Data Architecture as a Service

Empowering the Business to Build Compliant Pipelines

WAYNE W. ECKERSON

SEPTEMBER 2022

About the Author

Wayne W. Eckerson is an international thought leader in the data and analytics field since the early 1990s. He is a noted speaker, sought-after consultant, and widely read author. Eckerson has conducted groundbreaking research studies, chaired numerous conferences, and consulted with organizations around the world during his 25 years in the field. Eckerson has also written two books: *The Secrets of Analytical Leaders: Insights from Information Insiders* (2012) and *Performance Dashboards: Measuring, Monitoring, and Managing Your Business* (2005/2010) He has degrees from Williams College and Wesleyan University.

About Eckerson Group

Eckerson Group is a global research and consulting firm that helps organizations get more value from data. Our experts think critically, write clearly, and present persuasively about data analytics. They specialize in data strategy, data architecture, self-service analytics, master data management, data governance, and data science. Organizations rely on us to demystify data and analytics and develop business-driven strategies that harness the power of data. **Learn what Eckerson Group can do for you!**

About This Report

Research for this report comes primarily from numerous briefings with software vendors. This report is sponsored by Coalesce, who has exclusive permission to syndicate its content.

Table of Contents

Executive Summary

Data architecture-as-a-service (DAaaS) is a new self-service paradigm that empowers business users to create architecturally compliant data pipelines and repositories. By injecting architectural guardrails into commercial self-service data engineering tools, DaaS solves the problem of data silos, which wreak havoc on enterprise data consistency and trustworthiness.

DAaaS is not a new concept. Architects and engineers have long created development templates to foster reuse, accelerate development, and improve accuracy and efficiency. But DAaaS commercializes this approach, baking templates and automation into GUI-based development environments geared to business users.

DAaaS is a metadata-driven approach that auto-generates code and documentation. It also abstracts the underlying platform so data architects and engineers can change or update the platform without impacting business users. On the whole, DAaaS promises to govern self-service, foster reuse, reduce errors, and speed development.

Key Takeaways:

DAaaS enables:

> Compliance with architectural standards, minimizing data silos.

> Governed, self-service data engineering.

> The insertion of architectural guardrails into no-code development environments.

> Auto-generation of code, accelerating development and improving accuracy.

> Abstraction of underlying data platform and infrastructure.

Recommendations:

> Expect DAaaS-compliant products to emerge in the next several years.

> Prime candidates are commercial data integration, transformation, and ingestion products as well as dataOps and data warehouse automation tools.

> Ask your data vendors to support DAaaS concepts and principles.

> Experiment with adding reusable templates to existing data engineering tools to foster reuse and auto-generate code.

Understanding Data Architecture as a Service

The Answer to Data Silos

Ever since the dawn of data, there is one thing that has always plagued data managers: data silos. It seems that no matter what data managers do, data silos keep popping up everywhere, like a relentless game of whack-a-mole.

On the bright side, data silos reflect a voracious appetite for data among business users. But when the enterprise data team can't meet their needs fast enough, they take matters into their own hands. They hire data analysts, purchase self-service tools, and build renegade data repositories and IT shadow systems.

Data mesh. Today, we have an entire architecture—the data mesh—that formalizes data silos. Data mesh advocates say that "if you can't beat them, join them." But data mesh skeptics, like myself, argue that the same dynamics that spawn data silos will also make it challenging for a data mesh to, well, mesh. There must be governance and standards for technology, data, and process to make a data mesh architecture work.

> *The same dynamics that spawn data silos will make it challenging for a data mesh to, well, mesh.*

Unfortunately, governance is a dirty word for business leaders. They think governance processes and standards will slow them down. They strive to avoid data governance meetings at all costs. Nonetheless, it's critical that organizations establish governance practices with code reviews to ensure alignment with enterprise standards. But this takes time to set up and requires buy-in from all stakeholders, which is not guaranteed.

Self-Service Data Engineering

A different and complementary approach is data architecture as a service (DAaaS). Rather than make governance external to development processes, DAaaS embeds standards into development tools that data analysts and data scientists use to create data pipelines. Data architects configure the tools with templates and other "guardrails" so business users can't violate architectural standards and conventions that lead to the creation of data silos.

DAaaS embeds standards into development tools that data analysts and data scientists use to create data pipelines.

BI guardrails. Business intelligence (BI) tools have long offered DAaaS-like capabilities. BI architects build a semantic layer into the tool that predefines joins, defines metrics, and adheres to standard naming conventions. The model enables business users to query, pivot, and visualize data without knowing the intricacies of back-end databases and models. BI tools not only simplify ad hoc queries and analysis (i.e., self-service), but also help prevent business users from creating non-standard reports and dashboards.

Unfortunately, there have been no equivalent products for building data pipelines and repositories. For decades, ambitious data analysts, stymied by IT and data bottlenecks, have used Excel and Microsoft Access to mash together data sets. More recently, they've used data preparation tools from Alteryx, Tableau, and others. Unfortunately, these desktop tools are almost impossible to govern from an enterprise perspective.

Challenge. We can't expect data analysts to do the work of experienced data architects or data engineers. They don't necessarily know how to design, model, and implement robust, scalable data environments or build data pipelines that reuse standard data flows and naming conventions.

DAaaS enables data analysts to do their work effectively and at scale without involving data engineers for what are often repetitive time-consuming tasks. Data analysts don't need to know the architectural principles and standards required to generate conforming data sets. They don't need to know the difference between inner and outer join or write SQL code to handle slowly changing dimensions. DAaaS does these things for them. DAaaS offers a proverbial win-win: business users are empowered to build their own data domains while IT ensures the integrity of the enterprise data architecture.

DAaaS offers a proverbial win-win: business users are empowered to build their own data domains while IT ensures the integrity of the enterprise data architecture.

Benefits

Abstraction. Data architecture-as-a-service is a verbal twist on cloud processing environments, such as software-as-a-service or platform-as-a-service. This moniker conveys that it's possible to abstract architecture and build it into easy-to-use, customer-facing tools. When we abstract data architecture, we solve the most enduring data pain point: the proliferation of data silos that wreak havoc on data consistency and trustworthiness.

DAaaS abstracts the underlying platform so data architects and engineers can change or update the platform without impacting business users. Data pipelines continue to work, even if administrators migrate from Azure Synapse to Google BigQuery or upgrade to a major new version of either. And business users don't need to rewrite code or learn new features or functions. Most may not even be aware that the underlying infrastructure changed.

Self-service. DAaaS is the culmination of self-service, where business units liberate themselves almost entirely from enterprise IT. With DAaaS, a data architect bakes architectural requirements into self-service data engineering tools so data analysts can create their own repositories without undermining data consistency and trustworthiness. If done right, DAaaS reduces data bottlenecks, eases the burden on enterprise data teams, and empowers data domains to service their own data needs.

Reuse. From a developer perspective, DAaaS fosters reuse, accelerates development, and eliminates costly errors that arise when inexperienced developers try to create complex SQL code. It's a metadata-driven approach that uses configurable templates within commercial data pipeline products that automatically generate the SQL code. DAaaS enables data analysts, data scientists, and new data developers to leverage the work of experienced architects and engineers without having to learn the intricacies of data pipeline design and development.

The Market for DAaaS Products

Custom Building Blocks

In our consulting practice, we've seen enterprise data architects create data "building blocks" that departmental analysts use to create extensions to an enterprise data warehouse. The blocks contain governance guardrails that enable analysts to create their own data marts without deep knowledge of SQL, data structures, query logic, or schemas. However, it's a heavy lift for most enterprise data teams to create a self-service data infrastructure given competing demands for their time.

Commercial Products

Today, software vendors are starting to recognize the opportunity to offer DAaaS-enabled data pipeline development tools. These products come in a variety of shapes and forms. Most DAaaS products are cloud-based tools that support the development of data pipelines, including data integration, data transformation, DataOps, and data warehouse automation (DWA) tools.

Coalesce. One example is Coalesce, which sponsored this report. Launched in January 2022, Coalesce is a data transformation vendor that offers a SaaS alternative to dbt, a popular, open source data transformation toolkit. Founded by ex-WhereScape employees, Coalesce offers both GUI- and code-based development environments, a column-aware architecture that supports full data lineage, and built-in automation functions. Germane to this report, the product allows data administrators to insert architectural guardrails into its GUI-based pipeline development environment.

We expect many other vendors to introduce DAaaS-based products in the next two or three years. Every vendor wants to help businesses accelerate, if not automate, the development of data pipelines. DAaaS promises to remove the last remaining bottleneck between business users and data.

DAaaS Pipeline Development

A DAaaS product provides a GUI-based development environment for business users and a coding environment for developers. Business users point and click to create compliant data pipelines, while developers can view the SQL that is generated and alter it if needed. Architects or administrators configure data processing "nodes" that business users stitch together to create a visual workflow that defines the data pipeline.

In the example below, a business user has dragged source tables into a development environment and created a customer staging table and dimension table using a drop-down menu. The business user selects from a dozen or so node types which were pre-configured by an architect or administrator.

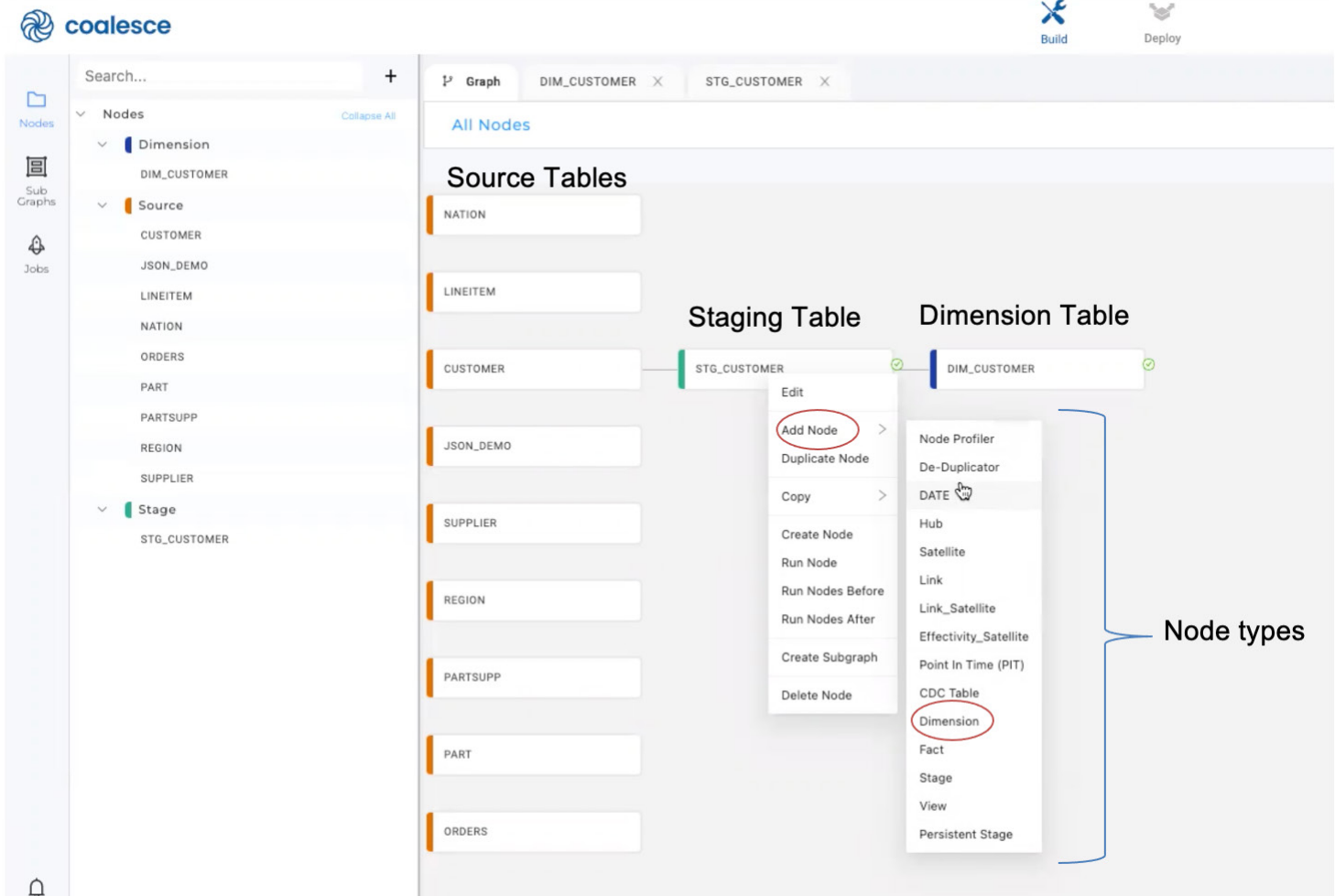**Figure 1. Selecting Nodes from a Drop-Down List**

Figure 1 courtesy of Coalesce

To define the customer dimension, the business user opens the node and views the available fields and sample data, if desired. The system automatically adds fields to the node, including a surrogate customer key and time-based fields to support future updates and changes to the field. The right-side panel presents the user with various options for configuring the dimension. For instance, the user can choose to create the dimension as a table or view depending on their business requirements. They can also define a join key and select columns to track changes to historical values (i.e., slowly changing dimensions). (See figure 2.)
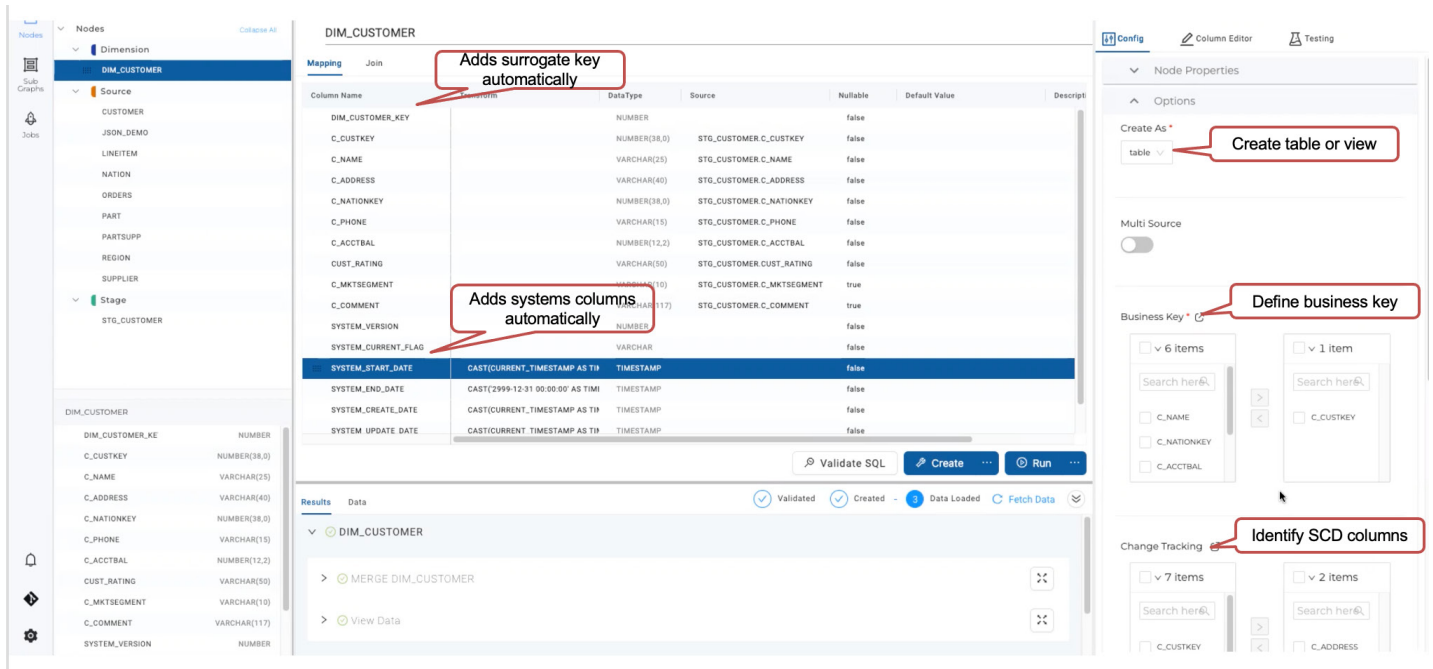
Figure 2. Configuring Options

Figure 2 courtesy of Coalesce

The right-side panel is defined by the architect in advance. They define in code all the options that get displayed in the panel. Once users select the options, the tool automatically generates the code as well as documentation about what was built. By abstracting database concepts, business users don't need to write or manipulate code or understand database processing at a deep level. In the example above, the tool saved business users from having to write 236 lines of SQL code. (See figure 3.)

Figure 3. Auto-Generated Code


```sql
⊘ MERGE DIM_CUSTOMER

number of rows inserted: 150000

1   MERGE INTO "CLS_PROD_DB"."DEMO"."DIM_CUSTOMER" "TGT" USING (
2     /* New Rows That Don't Exist */
3     SELECT
4       "STG_CUSTOMER"."C_CUSTKEY" AS "C_CUSTKEY",
5       "STG_CUSTOMER"."C_NAME" AS "C_NAME",
6       "STG_CUSTOMER"."C_ADDRESS" AS "C_ADDRESS",
7       "STG_CUSTOMER"."C_NATIONKEY" AS "C_NATIONKEY",
8       "STG_CUSTOMER"."C_PHONE" AS "C_PHONE",
9       "STG_CUSTOMER"."C_ACCTBAL" AS "C_ACCTBAL",
10      "STG_CUSTOMER"."CUST_RATING" AS "CUST_RATING",
11      "STG_CUSTOMER"."C_MKTSEGMENT" AS "C_MKTSEGMENT",
12      "STG_CUSTOMER"."C_COMMENT" AS "C_COMMENT",
13      1 AS "SYSTEM_VERSION",
14      'Y' AS "SYSTEM_CURRENT_FLAG",
15      CAST(CURRENT_TIMESTAMP AS TIMESTAMP) AS "SYSTEM_START_DATE",
16      CAST('2999-12-31 00:00:00' AS TIMESTAMP) AS "SYSTEM_END_DATE",
17      CAST(CURRENT_TIMESTAMP AS TIMESTAMP) AS "SYSTEM_CREATE_DATE",
18      CAST(CURRENT_TIMESTAMP AS TIMESTAMP) AS "SYSTEM_UPDATE_DATE",
19      'INSERT_INITAL_VERSION_ROWS' AS "DML_OPERATION"
20    FROM
21      "CLS_PROD_DB"."DEMO"."STG_CUSTOMER" "STG_CUSTOMER"
22      LEFT JOIN "CLS_PROD_DB"."DEMO"."DIM_CUSTOMER" "DIM" ON "STG_CUSTOMER"."C_CUSTKEY" = "DIM"."C_CUSTKEY"
23    WHERE
24      "DIM"."C_CUSTKEY" IS NULL
25    UNION ALL
26      /* New Row Needing To Be Inserted Due To Type-2 Column Changes */
27    SELECT
28      "STG_CUSTOMER"."C_CUSTKEY" AS "C_CUSTKEY",
29      "STG_CUSTOMER"."C_NAME" AS "C_NAME",
30      "STG_CUSTOMER"."C_ADDRESS" AS "C_ADDRESS",
31      "STG_CUSTOMER"."C_NATIONKEY" AS "C_NATIONKEY",
32      "STG_CUSTOMER"."C_PHONE" AS "C_PHONE",
33      "STG_CUSTOMER"."C_ACCTBAL" AS "C_ACCTBAL",
34      "STG_CUSTOMER"."CUST_RATING" AS "CUST_RATING",
35      "STG_CUSTOMER"."C_MKTSEGMENT" AS "C_MKTSEGMENT",
36      "STG_CUSTOMER"."C_COMMENT" AS "C_COMMENT",
37      "DIM"."SYSTEM_VERSION" + 1 AS "SYSTEM_VERSION",
38      'Y' AS "SYSTEM_CURRENT_FLAG",
39      CAST(CURRENT_TIMESTAMP AS TIMESTAMP) AS "SYSTEM_START_DATE",
40      CAST('2999-12-31 00:00:00' AS TIMESTAMP) AS "SYSTEM_END_DATE",
```

Figure 3 courtesy of Coalesce

Change once. Since DAaaS is template-based, if architects want to redesign how they manage or join tables, among other things, they update the template in one place and propagate those changes automatically to all pipelines built using that template. This accelerates design changes, abstracts development from underlying infrastructure, and helps turn the enterprise data team from a data bottleneck to a data accelerator.

Overall, DAaaS makes it possible for business users to construct complex data pipelines without having to write code or master complex database concepts, like slowly changing dimensions. This means they can build more complex data pipelines without IT involvement and minimal governance. And since DAaaS auto-generates code, it improves accuracy, fosters reuse, and auto-generates documentation—delivering more value to the business.

Evaluation Criteria

To prepare for the advent of user-friendly data development products, we've prepared a checklist of features and functions that we think every DAaaS-compliant product should support. Although every vendor will take a different approach to abstracting architecture, we think there are core principles that characterize a DAaaS product. As in any emerging discipline, this list is bound to evolve as customers implement these products in production environments.

Here are criteria for evaluating DAaaS products:

1. **Configurable.** Architects or data engineers can configure the product using templates, blocks, or other constructs to simplify the development of data pipelines.

2. **Multi-code.** The product supports no-code, low-code, and all-code development environments so that it can be used by all types of users.

3. **Metadata-driven.** The product automatically generated SQL code that developers can view and modify, if desired.

4. **Universal updates.** Developers can update a template in one place and ripple changes automatically to all solutions that leverage the code.

5. **Platform agnostic.** The product runs on multiple data platforms, adjusting SQL output as needed, and architects can migrate platforms without rewriting code.

6. **Connectors.** The tool connects to multiple source systems and loads it into multiple targets.

7. **Orchestration.** The product executes transformations and tasks according to a workflow designed within a directed acyclic graph.

8. **Data lineage.** The tool tracks lineage of data elements from source to target, both at a table level and column level.

9. **Data catalog.** Users can search and reuse existing data pipelines or data assets to speed development and promote reuse.

10. **Documentation.** The tool automatically generates documentation to speed onboarding of new developers and support auditors.

Next Generation DWA. The features above align with a prior generation of products that automated the generation of data warehousing structures. Like data warehouse automation (DWA) products, DAaaS products rely on metadata to automatically generate SQL code, simplify changes, eliminate errors, and auto-generate documentation, including lineage. The main difference is that DAaaS products are SaaS-based products that contain a no-code development environment geared to business users as well as an all-code environment geared to developers.

Conclusion: The Future is DAaaS

DAaaS is not a new concept. Architects and engineers have long created development templates to foster reuse, speed development, and improve accuracy and efficiency. But DAaaS commercializes this approach, baking templates and automation into self-service data pipeline products geared to both business and technical users.

For business users, DAaaS provides self-service data engineering tools that empower them to meet their own data needs without IT assistance while aligning with enterprise standards and guidelines. For architects, DAaaS enables them to embed standards into data pipeline tools to improve accuracy, minimize data silos, and abstract underlying infrastructure. For developers, DAaaS speeds development, fosters reuse, and auto-generates code they can review and manipulate.

There's a lot to like about DAaaS. Ask your data integration, ingestion, transformation, or DataOps vendors when they will begin supporting DAaaS principles. It's time our back-end data environments fostered the same degree of abstraction and governance as our front-end business intelligence environments.

About Eckerson Group

Wayne Eckerson, a globally-known author, speaker, and consultant, formed **Eckerson Group** to help organizations get more value from data and analytics. His goal is to provide organizations with expert guidance during every step of their data and analytics journey.

Eckerson Group helps organizations in three ways:

> **Our thought leaders** publish practical, compelling content that keeps data analytics leaders abreast of the latest trends, techniques, and tools in the field.

> **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate business requirements into compelling strategies and solutions.

> **Our advisors** provide one-on-one coaching and mentoring to data leaders and help software vendors develop go-to-market strategies.

Eckerson Group is a global research and consulting firm that focuses solely on data and analytics. Our experts specialize in data governance, self-service analytics, data architecture, data science, data management, and business intelligence.

Our clients say we are hard-working, insightful, and humble. It all stems from our love of data and our desire to help organizations turn insights into action. We are a family of continuous learners, interpreting the world of data and analytics for you.

Get more value from your data. Put an expert on your side. **Learn what Eckerson Group can do for you!**

Eckerson Group

RESEARCH • CONSULTING • ADVISORY

G E T • M O R E • V A L U E • F R O M • Y O U R • D A T A

About the Sponsor

Founded in 2020, Coalesce is the only data transformation tool built for scale. As the first platform to combine the speed of an intuitive graphical user interface (GUI), flexibility of code, and the efficiency of automation for data transformations, Coalesce customers benefit from increased data engineer productivity and insights. Based in San Francisco, Calif., Coalesce is backed by Emergence Capital, 11.2 Capital, GreatPoint Ventures, and Industry Ventures and supports customers worldwide.

Learn more at **https://coalesce.io/**.