Market Landscape Report

Modern Data Pipelines

Trends and Tools

BY KEVIN PETRIE

JULY 2023

RESEARCH SPONSORED BY

StreamSets
A SOFTWARE AG COMPANY

About the Author

Kevin Petrie is the VP of Research at Eckerson Group, where he manages the research agenda and covers topics such as data pipelines, data observability, machine learning, and cloud data platforms. For 25 years Kevin has deciphered what technology means to practitioners, as an industry analyst, writer, instructor, marketer, and services leader. Kevin launched, built, and led a profitable data services team for EMC Pivotal in the Americas and EMEA, and ran field training at the data integration software provider Attunity (now part of Qlik). A frequent public speaker and co-author of two books on data streaming, Kevin also is a data management instructor at eLearningCurve.

Kevin is the author of Eckerson Group's Decoding Data Software blog.

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

Eckerson Group provides independent and objective research on emerging technologies, techniques, and trends in the field. While we do not recommend vendors or products, we write sponsored profiles to help practitioners understand how offerings differentiate themselves.

Table of Contents

Evolution of Data Pipelines

The role of data pipelines has evolved over three decades to support increasingly complex, cloud-driven data environments. This evolution comprises three phases:

- **Phase 1: Load (1990-2010).** During this phase companies used basic ETL pipelines that loaded periodic batches of database records, perhaps hourly, daily, or weekly, into a central data warehouse for business intelligence projects such as operational reporting and dashboards.
- **Phase 2: Consolidate (2010s).** In the second phase, companies modernized their environments by migrating analytics workloads to new data warehouses in the cloud. They sought to consolidate data from a rising number of sources, including IoT sensors, log files, and SaaS applications as well as traditional databases, into a data warehouse for BI or data lake for data science.
- **Phase 3: Synchronize (2020s).** Despite efforts to consolidate, data environments grow more diverse than ever. Companies maintain some data on premises due to regulatory concerns, data gravity, and the sheer cost of moving it all. While data warehouses and lakes start to merge into lakehouses, companies often have multiple such platforms across two or even three cloud providers. Data pipelines must synchronize data across these distributed elements in real time to support BI and AI/ML projects, as well as merged workflows in which analytical outputs trigger operational action.

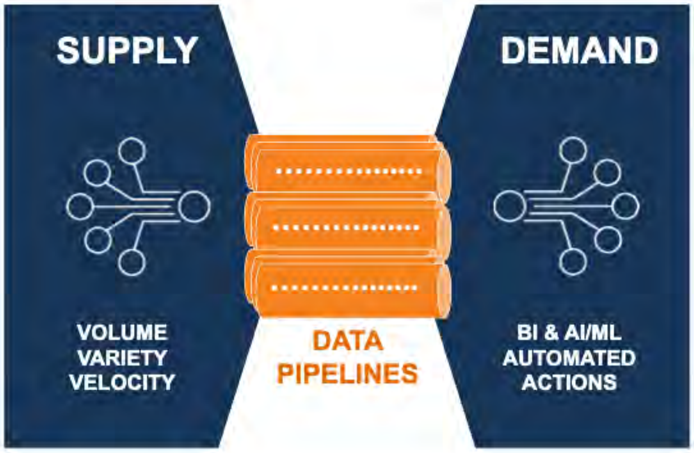
Data pipeline management has evolved over three phases: from loading (1990-2010) to consolidation (2010s) to synchronization (2020s)

Related Resources:

The Rise of Streaming ETL: Transforming Data in Flight for Real-Time Insight and Action

Market Drivers

Data Supply and Demand are Booming…



…Creating the Need for Many Data Pipelines

- Business requirements change
- Bottlenecks appear
- Complexity rises

With this evolution in mind, let's take a look at the forces pressuring modern data pipelines. On one side, the supply of data is booming. Digital transformation means that all kinds of business actions and interactions now throw off data, pushing up volume, variety, and velocity. On the other side, demand is also booming because business owners need more data from more sources, real time, to make better decisions and compete. Data pipelines must support this booming supply and demand while responding to changing business requirements, fixing bottlenecks, and reducing complexity. In this context, data engineers and analytics engineers struggle to build and manage the myriad data pipelines they need to feed modern business.

Data pipelines support booming data supply and demand while responding to changing business requirements, fixing bottlenecks, and reducing complexity

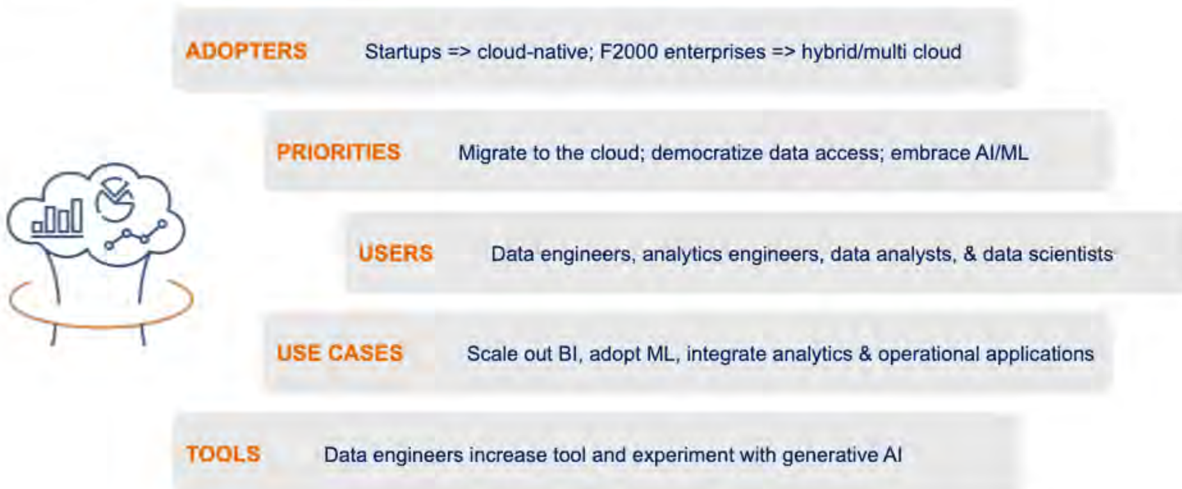
Market Trends

Five market trends shape the market for data pipeline tools: adopters, priorities, users, use cases, and tools.

> **Adopters**. Data pipeline adopters divide into two camps, each with distinct characteristics and requirements. Cloud-native companies born since 2010 embrace the latest pipeline tools with little if any need to worry about technical debt. Large and midsized enterprises, on the other hand, use pipelines to synchronize data across a mix of old and new elements in their hybrid/multi-cloud environments.

> **Priorities**. Data teams seek to keep migrating analytical and operational workloads to the cloud; democratize business access to data; and build competitive advantage with AI/ML projects.

- **Users**. Data pipeline users and managers include data engineers that ingest and transform data; analytics engineers that prepare and document data for analytics; analysts that serve themselves with basic discovery and preparation; and data scientists that define features and train AI/ML models.
- **Use cases**. The use cases for data pipelines break out into three loose categories: scaling out BI projects in terms of users, tools, datasets, etc.; testing and bringing AI/ML projects into production; and merging analytical outputs with operational applications to drive smarter action.
- **Tools**. Eckerson Group research shows that most practitioners expect to increase the number of pipeline tools they use in 2023. It also shows that more than 40% of practitioners already use ChatGPT to assist their data engineering work.



Use cases for data pipelines include scaling out BI projects; testing and bringing AI/ML project into production; and merging analytical outputs with operational applications

Definition of a Data Pipeline

A data pipeline refers to a workflow that ingests multi-structured data, schemas, and other types of metadata from sources to targets and transforms that data for analytics. Ingestion entails one or more of the following tasks:

- **Extracting or capturing** data from a source, such as one or many records from a database
- **Streaming** data messages in memory between sources and targets, for example to enable real-time transformation, delivery, and/or analytics

- **Loading** either batch data or incremental updates into a target such as a data lake
- **Appending** data to a target by adding it to existing datasets
- **Merging** data into a target by combining it with existing objects such as tables or files

Transformation, meanwhile, includes tasks such as the following. It can take place before or after the pipeline loads data to the target.

- **Filtering** data to identify and remove unneeded subsets such as columns, tables, or images, for example to protect personally identifiable information
- **Combining** multi-sourced data, for example to add columns to a table or join tables for a query
- **Formatting** data, for example by converting various tables to a single format such as Parquet
- **Structuring** data, for example by applying a schema to organize tables and columns in a database
- **Cleansing** data by removing duplicates, fixing errors, or taking other steps to improve data quality

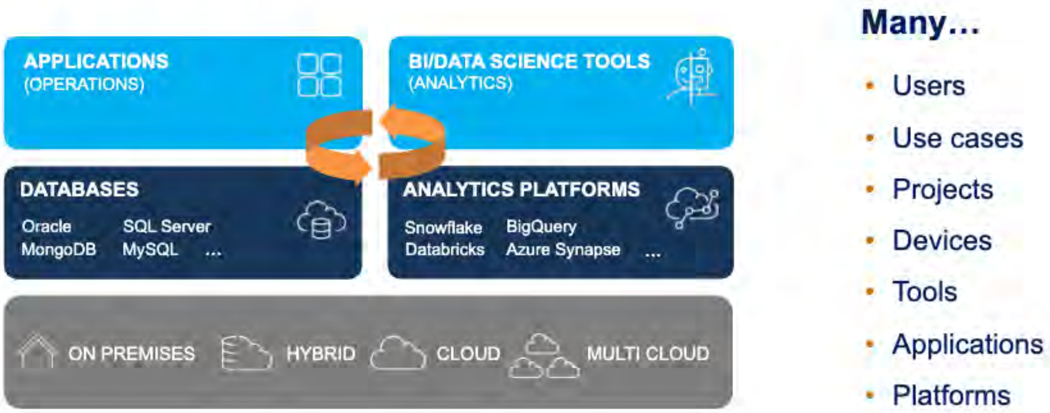
Modern pipelines span on premises, hybrid, cloud, and multi-cloud ecosystems that include various pipelines, languages, open-source projects, interfaces, tools and now AI bots, as shown in the examples in this diagram.

A data pipeline refers to a workflow that ingests multi-structured data, schemas, and other types of metadata from sources to targets and transforms that data for analytics

Related Resources:

Data Pipeline Design Patterns

Modern Data Environments



To understand the complexity of modern data environments, consider their many elements. Databases such as Oracle or MongoDB, applications such as Salesforce or Slack, analytics platforms such as Snowflake or BigQuery, and BI/data science tools such as Qlik or Tensorflow all must share data with one another. They depend on pipelines to deliver the right data to the right target in the right format—on time. Compounding the complexity, data engineers and other stakeholders must contend with multiplying users, use cases, projects, devices, tools, applications, and platforms.

Data engineers and other stakeholders must contend with multiplying users, use cases, projects, devices, tools, applications, and platforms

Market Definition

Now we define the four segments of the market for data pipeline management. They include data ingestion, data transformation, DataOps, and pipeline orchestration. A variety of commercial and open-source tools address one or more of these four market segments. Many commercial tools take a no code/low code approach by providing a graphical user interface that reduces or eliminates the need for manual scripting.

- **Data ingestion tools** help configure, manage, and monitor pipelines that ingest data using the various tasks described earlier.
- **Transformation tools** help design, build, and execute jobs that transform data in the ways described earlier.
- **DataOps tools** help continuously integrate and continuously deploy (CI/CD) pipeline code, test code functionality, and observe both pipeline performance and data quality.

> **Orchestration tools** automate the workflows that stitch together pipelines and the tools and applications that consume their outputs.

The market for data pipeline management includes data ingestion, data transformation, DataOps, and pipeline orchestration

Related Resources:
Modern Data Pipelines: Three Principles for Success
Why Enterprises Should Implement the Data Mesh with DataOps

Success Factors

The success of technology always depends on more than just cool tools. To understand which factors drive the success of data pipelines, consider the phases of pipeline design, operation, and adaptation.

> **Design**. To design effective data pipelines, data engineers and architects must first align with their company's business priorities by defining and stack ranking use cases according to their expected business value. As they design pipelines to support those use cases, they must scope likely future requirements and how to meet them. They also must embrace graphical tools that support a no code/ low code approach, increasing productivity.

> **Operate**. The more data and analytics engineers can automate the design, execution, and monitoring of pipeline processes, the more they improve efficiency and reduce risk. Engineers can further boost productivity by creating standard pipelines artifacts that they can assemble and reuse like Legos.

> **Adapt**. Data teams must monitor a consistent set of key performance indicators, tracking both technical operations and business health. Data observability tools can help them track KPIs as they

optimize pipeline performance and data quality. When it comes time to adjust pipelines, data teams should make modular changes so as to minimize risk.

Business and data teams must govern all three of these phases by defining the roles of various stakeholders and creating and enforcing rules that control their processes. Data pipelines, and the processes managing them, must integrate with governance-related tools such as catalogs, master data management platforms, and data access management tools.

To design effective data pipelines, data engineers and architects must first align with their company's business priorities by defining stack ranking use cases according to their expected business value

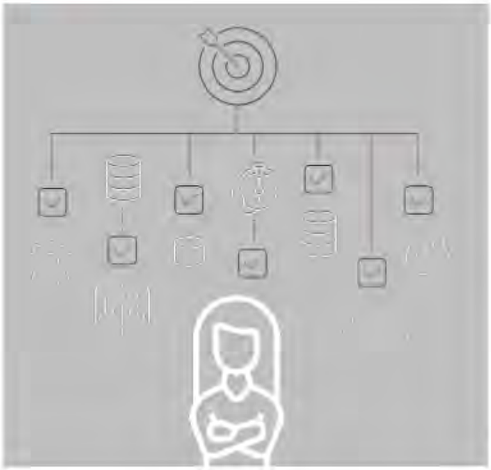
Related Resources:
Flexible Data Pipelines: Balancing Standard and Custom Approaches
How to Design Streaming Data Pipelines for Open, Distributed, and Elastic Cloud Platforms

Benefits

Effective data pipelines serve as the circulatory system for a healthy business. Data teams that design, operate, and adapt pipelines in accordance with the success factors described earlier can deliver tangible benefits to the business. Data analysts, data scientists, and application developers can deliver more analytics value in the form of new insights, faster insights, and new projects, all of which make the business more agile to compete. Data teams can build pipelines that increase data uptime and reduce operational risk. Finally, they consume resources such as cloud compute more efficiently and boost their own productivity.

Effective data pipelines serve as the circulatory system for a healthy business.

Best Practices



Treat your data pipelines as living, adaptable contributors to your business

Make incremental change rather than ripping and replacing tools or processes

Define tool evaluation criteria based on business objectives

Data practitioners and leaders should take an agile, incremental, and business-centered approach to modernizing their data pipelines and environment.

- **Adaptability**. Data engineers and other stakeholders should treat data pipelines as living organisms rather than static systems. New users, use cases, risks, and opportunities will require ongoing changes to your data environment. Such changes might include new sources, targets, platforms, and connections. Design your architecture and pipelines for adaptability by investing in open tools, APIs, and data formats.
- **Incremental change**. Data teams must maintain uptime and resiliency. They must adapt to new business and technical requirements by making incremental changes. Rather than ripping and replacing a platform or set of pipelines, you should phase in new elements in a modular fashion, one at a time, gauging progress along the way.
- **Business-driven evaluation criteria**. When it comes time to select a pipeline tool, be sure to have business objectives shape your evaluation criteria. Advanced, flashy tools might impress engineers, but drive up compute costs or training requirements for business users.

Data engineers and other stakeholders should treat data pipelines as living organisms rather than static systems. This requires an adaptable architecture with open tools, APIs, and data formats

Related Resources:

Best Practices in DataOps: How to Create Robust, Automated Data Pipelines

Challenges and Product Evaluation Criteria

The challenges facing modern data pipelines are five-fold. As described earlier, rising data volumes, driven by data transformation and proliferating data sources, put significant strains on traditional pipelines. An accelerating business environment, meanwhile, drives the need for real-time data delivery and access. In addition, data engineers, analytics engineers, data scientists, and analysts—especially the more business-oriented ones—can struggle to build the right pipeline scripts in SQL, Python, or other popular languages. All the while, data environments grow more diverse, adding elements to accommodate new business demands and projects. A final challenge is compliance with stringent regulations, for example to ensure privacy and avoid bias.

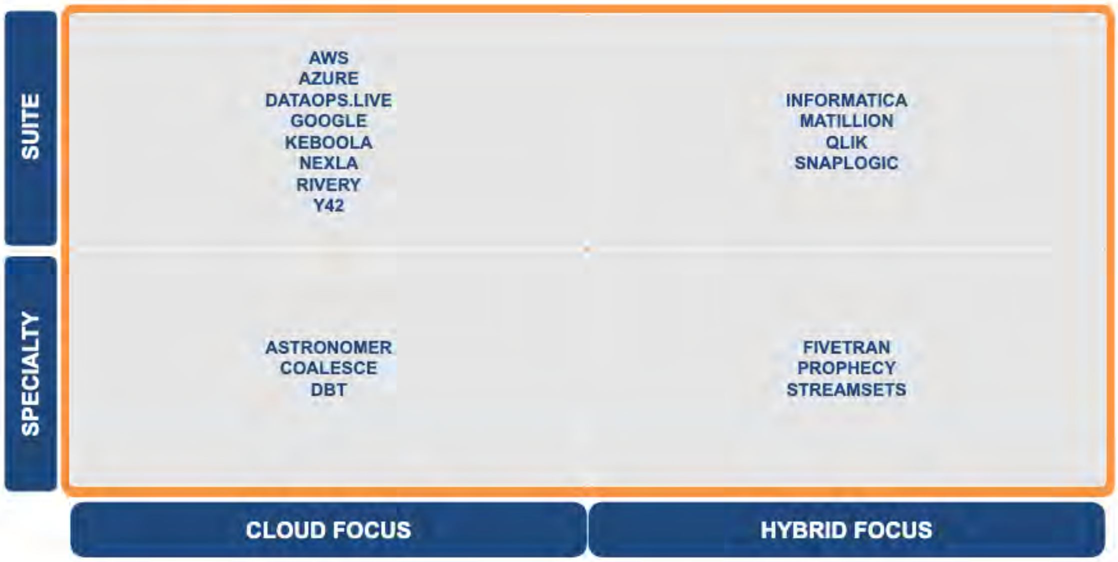
With these challenges in mind, Eckerson Group recommends evaluating data pipeline products according to five criteria.

- **Breadth of functionality.** Your data pipeline tool should support—either natively or via easy integration with third-party tools—multiple data integration patterns, such as ETL, ELT, ELTL, and reverse ETL.
- **Performance and scale**. Your tool should put minimal processing burden on your current architecture, avoid source agents, use log-based change data capture, and meet your business' most rigorous service level agreements (SLAs).
- **Ease of use**. Your tool should require minimal training, automate basic tasks, provide an intuitive graphical interface, easily change pipeline elements, and improve the productivity of your data team.
- **Open architecture**. Your tool should provide an open architecture that works with various sources, targets, processors, and programming languages. Open data formats and APIs can help ensure your tools interoperate with one another and give you the flexibility you need to migrate or integrate data as needed.
- **Governance**. Look for a tool that centralizes pipeline metadata, provides granular role-based access controls, masks sensitive data, tracks lineage, and audits user actions to assist compliance efforts.

Eckerson Group recommends evaluating data pipeline products according to their breadth of functionality, performance and scale, ease of use, open architecture, and governance

Related Resources:
Twelve Must-Have Characteristics of a Modern Data Stack

Product Framework



Eckerson Group recommends categorizing pipeline products according to the following framework.

- **Cloud-focused suite products**. Select products in this category to manage ingestion, transformation, DataOps, and pipeline orchestration with a suite for **cloud** environments.
- **Hybrid-focused suite products**. Select products in this category to manage ingestion, transformation, DataOps, and pipeline orchestration with a suite for **hybrid** environments.
- **Cloud-focused specialty products**. Select products in this category to focus on specific aspects of data pipeline management for **cloud** environments.
- **Hybrid-focused specialty products**. Select products in this category to focus on specific aspects of data pipeline management for **hybrid** environments.

The data pipeline market divides into four categories: cloud-focused suite products, hybrid-focused suite products, cloud-focused specialty products, and hybrid-focused specialty products.

Key Takeaways

Data environments get more complex by the day. New users, use cases, sources, targets, etc. put strain on the pipelines that must deliver timely, accurate data for analytics

The market for data pipeline management includes four segments:

> **Ingest:** extract and load data from source to target.

> **Transform:** filter, merge, and format data for consumption.

> **DataOps:** optimize pipelines with CI/CD, testing, and monitoring.

> **Orchestrate:** schedule and execute workflows across pipelines and applications.

Users of data pipelines include data engineers that specialize in pipelines and data analysts and scientists that prepare data to support analytics. They also include analytics engineers that transform and validate data for consumption by analysts.

Data pipeline tools divide along two axes: specialty vs. suite, and cloud vs. hybrid focus.

Suite cloud products help manage ingestion, transformation, DataOps, and pipeline orchestration with a suite for cloud environments.

Suite hybrid products offer these capabilities for hybrid environments.

Specialty cloud products focus on specific aspects of data pipeline management for cloud environments.

Specialty hybrid products have a similar focus on specific aspects of pipeline management for hybrid environments.

Product Profiles

ASCEND.IO

Ascend.io

As CTO of a personalized video platform in the 2010s, serial entrepreneur Sean Knapp watched his team struggle to integrate various point tools for data delivery. He founded Ascend.io in 2015 to untangle this problem with a solution for automating and orchestrating data pipelines. With $69 million of funding from backers such as Sequoia Capital, Ascend.io now helps customers such as Mattel, News Corp, and Harry's ensure timely, accurate, and efficient data delivery.

Ascend.io's platform automates the creation, control, and operation of change-aware data pipelines. It enables data engineers and code-writing analysts to ingest, transform, and share data between legacy and cloud endpoints. By "fingerprinting" all versions of data and pipeline code, they can identify and propagate any changes in a controlled fashion through a network of pipelines. They also can observe pipeline operations to optimize performance, data quality, and compute cost. Together these capabilities enable data teams to reduce the risk and effort of preparing data for analytics. They also can use Ascend.io to publish, consume, and reuse data products in data mesh environments that span multiple clouds.

Ascend.io differentiates itself in two ways. First, its platform addresses all four segments of Eckerson Group's definition of data pipeline management—ingestion, transformation, DataOps, and orchestration—as well reverse ETL, to help data engineers reduce the need for point tools. Second, Ascend.io's fingerprinting technology gives data engineers a granular view of metadata across the environment, enabling enterprises to streamline data pipeline automation while enforcing cloud consumption best practices.. Its ideal customers are enterprises new to data modernization and cloud-based analytics, as well as enterprises that need to streamline what they've already moved to the cloud.

ASTRONOMER

Astronomer

In 2014, Maxime Beauchemin, then a data engineer at fast-growing Airbnb, created Apache Airflow to orchestrate data workflows across their data stack. He sought to replace rigid and vendor-specific job schedulers with a more programmatic, multi-platform approach to data orchestration. As public adoption exploded, a trio of serial entrepreneurs based in Cincinnati–Tim Brunk, Ry Walker, and Greg Neiheisel–

founded Astronomer in 2018 to deliver enterprise support and packaging for Airflow. Now led by CEO Scott Yara, Astronomer employs 15 of the top 25 contributors to the Airflow project. With more than 16 million user downloads of Airflow each month, Astronomer's addressable market continues to grow.

Astronomer's Astro managed service seeks to help data engineers boost their productivity and make Airflow more reliable and scalable. Astro helps orchestrate workflows across hundreds of third-party tools, including offerings from vendors such as Fivetran, dbt, Snowflake, and Databricks. Astronomer's ideal customers are data and now ML engineers that use Airflow to write, schedule, monitor, and optimize the intricate task sequences that comprise modern data pipelines. Assisted by the acquisition of Datakin last year, Astro supports DataOps with data observability and lineage. Astronomer raised $283 million from investors that include Sutter Hill Ventures, Bain Capital, JP Morgan, and Salesforce Ventures, and has thousands of customers that include Marriott, Conde Nast, and StockX.

Coalesce

Modern enterprises struggle mightily to transform multi-sourced data at scale. Armon Petrossian and Satish Jayanthi, both veterans of the DW automation vendor WhereScape, founded Coalesce in 2020 to tackle this problem. Their mission: empower data engineers and architects to transform more data with less effort on Snowflake. Petrossian and Jayanthi came out of stealth mode two years later to release their first product. Now a series A startup, they have raised $32 million from backers such as Emergence Capital and GreatPoint Ventures. Initial customers include FCP Euro, TotalEnergies, and Burnco.

Coalesce's ideal customers are enterprise data teams that need to improve the productivity, scalability, and governance of their transformation jobs–and get data warehouses into production faster. Coalesce differentiates itself from other pipeline tools by using granular metadata to apply transformation jobs to columns rather than tables. This helps automate the creation of data objects such as dimension tables, perform impact analysis, and track lineage at the column level. Users can build, edit, package, and share various automation templates to improve team productivity, all through a graphical interface that minimizes manual scripting. Coalesce also differentiates itself with flexibility and extensibility, integrating with the GitHub software development platform and ingestion tools such as Fivetran.

DataOps.live

CEO Justin Mullen and CTO Guy Adams co-founded London-based DataOps.live in 2020 as a software spinoff from their data consulting business. Their stated mission: "to change the way the world builds, tests and deploys data platforms and data products," with an initial focus on the Snowflake Data Cloud. They also helped shape the philosophy of the TrueDataOps community. DataOps.live differentiates itself by providing a control layer that orchestrates, monitors, and controls the many elements of modern data pipelines. It best serves mid-sized and large enterprises that want to improve productivity with a unified developer and operator experience on Snowlake. Oher cloud data platforms are on their near-term roadmap.

The DataOps.live platform enables data engineers and data product managers to configure, build, test, and release data pipelines and data products in heterogeneous environments. This includes the ability to ingest and transform data natively or via third-party tools. For example, DataOps.live's latest orchestrator feeds metadata into Vaultspeed, then tests and automatically deploys its modeling outputs for data vault environments. To improve developer productivity, DataOps.live integrates with Kubernetes containers, the Streamlit library of python scripts, and Snowpark APIs for running diverse programs on Snowflake. DataOps.live's new release, now in private preview, supports "Data Products" as complex objects and traces their lineage. It shares metadata with observability tools such as Monte Carlo and catalogs such as Collibra, and helps orchestrate more than 25 other products. DataOps.live customers include Roche and OneWeb.

dbt

The steep rise of dbt started in 2016 when co-founders Tristan Handy, Connor McArthur, and Drew Banin decided it was time to remove traditional barriers between data analysts and data engineers. They launched dbt Labs to create the intermediary role of "analytics engineers" that help make data more readily available to analysts. These analytics engineers clean, model, test, deliver, and document datasets for consumption. They use the dbt open-source framework to handle all this work using standard software development practices such as modularity, portability, and CI/CD. For example, by writing SQL select statements they can build interrelated models that become data warehouse tables and views. Such capabilities target digitally-oriented companies whose data teams prefer scripting.

Capabilities like these enable 50,000 community users of dbt to support BI reporting, dashboards, ML models, and data-driven applications on cloud data platforms. dbt Labs has 3,000 customers, including both traditional enterprises such as JetBlue and Nasdaq and cloud-born startups such as Acorns and Vendr. Now at the Series D funding level, dbt Labs has raised $414 million from such backers as Altimeter, Databricks, Snowflake, and Salesforce Ventures. Its many integration partners include Fivetran for ingestion, Alation for cataloging, Monte Carlo for data observability, and Rudderstack for customer data management.

Fivetran

Cloud adoption in the 2010s exposed the architectural limitations of traditional data pipelines. ETL and change data capture tools, designed to extract records from databases on premises, struggled to move data from new SaaS applications to cloud data warehouses in a simple and reliable way. George Fraser and Taylor Brown launched Fivetran in 2012 to fill this market need with an automated ELT solution. Fivetran targeted data engineers with large and mid-sized enterprises that sought a cost-effective, no-code graphical interface for ingesting multi-sourced data into cloud data warehouses. Fraser and Brown, friends since childhood, stole market share from incumbent pipeline vendors as they raised $728 million in funding and grew to more than 1,100 employees.

Fivetran beefed up its capabilities for database replication in 2021 by acquiring HVR, which specializes in high-volume data replication from tricky legacy systems such as Oracle, Db2 z/OS, and SAP. Fivetran also integrates with dbt to help users automatically transform data after loading it into the data warehouse. Data or analytics engineers can use Fivetran's pre-built dbt packages to create data models, define relationships between them, then materialize those models as analytics-ready tables and views in the data warehouse. Fivetran also offers Quickstart transformations that enable users to orchestrate and manage model runs within its graphical interface, without the need for a separate dbt project, Git repository, or third-party tools. Fivetran now has thousands of customers, including Conagra Brands, Autodesk, and Morgan Stanley.

Informatica

Informatica Cloud Data Marketplace turns assets from the Informatica Catalog into data products in the Cloud Data Marketplace. It's a service in the Informatica Intelligent Data Management Cloud.

The largest vendor profiled here, Informatica offers data pipeline tools as part of a comprehensive portfolio for data management and governance. The Redwood City, CA-based company started as an ETL provider in 1993. It grew organically and via acquisition to address critical segments such as cataloging, data quality, data privacy, and master data management, all integrated as modules on an AI-driven platform. Its data pipeline capabilities span ingestion, transformation, DataOps, and pipeline orchestration across hybrid and multi-cloud environments. Informatica seeks to standardize and accelerate data delivery with metadata, automation, and AI. For example, it guides data engineers and data analysts with prompts and recommendations as they automate the design, creation, testing, deployment, and monitoring of pipeline code.

Informatica's ideal data pipeline customer is the data engineering team within a large or mid-sized enterprise that needs to modernize its hybrid environment and shift analytics workloads to the cloud. KLA, for example, uses Informatica to migrate and continuously synchronize on-premises Oracle ERP records to a Snowflake data warehouse. Twitch, meanwhile, uses Informatica to integrate both data and applications to support its interactive live streaming service. Informatica has more than 5,000 customers, makes more than $1.5 billion in annual revenue, and trades on the New York Stock Exchange.

MATILLION

Matillion

In 2011 CEO Matthew Scullion and CTO Ed Thompson hatched plans to fix a problem they'd lived as consultants: enterprise BI projects needed faster, simpler access to structured data. Seated at Matthew's kitchen table, they founded Matillion as a cloud-based solution to this problem. They started with a BI service for AWS, then pivoted to help accelerate and simplify data transformations on cloud data warehouses. Matillion launched its flagship ETL product for Amazon Redshift in 2015. Over time it expanded to other platforms, grew to more than 650 employees, and raised $290 million from investors such as Databricks, Snowflake, and Battery Ventures.

The ideal customers for Matillion Data Productivity Cloud are traditional and cloud-born enterprises that need to integrate the many moving parts of modern data pipelines: proliferating data sources, frequent schema changes, you name it. Data engineers, architects, and scientists use Matillion to ingest data from hybrid sources into cloud data warehouses, then perform no-code, low-code, or scripted data transformations based on preference. They also can orchestrate data pipelines and synchronize updates across hybrid and multi-cloud environments. Matillion differentiates itself with flexibility, ease of use, and an elastic cloud-native architecture. Its 1,500+ customers include names such as Cisco, DocuSign, Pacific Life, Slack, and TUI.

Nexla

One element of the data mesh has gained more market traction than others: creating and sharing data products. However, implementing data products has proven difficult because it requires time and resources. Enter Nexla, which aims to make data products easier by automating various aspects of data engineering. CEO Saket Saurabh started Nexla in 2016 along with several alumni of his prior venture, the mobile advertising platform Mobsmith, because they believed enterprises needed a new way to minimize the silos and friction of data delivery. Saurabh and team raised $15 million in seed and series A funding from investors such as Industry Ventures, Blumberg Capital, and Correlation Ventures, and now serve customers that include Doordash, LinkedIn, J&J, and American Express – Global Business Travel.

Nexla's ideal user is a large or mid-sized enterprise that seeks to foster team collaboration and deliver timely, consistent, and reusable data products to the business. Nexla automates data integration, preparation, monitoring, and discovery in environments that span data warehouses, databases, files, and various other end points. It offers a no-code method to prepare, inspect, and publish modular data products, called Nexsets, that comprise virtual views, schema, and other metadata. Data engineers can assemble and reassemble combinations of Nexsets like Lego blocks to support various analytics projects with evolving business requirements. Nexsets encompass batch and streaming delivery methods, as well as ETL, ELT, reverse ETL, or API integration.

Qlik
LEAD WITH DATA

Qlik

In 2005 an Israeli company called Attunity released new change data capture (CDC) software for loading database records into data warehouses in real-time increments rather than periodic batches. By eliminating duplicative batch processes, CDC boosted efficiency and accelerated analytics. This innovation helped enterprises ingest data into Hadoop data lakes, then cloud data warehouses such as Snowflake, and now lakehouses such as Databricks–as well as myriad targets on premises. Attunity gave DBAs and data engineers a graphical interface to simplify pipeline configuration, and automated the transformation of data once it arrived at the analytics target. In 2019 Qlik acquired Attunity to couple these pipeline capabilities with its Qlik Sense business intelligence platform.

Qlik's ideal customers are the data teams within mid-sized and large enterprises that need to modernize their hybrid environments for real-time analytics. Qlik helps data analysts and engineers automate the configuration, execution, and monitoring of pipelines that ingest and transform data for analytics. It differentiates itself in the data pipeline market with ease of use and deep integration with complex legacy systems such as mainframe, SAP, and IBM i. Mid-sized and large enterprises use Qlik to migrate operational systems to the cloud and continuously synchronize real-time operational data with cloud analytics platforms–as well as a variety of on-premises targets. Once the data arrives at the target, Qlik automates formatting and table consolidation within data lake landing zones or cloud data warehouses. It also catalogs that data, analyzes it with Qlik Sense, and triggers automated application tasks based on analytical outputs.

Prophecy

Prophecy

In 2017 Raj Bains, a longtime product manager and engineer for analytics platforms, launched Prophecy to help enterprises simplify the next generation of data lakes—soon to be called lakehouses. Bains and co-founders Vikas Marwaha and Maciej Szpakowski built a graphical user interface and platform for transforming data on Databricks and other popular data platforms and data warehouses. Since raising $38.5 million, Prophecy has grown to more than 80 employees, and now serves customers across industries including the Texas Rangers, HealthVerity, Waterfall Asset Management, and a Fortune-50 pharmaceutical company.

The ideal Prophecy customer is an enterprise that needs to modernize many homegrown data pipelines for analytics and machine learning on the cloud. Prophecy helps such customers democratize data engineering — empowering all data users to create, publish, and subscribe to data products that combine data, metadata, and pipeline code. Its visual interface and extensible framework enables data engineers, data scientists, and data analysts to build transformation pipelines with Apache Spark and now SQL using agile design practices.

Prophecy differentiates its platform with visual design, extensibility and portability based on open standards, and breadth of functionality. It addresses four primary use cases: migration, data engineering, self-service, and lakehouse management. While Prophecy focuses on the Databricks Lakehouse and Spark, it also supports SQL data warehouses, Snowflake Data Cloud, BigQuery and on-premises sources. It standardizes on GitHub for software development best practices, Airflow and Databricks Workflows for workload orchestration, and dbt Core for SQL transformations.

Rivery

Rivery

For two decades Itamar Ben Hemo helped enterprises build and manage data warehouse environments. So when he co-founded Rivery with longtime colleagues Aviv Noy and Alon Reznik in 2017, he understood the need to streamline DataOps processes for CI/CD and testing. Ben Hemo and team devised a low-code/no-code ELT tool that helps data engineers build and orchestrate pipelines for the many cloud-based endpoints on which modern enterprises depend. Now a Series B venture, Rivery has $50 million in funding from backers such as State of Mind Ventures and Entrée Capital. Its customers, numbering in the hundreds, include Bayer, American Cancer Society, and Buzzfeed.

Rivery is ideal for hybrid and cloud-native enterprises that need to integrate multi-sourced data on cloud platforms for analytics. It automates data ingestion, from SaaS applications, cloud databases, and digital ad platforms in particular, then speeds the transformation of that data on analytics targets such as Snowflake and Databricks. Data engineers and even analysts use Rivery to design pipelines, build them with CI/CD practices, then test and debug those pipelines before deploying them in production. They also use Rivery to orchestrate pipelines, automating their workflows and interactions with applications. Rivery differentiates itself with starter templates that help speed data delivery by combining multiple data sources into analytics-ready models. Another differentiator is automated and transparent workflows that combine ingestion and Python-based transformation, with no need for a separate tool such as Airflow.

StreamSets

Streamsets

StreamSets was born in 2014 when CEO Girish Pancha decided data ingestion was "the single biggest barrier to a successful analytics platform," especially "when the data is constantly shifting underfoot." Pancha and Arvind Prabhakar, both alumni of Informatica, founded StreamSets to help enterprises overcome this barrier by automating the extraction and loading of data. They focused in particular on detecting and managing "data drift"—those gradual changes to source schema, metadata, and semantics that can force teams to manually reconfigure data pipelines. Streamsets raised $77 million from investors such as Battery Ventures and New Enterprise Associates (NEA) and amassed 150 customers before its acquisition by Software AG in 2022.

StreamSets' ideal users are data engineers that need to simplify how they design, test, deploy, and monitor the pipelines that feed cloud analytics platforms. It automates the configuration and control of basic tasks while enabling data engineers to design, import and customize complex transformation code to support advanced use cases. StreamSets' consolidated interface enables users to design and build batch, streaming, CDC, ETL, and ELT pipelines. Its pipeline diagrams, drag-and-drop processes, and consolidated views can help data teams centrally manage diverse pipelines, standardize and reuse pipeline elements, and reduce the time they spend managing data drift. StreamSets, whose customers include Shell, IBM, and Availity, also complements the application integration features of parent Software AG.

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

GET • MORE • VALUE • FROM • YOUR • DATA

About the Sponsor

At **StreamSets,** a Software AG company, we believe in the audacious, ambitious goal of teasing order out of the chaos of modern data. We help our customers achieve that goal by ensuring data engineering teams thrive in today's world of constant change. StreamSets brings enterprise-proven DataOps capabilities to modern data integration, enabling continuous data for the modern data stack.

The StreamSets vision for modern data integration is guided by DataOps, a set of practices and technologies that operationalizes data engineering and integration to ensure resilience and agility despite constant change. StreamSets technologies are architected with a modern approach to data integration and data pipeline operations.