	Done
Generate ETL	Done
Generate metadata for ETL generation	Done
Generate ETL	Done
Generate metadata for ETL generation	Done
obtaining generated files	Done

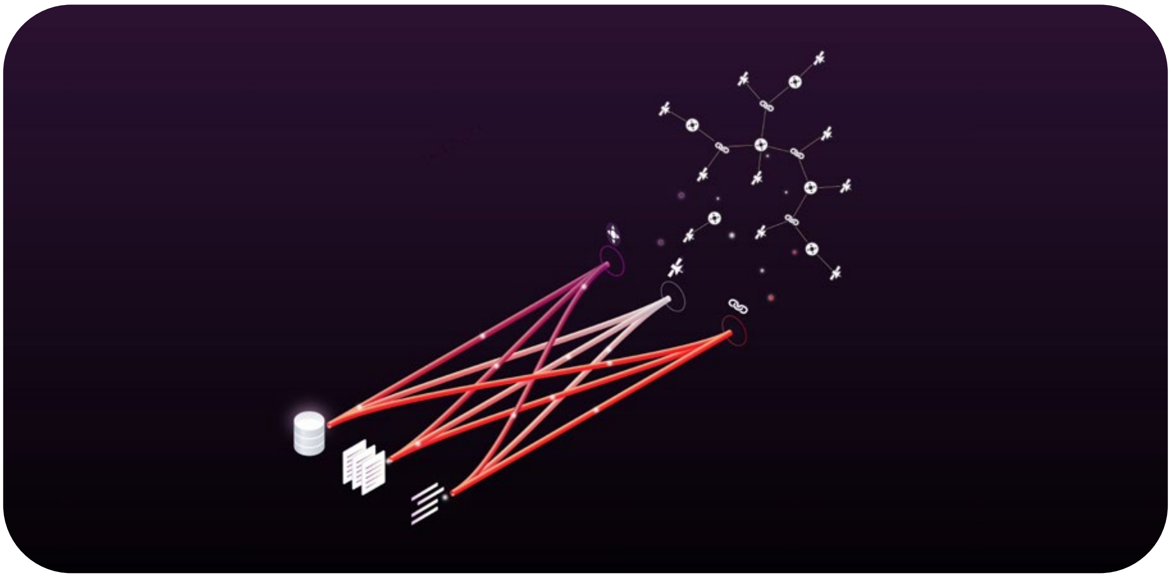
Any additional detail produced in the build stage can be shared with the appropriate parties. This includes, potentially:

- Data Integration detail
- Workflow orchestration detail
- Data Governance detail

Your data warehouse is ready to go!

Orchestrated Data Loading

VaultSpeed connects with your Airflow, Matillion, or ADF scheduler to start loading source data in the data warehouse, transforming them into analysis-ready data on the spot. Teams can instantly dive into business intelligence.



Continuous Automation

When sources are added, or technologies get upgraded, the only thing customers need to do is to adapt parameters and incorporate the changes in your existing model.

Conclusion

The combination of automation and Data Vault 2.0 enables VaultSpeed to significantly accelerate creation and maintenance of the data warehouse component of your analytics solution and produce the data warehouse component implementation detail for data integration.

Further implementation detail related to workflow orchestration and data governance can be produced where required.

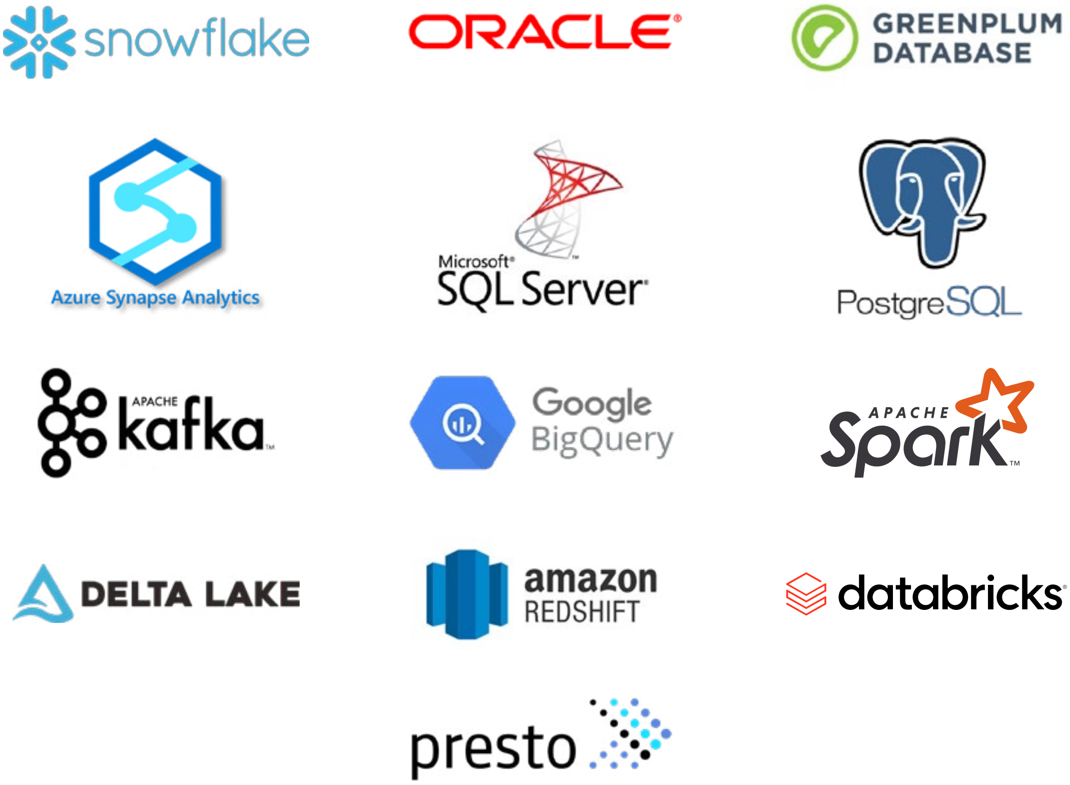
New data sources can be fast-tracked. The data model can be updated quickly and the relevant implementation detail for the integration component can be produced.

VaultSpeed helps to enforce a degree of standardization of approach regarding data model design and naming conventions.

Supported Technologies

Data Warehouse Solutions

Creation of schema for the following data warehouse is supported



Data Integration Solutions

Creation of implementation detail for the following data integration solutions is supported.

MATILLION

ORACLE FUSION MIDDLEWARE 12c

talend

HVR

ORACLE GOLDEN GATE

debezium

ATTUNITY

Stitch
A Talend Company

Data Workflow Solutions

Creation of the implementation detail for the following data orchestration solutions is supported.

MATILLION

Apache Airflow

Azure Data Factory

ASTRONOMER

Google Cloud Composer

Data Governance Solutions

Creation of the implementation detail for the following data governance is supported.

Collibra

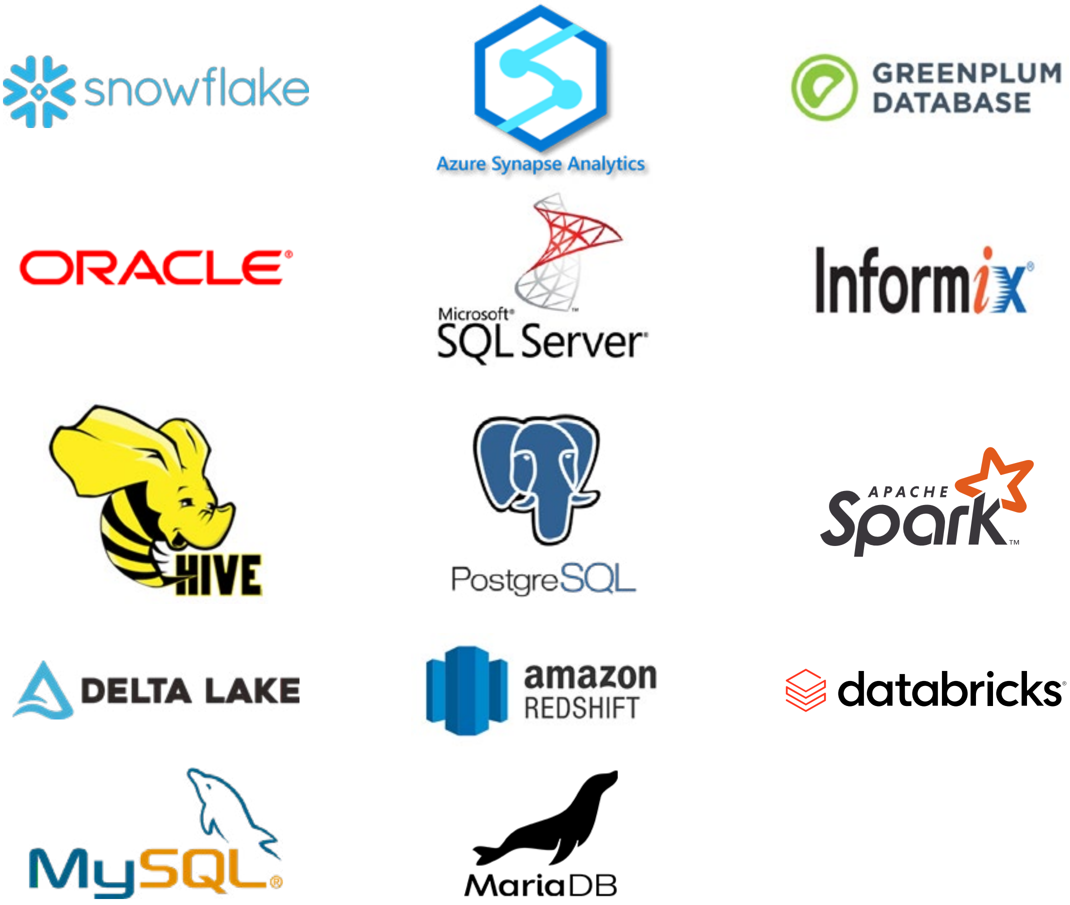
ellie

Azure Purview

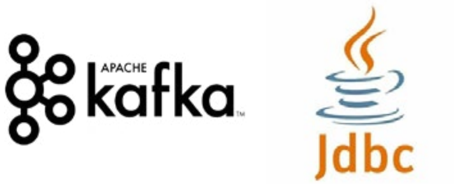
ADEPT by
eoncollective

Data Sources Metadata Harvesting

Metadata harvesting for the following data sources is supported.



With our Extended agent you get some additional sources:



* With the generic JDBC source type, you can connect any source trough jdbc. This means that, as long as you have a jdbc driver for it, you can read metadata from that source. A vendor that specialized in jdbc drivers is https://www.cdata.com/drivers/

Solution checklist

Functional considerations

Controlled delivery

Evaluation criteria:

Does the solution provide automated ETL & DDL code generation & deployment? □

Is it compliant with standards & guidelines? □

Does the solution provide tested logic? □

Does the solution offer error & exception handling? □

Is full integration with scheduling/CDC/governance tools provided? □

Does the solution provide a controlled implementation of a physical data model? □

Low technical debt

Evaluation criteria:

Does the solution eliminate error-prone manual work? □

Is the tool Data Vault 2.0 Certified? □

Does the solution provide automated code for ETL & DDL build & adaptation? □

Is the solution resilient to changes in sources, tech, staff? □

Always in sync

Evaluation criteria:

Does the solution provide code generation to handle model changes? □

Does the tool generate data migration logic? □

Does the solution offer smart version management? □

Are automatic software updates provided? □

Speed

Evaluation criteria:

Does the solution have an intuitive graphical interface? □

Does the solution provide ready-to-use templates? □

Does the solution provide no-code development? □

Are open standards used to accelerate knowledge transfer? □

Does the solution support data-driven automation? □

Does the solution provide a controlled implementation of a physical data model? □

Data Integrity

Evaluation criteria:

Does the solution avoid gaps in the data and provide historization & integration? □

Are business taxonomy & model mapping provided? □

Does the solution offer an integrated business model? □

Does the solution support the logical data warehouse? □

Does the solution support model-driven automation? □

Is it Data Vault 2.0 Certified? □

Security

Evaluation criteria:

Does the solution use a secure local agent? □

Does the solution offer private network connectivity? □

Does the solution leverage with Single Sign-On capabilities? □

Does the solution only process metadata from source systems? □

Does the solution provide dedicated infrastructure on demand? □

Is the tool secure by design? □

Is Data Security Perimeter sustained? □

No lock-in

Evaluation criteria:

- Does the solution offer yearly subscription? □
- Does the solution DDL, ETL, workflow code ownership transfer? □
- Does the solution avoid charges for running code? □
- Does the tool provide full metadata export? □

The VaultSpeed Advantage

Who we are

VaultSpeed has a long history in data warehousing consultancy. Our team consists of industry veterans who have been active in data integration and data warehouse automation for over a decade.
We are headquartered in Leuven, Belgium, with offices in Lithuania and in the US.

Partners

VaultSpeed has established a global partner network providing access to local knowledge and expertise in markets in the US, India, South Africa, UK, Japan, Germany, and Benelux.

Mission

We aim to achieve world-class business & decision intelligence, governance, and competitiveness by delivering no-code data integration platforms that provide speed and resilience to change.

Investors

VaultSpeed is backed by established investors, including Fortino Capital and Co- Foundry (Cronos).

Clients

VaultSpeed has acquired knowledge and expertise globally in the leading industry verti¬cals, including financial services, manufacturing, government, and utilities.

vaultspeed.com

Visit our site
vaultspeed.com

Contact sales
sales@vaultspeed.com

Book a demo
vaultspeed.com/book-a-demo

Join our community
community.vaultspeed.com

Sluisstraat 79 03-01
3000 Leuven
Belgium

VAULTSPEED