Data Management 2022

Trends, Technologies, Teams, and Organizations

Rivery

In 2022, data professionals will face new opportunities and challenges as they navigate a world where data is growing in size and complexity. The rapidly-changing data landscape is driving novel trends that will impact data-driven technologies, teams, and organizations in 2022 and beyond.

Amid these fast-moving developments, companies will have to make key decisions about the modern data stack, personnel, data mesh, and so many other emergent possibilities. That’s why we’ve developed this new eBook for you. We’ve created a single resource for all the hot topics and new capabilities in the data space in 2022.

Read on to learn what the future holds for data management in 2022!

The Trends & Technology that Will Define Data Management in 2022

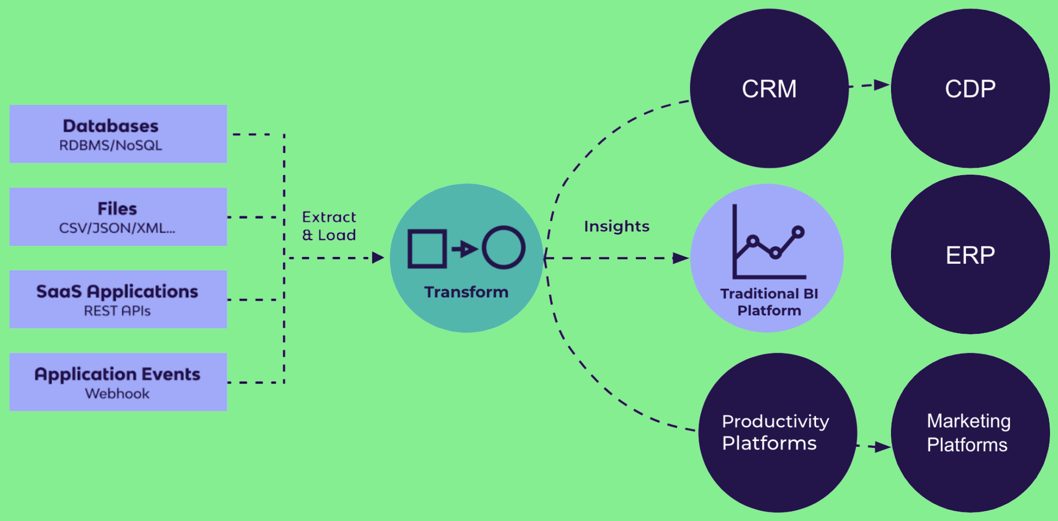
1. Reverse ETL

Since the advent of cloud data warehousing, business users have leveraged BI dashboards to make decisions and drive initiatives.

However, the BI process remains full of bottlenecks and inefficiencies, from measuring the wrong metrics, to slow roll outs, to low utilization. BI dashboards also do not allow business users to manipulate and operationalize data in business processes.

Reverse ETL bypasses dashboarding altogether by pushing data directly into 3rd party systems (CRM, CDP, ERP, etc.) for direct usage. In 2022, reverse ETL will become a key component of the modern data stack for many companies.

REVERSE ETL Process



True Business Wide Data Democratization

ETL and ELT both transfer data from third-party systems, such as business applications (Hubspot, Salesforce) and databases (Oracle, MySQL), into target data warehouses. But with reverse ETL, the data warehouse is the *source*, rather than the target. The target is a third-party system. In reverse ETL, data is extracted from the data warehouse, transformed inside the warehouse to meet the data formatting requirements of the third-party system, and then loaded into the system for action.

By pushing data back into third-party systems such as business applications, reverse ETL operationalizes data throughout an organization. Reverse ETL enables any team, from sales, to marketing, to product, to access the data they need, within the systems they use. The applications of reverse ETL are numerous, but some examples include:

- Syncing internal support channels with Zendesk to prioritize customer service
- Pushing customer data to Salesforce to enhance the sales process
- Adding product metrics to Gainsight to improve the customer experience
- Combining support, sales, and product data in Hubspot to personalize marketing campaigns for customers

What teams really want is to access data *within* the systems and processes that they're already using. This is exactly what reverse ETL enables you to do. With reverse ETL, business users can actually harness data in an operational capacity. Teams can act on the data in real-time via change data capture (CDC), and use it to make key decisions, while leveraging BI dashboards as supplementation.

2. Analytics Engineer

As the gap between business and technical teams continues to grow, a new role has emerged to bridge the divide.

The Analytics Engineer is a relatively new position that fills the gap between data engineers and data analysts. The role has grown in popularity recently, and in 2022, that trend will continue to grow.

Analytics Engineers build data models based on business requirements to produce analytics for stakeholders and teams. Unlike data analysts, analytics engineers do not analyze data; they transform, test, deploy, and document data for business users. Analytics engineers apply version control, CI/CD, and other SE best-practices to the analytics code base to ensure the quality, consistency, and value of data. They are, in short, the role that closes the chasm between business and IT.

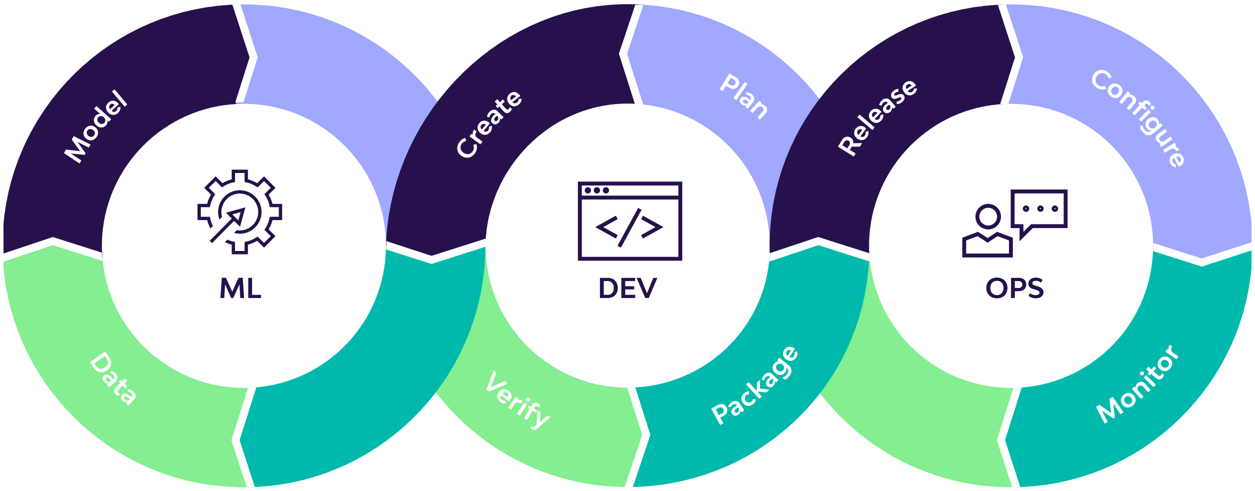
And this is what makes Analytics Engineers so important for data-driven companies. MPP databases have abstracted away the difficult computer science problems, democratizing the ability to process large amounts of data. Today, what companies need is a "first-hire" role that can set up and activate the modern data stack (more on this topic later).

An Analytics Engineer is someone who is technical enough to launch a cloud data warehouse and data pipelines, but also someone who has the business sense to translate and materialize the analytics requirements of users. Analytics Engineers can set up basic data infrastructure and visualizations — an appealing prospect for companies that do not yet need deep engineering firepower.

3. MLOps

In 2022, advances in data management will enable teams and companies to unlock new capabilities.

This includes functions such as MLOps. MLOps is a set of practices and protocols that deploys and maintains machine learning models in production. Just as DevOps improved software engineering, and DataOps improved data engineering, MLOps is improving data science.



MLOps applies to the entire ML lifecycle, from data gathering, model creation, CI/CD, and orchestration to deployment, health, diagnostics, governance, and business metrics. Key components of the MLOps cycle — such as data gathering and data transformation — are streamlined by data management platforms.

4. Active Metadata

Metadata has always played a key role in data management. However, with the emergence of the modern data stack, metadata management has lagged behind other components.

Companies can set up a data warehouse in less than an hour, but building out data cataloging and other metadata-centric solutions often takes months. Meanwhile, friction between cross-functional teams, from engineers to salespeople, has made metadata management more important than ever.

Enter active metadata. Active metadata refers to data that defines data, including data about what happens to the data. Combining technical, operational, business, and social metadata, active metadata is capable of driving faster, action-based data management deployments, from alerting to operationalizing insights.

> "*One of the strong themes for Gartner is the idea of active metadata,*" TopQuadrant CEO, Irene Polikoff, acknowledged in a recent article. "*One aspect of that is it's directly actionable; it's actually used in real-time by operational systems to do various things.*"

Active metadata enables quicker and more scalable solutions for data cataloging, lineage, discovery, and governance, enhancing the speed and flexibility of the modern data stack. That's why Garnter recently retired its Magic Quadrant for Metadata Management Solutions and introduced the Market Guide for Active Metadata. And in 2022, this trend will continue to become more pronounced.

5. Data Governance

In 2022, a governance framework will no longer be a passive set of protocols, but a dynamic factor in analytics and decision making.

For the past decade, companies built data governance frameworks for data privacy, regulatory compliance, and other areas of risk management. But now, with the advent of active metadata, data governance can also function in an operational capacity.

Data governance frameworks that leverage active metadata and operational functionality will enhance data modeling, data stewardship, AI/ML, and more across an organization. Active metadata will also enable data governance frameworks to offer transparency of data profiles, classification, quality, location, lineage, and context. Moreover, advances in data intelligence will democratize data while ensuring security.

Governance solutions will facilitate real-time communications between different systems in a data fabric. New "metadata graphs" will enable more granularity in data lineage and BI, anomaly analysis, and impact analysis. Additionally, governance frameworks will leverage smart inferences based on active metadata to automate compliance measures for similar sources and users.

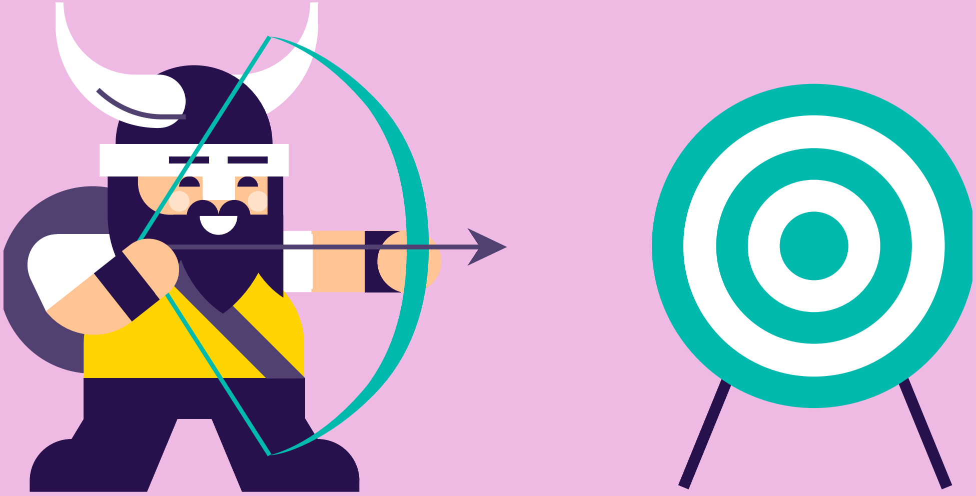
The days of passive data governance are over. In 2022, active metadata will make data governance an operational force.

6. Data as a Product

With the introduction of the data mesh (more below), the concept of "data as a product" has boomed in popularity.

A "data product" is typically a digital product or feature that uses data to achieve its end. But data mesh also discusses the productization of data — the application of the product life cycle to data deliverables.

In this new paradigm, data assets are treated as products, and data consumers are treated as customers. This framework applies product thinking to datasets, and ensures that they retain the discoverability, security, explorability, and other facets needed to remain leverageable entities.



Hyperspecialization vs. Data Mesh:

Bridging the Divide

Rivery

As business complexity and data size increases, companies must reassess how they can extract value out of data at scale.

Creation, capturing, copying, and consumption of data went up by a whopping 5000% between 2010 and 2020, and 63% of companies cannot gather insights from their big data.

As companies confront this deluge of data and advanced use cases, the amount of niche data roles continue to multiply. However, as data operations grow, this practice is likely unsustainable. But a competing vision has emerged: data mesh, a decentralized, domain-specific architecture. However, is the data mesh framework feasible? Or is hyperspecialization the only way forward?

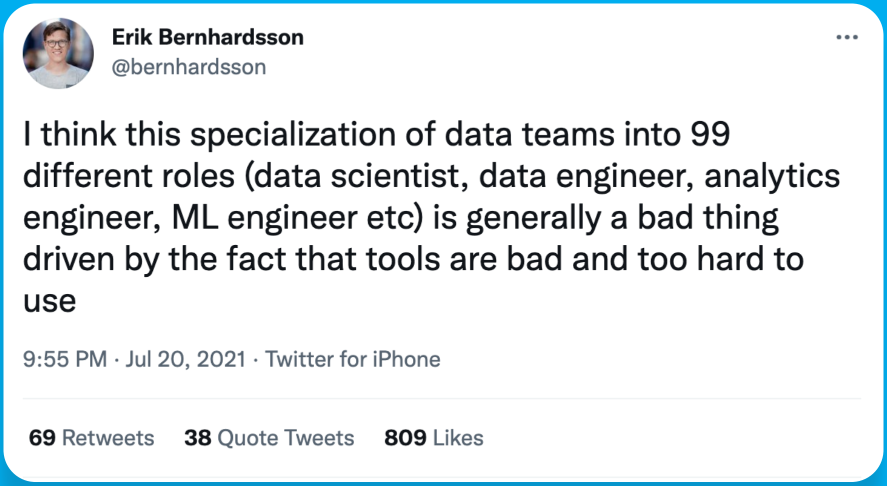
Hyper-Specialization: Is Tooling the Problem?

In the data space, hyper-specialization refers to the proliferation of all the different "expert" roles within a data team, as opposed to smaller teams with more generalized personnel.

At many companies, hyper-specialized data teams build and operate the organizational data platform. Each role is "hyper-specialized" — from data engineer, to data scientist, to data architect — for specific tasks.

But these hyper-specialized data teams are often siloed from the other business units of the organization, and even from each other. They lack domain-specific knowledge about how other teams operate and what the key business objectives are as a whole. This creates a fundamental misalignment between the data team and data customers.

This status quo is far from ideal, but perhaps hyper-specialization is just the cost of today's complex modern data stack? However, some industry commentators have other explanations. Erik Bernhardsson posited that *data tools* are leading to hyper-specialization, in one of last year's most pot-stirring tweets.



Erik Bernhardsson
@bernhardsson

I think this specialization of data teams into 99 different roles (data scientist, data engineer, analytics engineer, ML engineer etc) is generally a bad thing driven by the fact that tools are bad and too hard to use

9:55 PM · Jul 20, 2021 · Twitter for iPhone

69 Retweets **38** Quote Tweets **809** Likes

Bernhardsson went on to clarify in a blog that he does support specialization in general. But he also said that many tools *"require so much knowledge to use"* leading to "*wasting way too much time debugging YAML, waiting for deployments, or begging the SRE team for help.*" This raises an interesting question - are tools causing hyper-specialization? Or are they just making engineers miserable with manual, mind-numbing tasks?

Data Mesh: The Alternative Vision

In the past several years, a number of solutions and frameworks have attempted to minimize hyper-specialization. One of these proposed solutions is the **data mesh.**

Coined by Zhamak Dehghani in 2019, the data mesh is one of the hottest concepts in data architecture. Data mesh is an organizational paradigm that addresses the "failure modes" of data lake architecture, including "siloed and hyper-specialized ownership."

Data mesh supplants centralized data lakes/data warehouses with decentralized distributed architectures that enable self-service for domains within an organization. Some of the core concepts include:

- ✓ **Decentralized data ownership -** Data ownership is decentralized and handled by the business domains that source and use the data, rather than a centralized data team.
- ✓ **Distributed mesh -** Data warehouses and lakes are replaced with a mesh of data accessed through shared protocols.
- ✓ **Unitary data and code -** Data is not a subsidiary of code; data and code are treated as a single unit.
- ✓ **Federated model -** Data governance is no longer top-down and reliant on human labor; it is powered by computational policies intertwined in the mesh.
- ✓ **Data as product -** Data is served to delight customers; data is not an asset to connect to.

As a federated, self-service system, data mesh enables data customers to own and access data directly, decreasing the need for walled-off, hyper-specialized teams. However, data mesh is a relatively new concept, still under adoption, and also has some critics. So, in a world between the present conundrum (hyper-specialization) and the future ideal (data mesh), how can companies start making progress right now?

Simplify Your Data Stack

The modern data stack is complicated and fragmented. Companies leverage multiple tools to power ingestion, transformation, orchestration, reverse ETL, and other data processes.

This hodgepodge of tools causes slow time-to-value, poor scalability, inefficient projects, and high overhead costs, including more personnel.

Returning to Bernhardsson's tweet, there's an element of truth to his claim about data tools. Tools are not the sole cause of hyper-specialization, but talk to any data professional, and they will tell you that fragmented technology is a contributing factor. Now, tools themselves have become hyper-specialized, complicated to manage, and fractured.

As companies make the years-long move to the data mesh, perhaps an immediate way to diminish hyper-specialization is to **simplify your data stack**. Adopt end-to-end tools that do not require several different solutions and different team members to operate. By simplifying your data stack, you can make improvements to hyper-specialization now and be prepared for the data mesh in the years to come.

The Modern Data Team for the Modern Data Stack

Rivery

As the modern data stack continues to evolve, the modern data team is changing right alongside it.

By eliminating complicated engineering problems, the modern data stack offers the potential for a self-service data infrastructure that empowers business users to access data with ease. At the same time, key technologies — such as reverse ETL — are unlocking the business potential of data like never before.

In this fast-changing landscape, data teams are adjusting to help organizations maximize the value of data. Data teams are leveraging new personnel, processes, and technologies, to move from a centralized, top-down data hierarchy to an embedded horizontal model that enables business users to own data domains.

Read on to learn more about what the modern data team will look like in 2022.

What is the Modern Data Stack, Anyway?

The modern data stack is a term that gets mentioned so much, it starts to become divorced from its textbook definition.

The definition, of course, has been tweaked by many commentators. But typically, the "modern data stack" refers to a cloud-native data platform that reduces the complexity of older OLTP-based systems. In most definitions, the modern data stack includes these core components:

- ✓ Cloud data warehouse
- ✓ ETL/ELT platform
- ✓ BI tool

The modern data stack enhances scalability and automation by utilizing managed services (such as SaaS), OOTB platforms, and SQL-native architectures. The modern data stack is comparatively low cost. Functionally, the modern data stack is easy to use and set up, with launchtime measured in hours rather than days or months.

The modern data stack has impacted organizations in many ways, especially by reducing engineering burdens. By minimizing complex data architecture and system maintenance, the modern data stack allows data teams to focus less on infrastructure issues and more on serving valuable data to customers.

And data customers themselves are gaining more agency with the modern data stack. Data customers can now launch data infrastructure autonomously, run SQL queries against a cloud data warehouse, and much more. This moves customers closer to self-service, creating a horizontal data culture that leads data teams to become data enablers rather than data gatekeepers.

Data Teams in 2022: Key Functions

This change leads us to a key question: In this environment, what exactly is the role of data teams in 2022?

It might be tempting to jump straight into naming team personnel. But before roles can be discussed, functions must be defined. And in many companies, the functions of a data team will change in 2022.

The concept of "data as a product" — an idea with roots in the data mesh — has the potential to reorient the way data teams operate. A data product encapsulates everything that a data customer needs to generate value from a business entity's data. Under the "data as a product" framework, the data team is a product team, and applies the product life cycle to data deliverables. By productizing data, data teams can help businesses improve decisions and streamline business processes among customers.

Another key change for data teams is the move from BI dashboarding toward the operationalization of data. In particular, the emergence of reverse ETL is modifying priorities for data teams. By building reverse ETL pipelines, data teams can push data from a cloud data warehouse back into third-party systems (CRM, ERP, CDP). This allows business users to operationalize *actionable* data in their business processes, rather than relying on top-level dashboards. This year, data teams will focus more on putting data directly in the hands of customers.

In 2022, data teams will focus not just on time-to-delivery and data bottlenecks, but also knowledge gaps. With the data mesh, data teams can offload scoping and requirements onto business users, the stakeholders who know their specific domains the best. Horizontal data cultures can also de-silo the data team and reveal the key business objectives of other teams. These changes do not simply guarantee the speed of data; they also improve the quality of the data for business users.

Overall, in 2022, data teams will act more as facilitators, building a self-service data architecture that grants agency to business users in the modern data stack. This framework is as much a matter of efficiency as it is of business value. With the sheer volume of data streaming ever-faster into the data stack, data teams do not have the resources to drive every project. Business users must operate with some autonomy in the stack if companies want to unleash the full potential of this vast trove of data.

Personnel: Data Team Roles in 2022

As the modern data stack moves more toward a data mesh and self-service paradigm, roles and responsibilities on the data team will change.

Data engineers, data analysts, and BI professionals will continue to play key roles. But the outgrowth of other positions, such as the data architect, reflects the realities of a modern data stack that is moving away from IT supremacy.

Here is an overview of some of the key positions for a modern data team in 2022:

Position	Role	Responsibilities	Skills
Executive (CDO)	Drives the team to produce business-ready data for data consumers and leadership.	Team leadership, data governance, driving & proving results	Python, SQL, Google Presentations, Business dashboards
Data Architect	Builds a data governance framework for the organization's data systems.	Promoting trust in data, enforcing policies, implementing security	Python, Perl, Java, SQL, Relational databases, ETL
Data Quality Analyst	Improves the quality and reliability of data for consumers.	Developing systems & processes for delivering high-quality data	ETL, SAS, JavaScript, SQL, Excel, MySQL, Unix/Linux
Data Engineer	Builds, deploys, and maintains the organization's data infrastructure.	Creating data pipelines & data models, delivering structured data	ETL/ELT tools, SQL, NoSQL, Python, CDW, Database architecture
Data/BI Analyst	Manipulates, models, and visualizes data for data consumers.	Dashboarding, producing analytics & business insights	BI tools, SQL, Tableau, Looker, CDW, Database, ETL
Data Scientist	Produces advanced analytics and predictive insights for data consumers.	Designing predictive models, algorithms, advanced math & CS	R, Python, SQL, SAS, Apache Spark, ML tools, D3.js

While some of these roles focus on governance and other key architectural issues inherent to self-service, the list reflects a data team in flux.

Complete business user autonomy remains unrealized for many companies, and many issues are still resolved through IT, often manually and at a great expense to the company. In this regard, the modern data team is still catching up to the promise of the modern data stack.

Data Citizens: Data Team Leaders with Domain Expertise

The final component of the modern data team is not an official part of the team at all. Business users with data skills and domain expertise will be crucial in 2022.

These "data citizens" can leverage low-code and no-code solutions to deploy data pipelines, format data, and bypass overworked data teams. This allows data citizens to perform ETL/ELT tasks and get data to team members with no coding skills.

To optimize massive volumes of data, companies will have to empower team-level data leaders. Data teams are still too overwhelmed with manual data tasks and maintenance issues to extract the value of data for every business domain. And, the bigger the company, the more pronounced this problem tends to be.

This is why data literacy initiatives will remain important in 2022. But data literacy may need to expand past training employees to simply read and understand data. Companies should also consider introducing tooling and technical training to a cohort of engaged business users. That way, a company can build a horizontal data culture from the ground up, merging domain expertise with data autonomy across teams.

Modern Data Teams Will Evolve in 2022

In 2022, modern data teams must evolve alongside their modern data stacks to position their companies for success.

But as the data mesh, self-service, data citizens, and other factors disrupt the current data team framework, the core functions of data delivery and maintenance remain as critical as ever.

As the data landscape evolves, unexpected developments will surely bring more changes to the data team and that is why we will see you again for Data Management 2023!

About Rivery

Rivery is the modern way to build, manage, and monitor data pipelines. Rivery's SaaS platform provides an end-to-end solution for Ingestion, Transformation, Orchestration, and Data Operations.

Data Ingestion

Rivery's universal support for any data source empowers you to ingest all your data in the format and frequency of your choice. Efficient data ingestion and governance begins with control over all your data sources.

Data Transformation

Produce the data your team needs at any time in any format. Rivery's powerful transformation layer refines raw data into business-ready inputs that fuel superior insights, analysis, and decision making.

Data Orchestration

Rivery enables your team to seamlessly connect and orchestrate all your data sources in the cloud, from both in-house and third-party platforms. Create the perfect data ecosystem, with robust, automated processes.

Data Operations

With Rivery, data management is about more than just handling data. Rivery eliminates manual data management tasks and gives teams the power to control how they use their data, for any project, opportunity, or company.

Talk to an Expert

Learn More
rivery.io

For questions:
contact@rivery.io

Follow us: