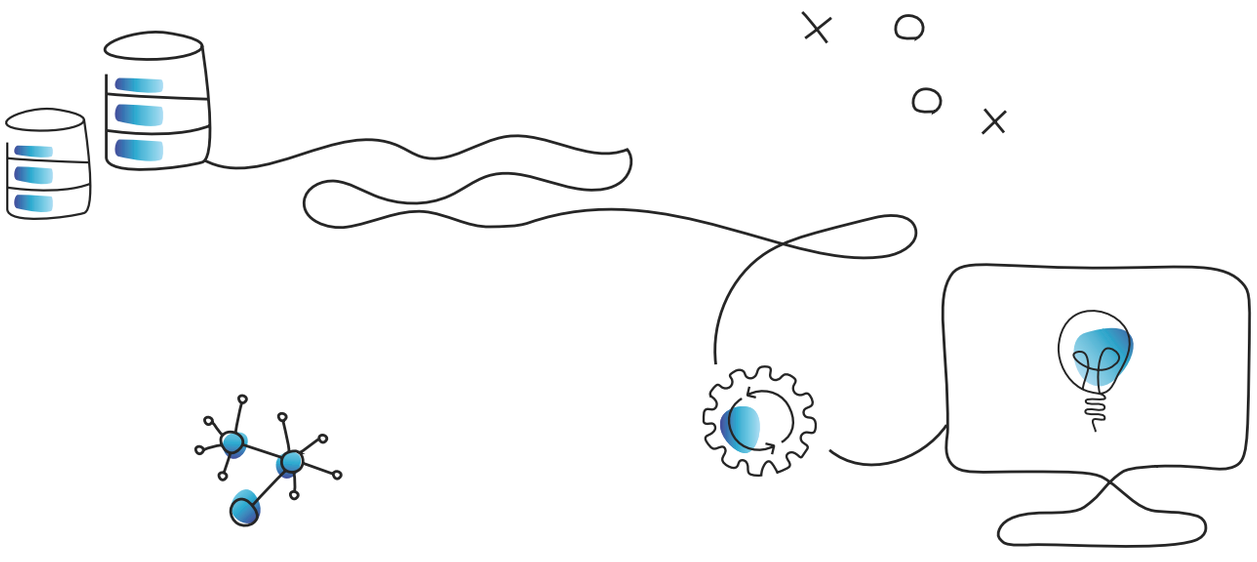
TEAM8™

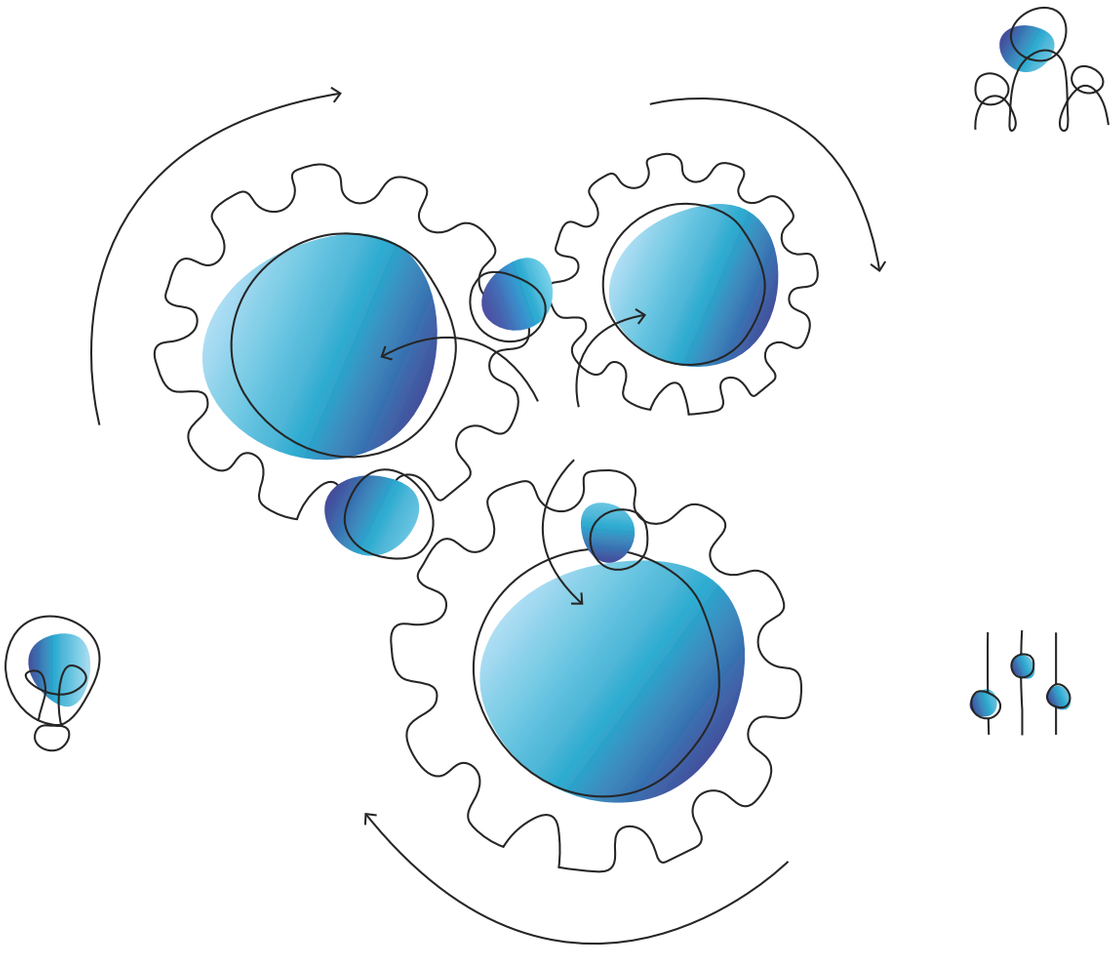
A CDO'S GUIDE

Making Data Work: On the Need for Next-Gen Automated Data Fabric



By the Team8 Data Village

July 2023



The Team8 Data Village is a community of data experts and thought leaders from the world's leading enterprises. The primary focus of the Village is to facilitate collaboration among the world's most prominent companies with the goal of sharing information and ideas, conducting intimate discussions on industry and technology trends and needs, and generating value for all parties.

By helping Team8 to identify real pain points and understand the requirements of large organizations, members of the Village are first in line to leverage solutions that are purpose-built by Team8's portfolio companies to support their needs.

To contact the Team8 Data Village, please email **datavillage@team8.vc**

DISCLAIMER: **These materials are provided for convenience only and may not be relied upon for any purpose. The contents of this document are not to be construed as legal or business advice; please consult your own attorney or business advisor for any such legal and business advice. The contributions of any of the authors, reviewers, or any other person involved in the production of this document do not in any way represent their employers.**

This document is released under the Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

WRITTEN BY

Aviad Harell

Managing Partner,
Team8

Bobi Gilburd

Chief Ideation Officer
Team8

Tom Sadon

Director of Marketing,
Team8

Kalman Heims

Strategy & Business
Research Manager, Team8

Many Team8 Data Village members, and others from the wider community, assisted in the writing, reviewing, and editing of this document.
These are the ones who could share their names publicly:

Angeli Möller, PhD
Head of Data and Integrations
Generating Insights, Roche

Erika McBride
Data Science Manager and AI
Evangelist, Paychex

Dr. Florian Wolf
Founder and CEO,
Mergeflow AG

Sean Rober
Head of Data,
Avenue One

Making Data Work: On the Need for Next-Gen Automated Data Fabric

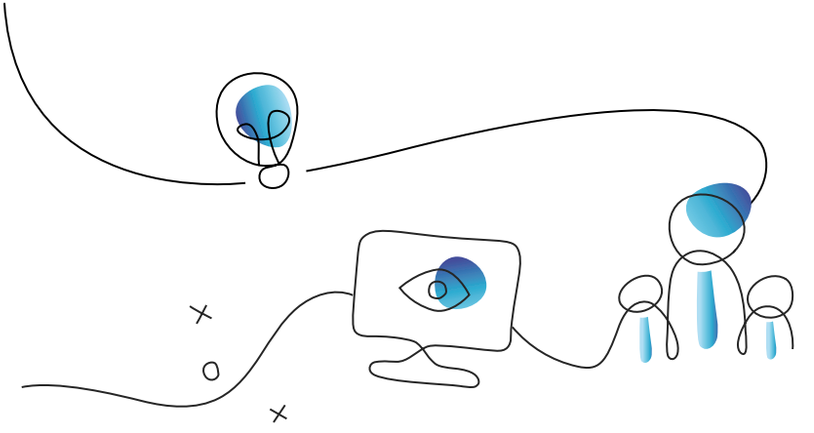
Data is at the center of business operations. Everyone wants more data to empower their business, and the good news is, there's more and more data to leverage. Data and analytics can be used to generate new revenue streams, increase existing revenue and reduce operational costs. Unfortunately, many companies find that they hit roadblocks when trying to extract value from data at speed and at scale.

Per the 2023 survey out of NewVantage Partners, **76.1% of organizations do not consider themselves to be data-driven** even though 87.8% of organizations increased their data investments in 2022, and 93.9% plan to invest even more in 2023.

As the backbone of all business processes, the value of data does not come from its size but from its ability to drive informed decision-making and support smart data-driven business processes. So what is keeping more than three out of four organizations from unlocking this value?

Answering this question becomes fundamental for businesses to stay competitive in today's fast-paced environment, especially as the same report cited above claims 91.9% of the 116 Fortune 1000 executives surveyed saw "measurable business value" from investing in data and analytics.

Below, we explore how to get the most out of data and propose a next-gen automated data fabric as the answer.



Why Is Data Not Working?

Data works only if it successfully drives business insights and processes.

The vast majority of organizations questioned in the NewVantage Partners survey believe that the principal challenge to becoming data-driven is not with technology limitations, such as cost of storage or difficulties in scaling up transformations; nor did they feel investment was lacking.

Instead, 79.8% highlighted cultural issues, including a lack of the right business processes to make the most of the data collected. Here are some of the obstacles organizations are facing today.

Need for Data Specialists

Siloed Data

Lack of Context and Format

Lack of Observability

Need for Data Specialists

Data requires specialized knowledge to be analyzed, and the experts required are in short supply. According to Gartner's 2023 Agenda CDAO survey, close to 40% CDAOs cite "skills and staff shortages" as the top roadblock inhibiting the success of their D&A initiative.

Time-to-insight is lengthy. Building out each query itself takes time. Furthermore, technical teams composed of data engineers and analysts must be involved to collaborate with business analysts when answering every business question.

This high barrier to data literacy causes data to be inaccessible without the right people in place.

Siloed Data

Organizations tend to have their data in silos scattered across different systems and departments. This makes it inefficient, costly, and time-consuming to access data and analyze it properly to deliver value.

Technical teams cannot see what work and queries have already been executed; data engineers and analysts often find themselves rebuilding pipelines and queries to answer similar business questions.

This duplication causes low productivity and efficiency. It may be difficult to share or un-silo data due to non-technical reasons, either real or perceived. For example, there may be incentives to not share data ("I'm the gatekeeper, so I am in control and indispensable"). Or there could be compliance issues, such as different privacy standards across teams or entities within the same multinational organization.

Lack of Context and Format

When data lacks context and format, it is not only from a schema perspective but also from a lack of definitions or support for multiple languages.

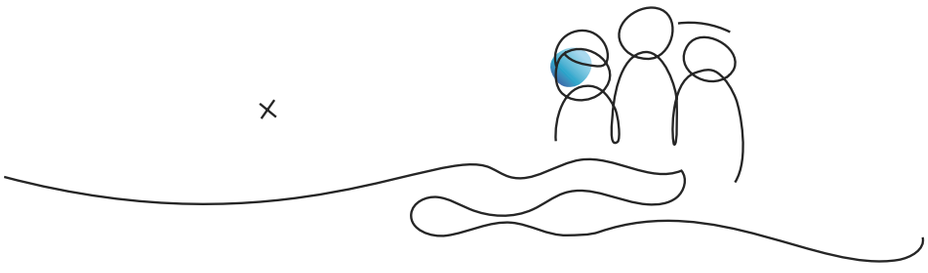
A technical team who has to answer "What is churn?" may find it difficult to figure out the correct definition for their specific business unit. Meanwhile, different teams may also be answering this question as well and coming up with completely different results.

This incoherence creates analyses that are inaccurate, inconsistent, and biased.

Lack of Observability

Data cannot be trusted when there is not full visibility on how it is sourced, stored, and transformed throughout the data stack and how it is used by end users, a full end-to-end data lineage.

On top of the expected data quality issues (e.g., missing values, unexpected entries, unexpected formats, infrequent updates, data decay), this causes business teams to often turn to gut decisions rather than data-driven decisions.



Because of these shortcomings with existing processes, data teams are struggling to prove their ROI to organizations, and their performance is becoming increasingly scrutinized as companies return to efficiency.

Data-driven processes lie at the core of business efficiency. When they don't work, the resulting business processes and insights cannot be fully trusted nor automated. Without the ability to access and analyze data effectively, decision-makers cannot derive critical business data insights in a timely fashion that support business growth and success.

To make data work, a new process needs to be set up so that data can be seamlessly accessed and leveraged, a process that ensures connectivity and coherence in the data stack.

How Do We Make Data Work?

Up until now, companies have made major investments in improving their data stack, resulting in technically advanced systems that offer fast retrieval, detailed access control, full security, and some level of deduplication. From an end-user perspective, relevant data has been collected and augmented, and predictive analytics powered by AI built on top.

So, what's missing?

In our analysis, we have discovered that visibility and reusability have not been achieved in the current data stack. To make data work, we need to ensure that:

- **Data is seamlessly accessible, understandable, and visible.**
- **Full collaboration is the cultural standard.**

To make this happen, business users need to be brought into the data stack to provide the business acumen required to define data entities and questions.

This has been the aim of the metric store, or metric layer. This is an autonomous layer sitting between the raw data warehouses/lakes and the data users, who act via BI/AI tools that support building business logic. This allows business dimensions, measures, and metrics to be defined in a business-friendly language.

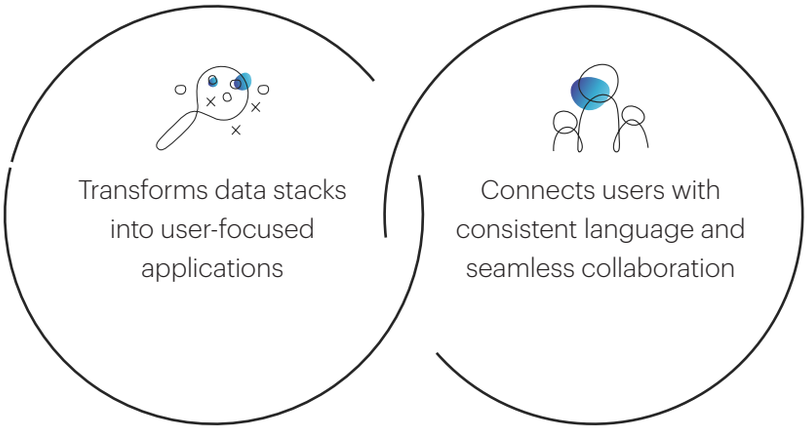
Metric layers have seen a surge in popularity recently, but the concept of embedding BI into other applications has been around for many years. As such, the metric layer has made a step toward improving the visibility and reusability of data, but it has not managed to solve the problem of not adding value.

An entirely new way of interacting with data is needed to make data work. This requires a paradigm shift that achieves two things:

1. Transforms data stacks into user-focused applications. By starting from the value stream and then going back to underlying technical processes, data can actually be linked to actionable insights to enable better decision-making.

2. Connects users with consistent language and seamless collaboration. No longer will working with data be frustrating. Teams will have a streamlined and guided activity to fully leverage data by asking the right question at the right time.

A paradigm shift to make data work - **support business processes and users in taking the best actions and decisions possible**



The primary goal of the data fabric will be to support business processes and users in taking the best actions and decisions possible.

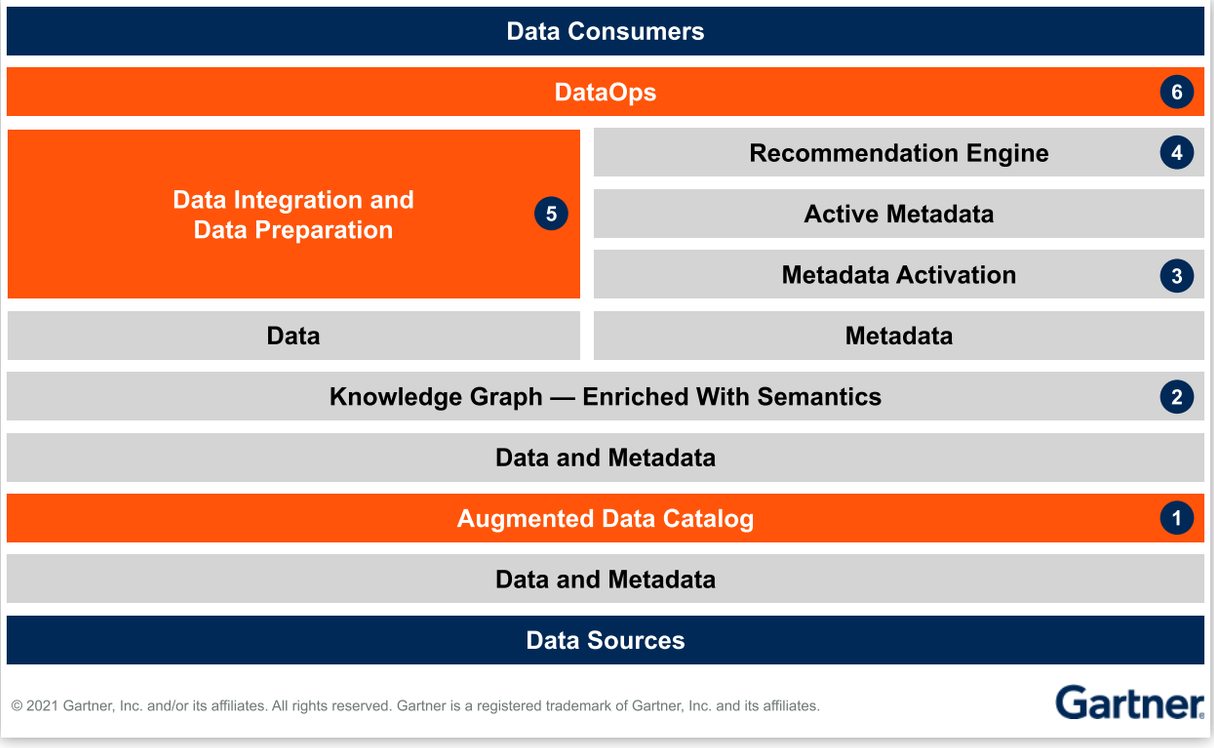
What Could Next-Gen Automated Data Fabric Look Like?

A data fabric is defined as a structured and organized representation of data that is used to describe the meaning and relationships between various data elements.

In their Top Trends in Data and Analytics, 2022, Gartner predicts that by 2025, active metadata-assisted automated functions in the data fabric will reduce human effort by a third while improving data utilization fourfold.

Yet, they refer to data fabric as a process with "composable" and "interoperable" architecture. This enables gradually adding technology components that are not necessarily delivered together immediately as a platform.

Data Fabric Need to Be Designed - **With "Composable" & "Interoperable" Technology parts**



Source: The "Practical" Data Fabric — How to Architect the Next-Generation Data Management Design, Ehtisham Zaidi, VP Analyst, Gartner

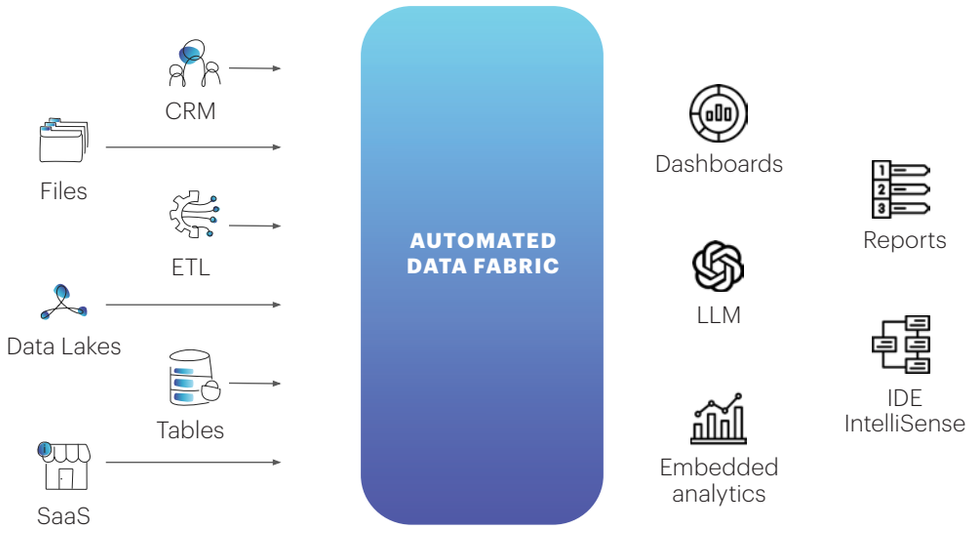
Some data fabric approaches focus on recommending design principles and architectures, such as those defined by IBM, that provide processes and implementations for a better data journey. Microsoft Fabric presents a mature approach to data fabric, aiming to eliminate data silos and enable easier unlocking of new insights.

The next-gen data fabric needs to step further and productionize the concept into a universal solution as a fully-automated independent layer. This layer should sit between data and business users that provide a consolidated and shared view of data (and their meaning) across the organization. It serves as a common language for different technologies and systems to communicate with each other and allows for more efficient and accurate data collection, analysis, and utilization.

How it works and functionality

The next-gen automated data fabric is powered by bi-directional metadata from across the organization. This means that the input data is coming from an organization's ETL, queries, usage patterns, and dashboards to name a few. This input data can be collected from any source and format including systems, devices, and people along with both structured and unstructured data.

Applications on top of the **Automated Data Fabric**



The next-gen automated data fabric requires advanced functionalities to achieve its purpose. These include:

- Ingesting metadata from across the organization (ETL, Queries, Dashboards, Usage patterns).
- Having data catalog capabilities so that the fabric can understand the physical structure of data in a warehouse or data lake.
- Having metric store and feature store capabilities, to the extent necessary to be exposed to the output of BI and AI tools.

- Having lineage capabilities to be able to connect the dots between data entities.
- Knowing every data set and interaction with data within the organization.
- Inferring and extracting context automatically by learning how the organization uses data.
- Connecting business questions to the correct data.

Since data is stored in tabular format and most data queries are in SQL, which is a structured language, an automated data fabric can use advanced data modeling and AI to automatically:

- Learn the mapping of schemas in the data warehouse to meaningful business entities.
- Learn the meaning/correlation from all queries asked to define an interconnected user behavior.

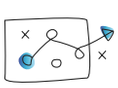
Let's take the same example of a team answering "What is churn?" When accessing the data fabric, the team will look for "churn" and see existing definitions with full lineage, timestamped reproducible usage, and suggestions for answers.

The envisioned next-gen automated data fabric can make data work by providing:

How the Next-Gen Automated Data Fabric Can Make Data Work?

Single source of truth:
Improved data accuracy and consistency

Precision at speed:
Faster and more efficient data collection and analysis



Elasticity:
Flexibility and scalability

Democratization:
Enhanced user experience and engagement

End-to-end impact:
Support for downstream use cases such as embedded analytics

Informed decision-making:
Data insights to power better business decisions and processes

By introducing this thesis on the Automated Data Fabric, we at Team8 aim to utilize its self-learned suggestion engine to enable efficient, fast, and smart data analysis to ultimately leverage data at its full potential. This solution will serve the diverse types of data consumers and make data insights accessible to all—from novices to experts.

Want to join us in shaping the future of the Automated Data Fabric?

Contact us: datavillage@team8.vc