The Evolution of Data 3.0

Modern

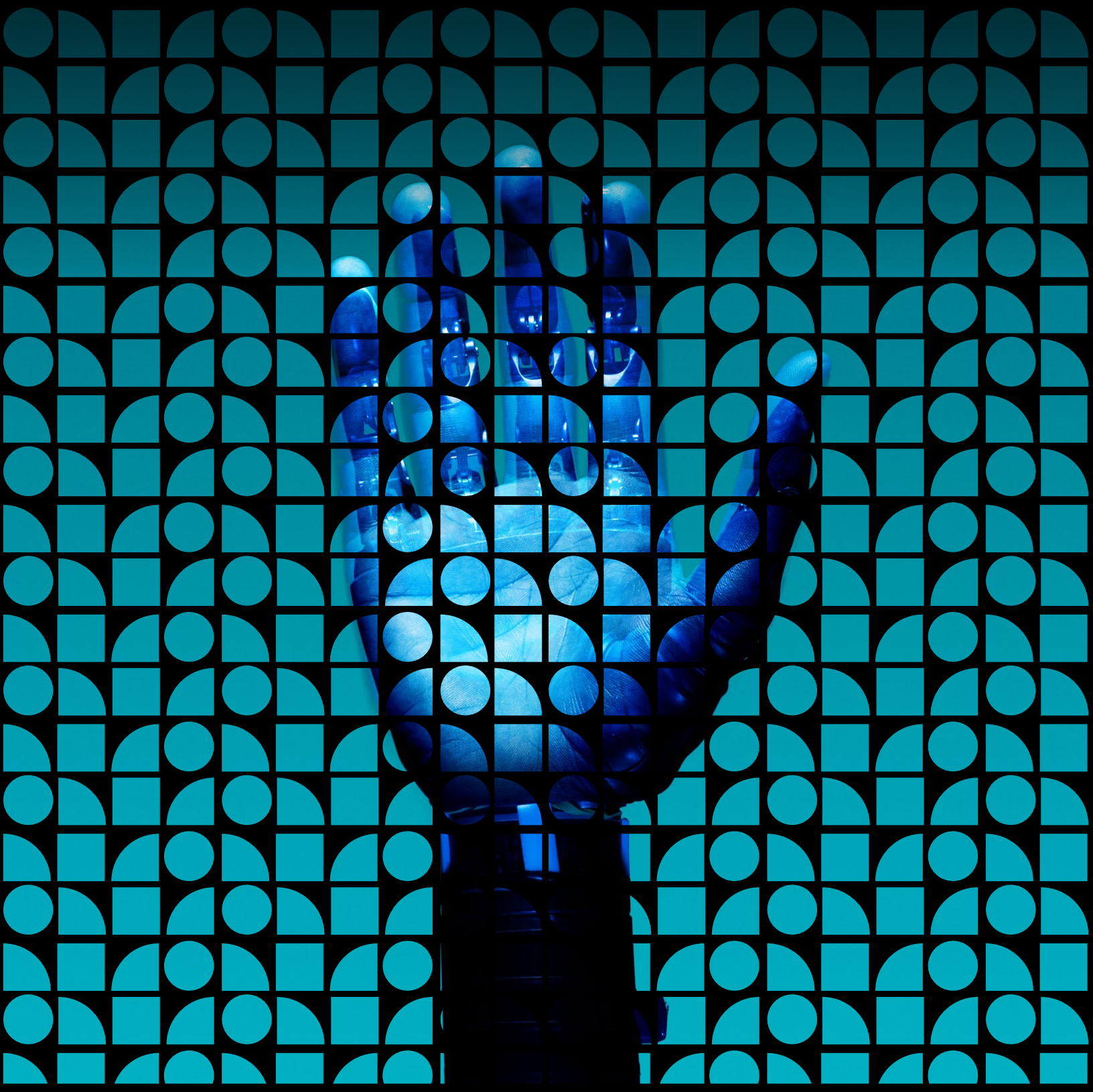


Table of Contents

TheModern DataCompany

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.IO

Introduction

Today, massive amounts of data are collected, processed, and stored for a range of analytical purposes around the world. Every customer, device, transaction, email, and image leaves a data trail. At present, this data is growing too big, changing too fast, and becoming hyper distributed. The traditional ways of doing integration and analytics are no longer viable or scalable. It is not feasible to create millions of data pipelines and to continue moving large amounts of raw data to a data lake or a centralized data warehouse.

Over the past decade, traditional data management and data analytics have undergone major transformations. In this white paper, we will walk through the evolution of data and data management techniques, Data 1.0 to today's Data 2.0—the transition of ETL (extract, transform, load) based data warehouses to ELT (extract, load, transform) based cloud native data platforms. We will also explain why organizations need to look toward Data 3.0 to address data needs that are beyond analytics.

Understanding the Technicalities

Both the ETL and ELT data transformation/integration process involves the following three steps:

- **Extract**

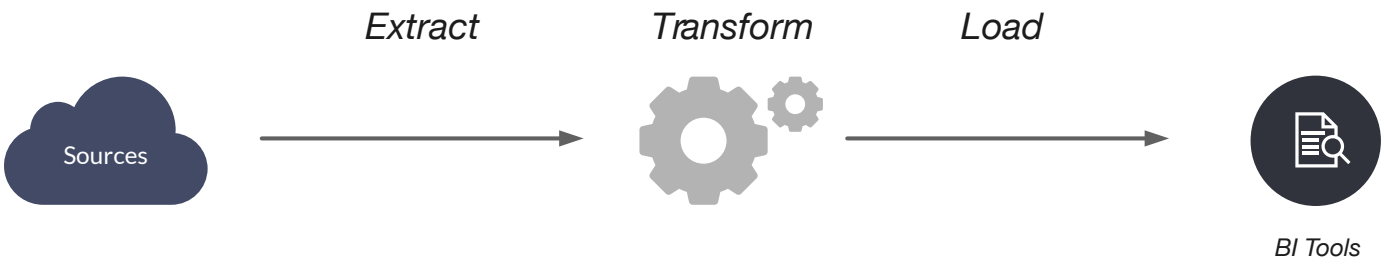
  Extraction refers to the retrieval of data from any data source system.

- **Transform**

  Transformation refers to the process of structuring and enriching the raw data for integration with the target system.

- **Load**

  Loading refers to the process of laying the data into a storage system for analysis by various business intelligence (BI) tools.



Evolution of Data

Data 1.0

In this phase, data sources and volumes were relatively small. Data was structured as customer data, sales data, financial records, and often came from internal sources. Everything was stored in enterprise warehouses, data stores, or marts before analysis. Analytical models were created using a "batch" process over the course of several months, and a great majority of analytical activity was either descriptive or hindsight analytics. Data access and usability were limited to a select few and IT, while business teams spent the majority of their time preparing data for analysis and relatively less time on the analytics itself.

This phase saw the rise of data warehouses which contained enterprise data in one huge repository. In this architecture, data is located below the applications layer. If a user wanted to read data, they would need to go through the application to access specific data. A single unified view of all the datasets across the organization was just not available.

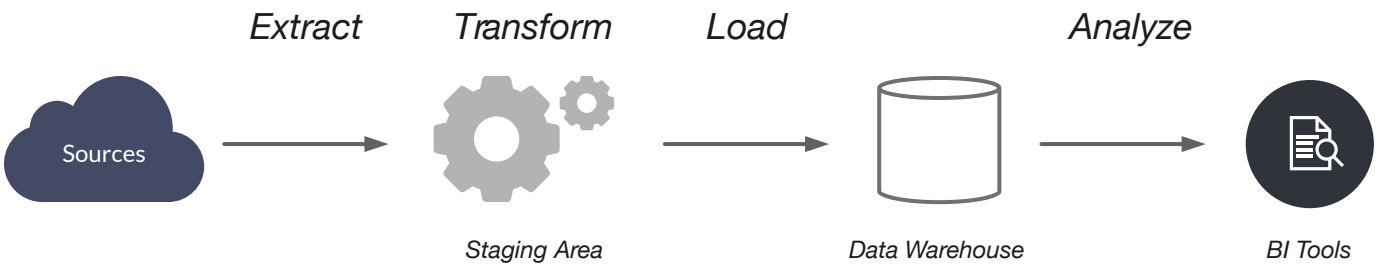
During Data 1.0, big data platforms relied on an ETL-based integration approach.

1. Data is extracted from various homogeneous or heterogeneous data sources.

2. It is then deposited into a staging area, where it goes through a cleansing process, gets structured, enriched, and transformed.

3. It is finally stored in a data warehouse where various BI tools can use it for analysis.

ETL-based data integration facilitates OLAP (Online Analytical Processing) analysis at scale to power hindsight analysis effectively. In other words, this type of data integration is suited for creating and viewing reports only after the events have happened or developed.



Challenges Posed by the ETL Data Integration Approach

ETL as a data integration approach lasted for almost two decades. However:

- Traditional ETL tools require disk-based data loading and transformation, which limited data growth to database technologies that were available—at a time when organizations saw their data grow at an unprecedented rate.
- By using ETL, data citizens and BI users experienced lengthy wait times because querying data and its analysis was a slow and inefficient process.
- Real-time access to data was not available until the entire ETL process had been accomplished.
- ETL data pipelines demanded high maintenance and issues with broken data pipelines were constantly reported.

The ETL data integration approach may have been straightforward and simple but it involved many input and output activities, which ultimately meant it was too compute-intensive and costly. Every time data teams encountered a new challenge, a lot of data had to be parsed and that means a lot of time was not utilized as it should have been.

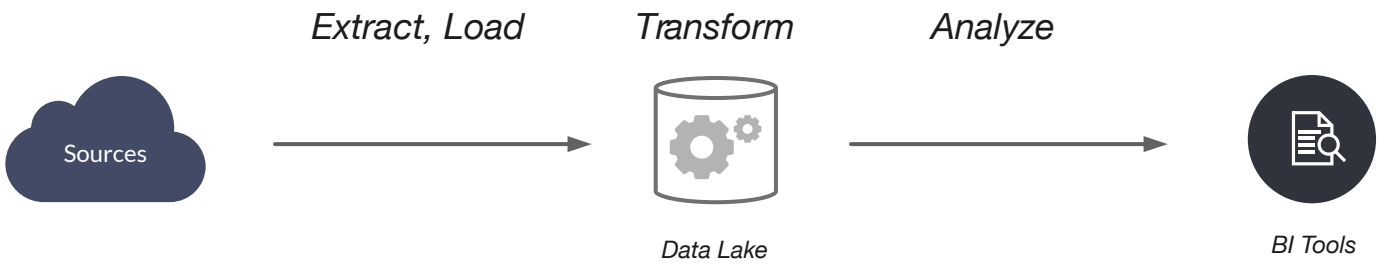
Data 2.0

Today, we are experiencing Data 2.0, a point in which organizations have started collecting data from multiple external sources. The volume and complexity of data have also grown exponentially. However, the advent of new cloud-based data warehouses, data storage, and compute capabilities has enabled organizations to scale these solutions at will. Hadoop, an open source software framework for fast batch data processing, took over the analysis of big data that couldn't be analyzed fast enough with previous tools.

During this phase, cloud-native data platforms relied on an ELT-based data integration approach which enabled, to some extent, a single view of the data across an organization. This allowed organizations to build data analytics as a system and made data analysis more streamlined.

During Data 2.0, cloud-native data platforms relied on an ETL-based integration approach.

1. In ELT, data extracted from various data sources immediately goes into a data lake storage system.
2. Data Lakes are special data stores that accept any kind of data (structured, unstructured or semi-structured), unlike OLAP data warehouses. This means the application of business logic driven transformations was no longer necessary before it is loaded into the data lake
3. Data is transformed on an "as-needed" basis for analytical purposes.



Major Advantages of the ELT-based Cloud-Native Data Platforms

High data availability - ELT can load all the data once into a single system. This facilitates quick deployment of powerful relational models to manipulate information and empowers data citizens to determine which data to transform and analyze.

Decoupled storage and compute - ELT processes are cloud-based and hence data sizes do not matter. Organizations can independently scale storage and compute on an as-needed basis.

Single view of the data - ELT based integration process works hand-in-hand with data lakes which can support all kinds of data. This enables the ability to load all data into the system at once and the creation of a single view of data across multiple systems, organization wide.

Limitations of Data 2.0

Despite the major evolution of the entire data integration process in Data 2.0, control of that data remains with the vendors—in addition to several other limitations.

- ELT-based cloud-native data platforms can solve the data analytics needs for many organizations, but they have yet to address operational data needs. For example:

  Using data analytics, an e-commerce company can learn the details of users who did not finish the purchase and abandoned their cart. They can also understand the cart abandonment rate and incentivize shoppers with a discount coupon to finish the purchase. However, these insights will not enable the e-tailer to reduce the number of cart abandoners in the first place. The operational needs of the organization cannot be addressed in real-time.

  The discount is a catch-all that may not be compelling for every shopper. To personalize the retargeting incentive, the e-commerce company needs more data, such as order history or wishlist creation. The improvement of cart abandonment rates will lead to higher sales, less churn, and better customer engagement. This is the impact operational analytics can have. It helps organizations make strategic decisions that are backed by real-time data.

- Data governance is a big concern for organizations in Data 2.0. ELT requires data to be loaded before redactions or masking of sensitive information can happen. Organizations must be hyper vigilant when it comes to compliance with privacy regulations and rules like HIPAA, PCI, and GDPR.

- Data pipelines must be manually configured and built for every data source and broken pipelines have to be re-engineered and fixed every time. This creates delays and hinders near real-time or predictive data analytics.

- To tackle unstructured data, organizations turned to a new class of databases known as NoSQL. This requires additional costs and resources with niche skills.

In Data 2.0, data is still not the connective tissue of the organization.

Data 3.0

Previous eras of data management created a single source of data across multiple systems organization-wide. However, data management is not only about the removal of data prisons, it is also about the activation and operationalization of data. To achieve all three, organizations must re-examine their data engineering, data governance, and data observability.

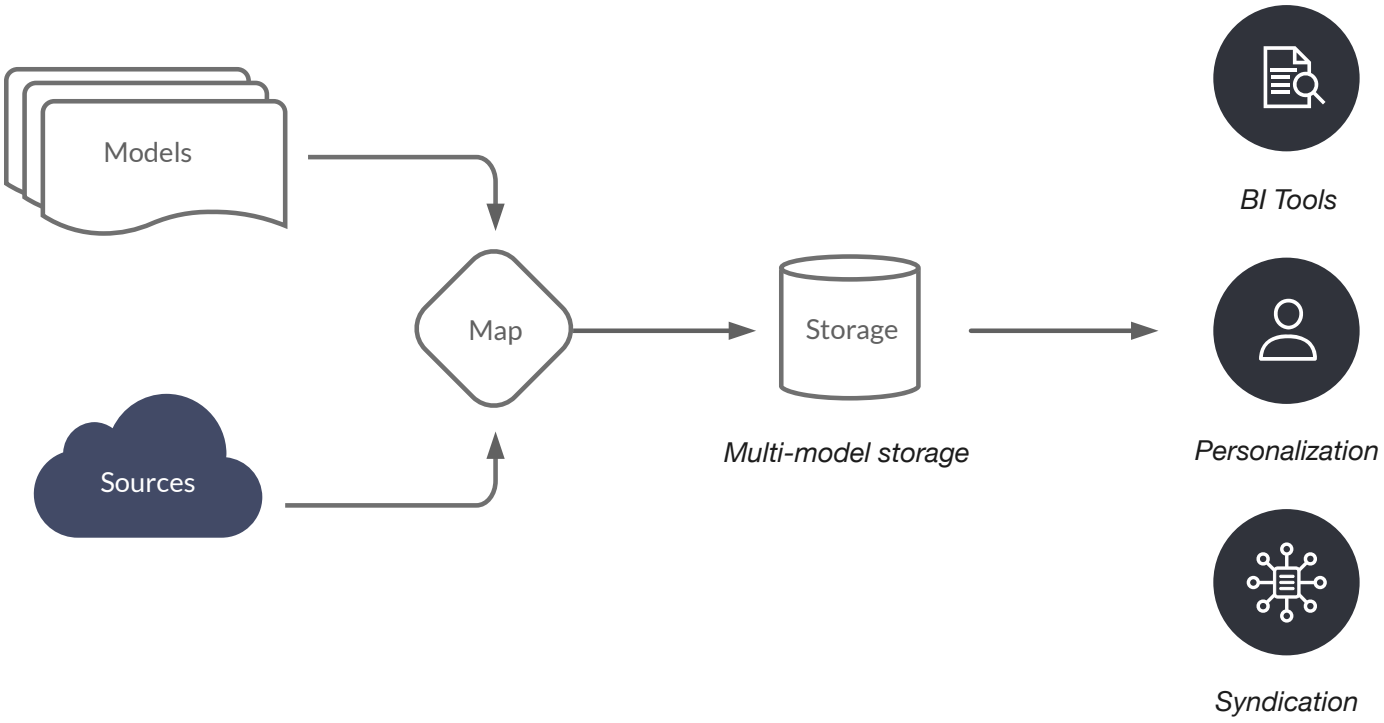
Automate the data engineering processes

Having additional pipelines built manually for every data request is simply not scalable. From the start, at data ingestion to transformation, all data engineering processes should be automated using ML or AI approaches. Today, they are done using ETL or ELT and it remains a point-solution approach. Organizations need to begin moving towards an MML approach where there is modeling, mapping, and loading of data. The more automated the data engineering is, the easier it is to scale data operations.

Do data governance the right way

Operationalizing data requires thinking beyond data analytics. It first requires a modern data governance engine that can handle data operationalization and scaling. The current role-based access control (RBAC) behind many data governance products is too rigid, not allowing for the creation of a data environment where data access and usage has to be closely monitored. The current data governance tools are outdated.

Organizations need attribute-based access controls (ABAC), not role based access controls. They need to treat data security and governance compliance at the atomic level. Infosec teams should be able to automate and deploy these policies across data pipelines without needing to reconfigure or re-engineer the pipelines.



Constantly observe the data

Many organizations do not know what happens to their data inside the black box data products that they rely on. With thousands of data pipelines that need to be monitored, it is critical that an organization's data products provide extensive observability by default. The products must emit all data logs all the time to enable internal teams to consume alerts and act upon them. For example:

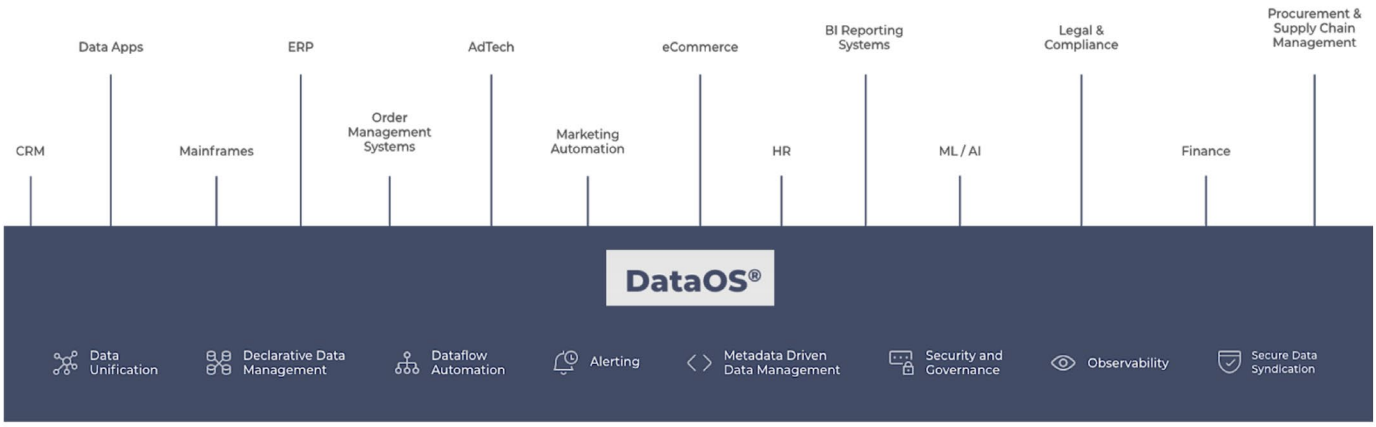
If sensitive data (e.g, PII or PCI) is present, infosec teams should have rules and workflows in place that require a manual approval process of the data classification before anybody can use the data.

Through our experience, we have observed that organizations spend a lot of time addressing data engineering, data governance, and data observability issues rather than focusing on business outcomes. Organizations should empower their business teams with the best data tools capable of harnessing the true value of their data.

DataOS - The Next-Gen Data Management Platform

DataOS enables the ingestion of any data (e.g., batch, stream, structured, or unstructured) at any volume and cadence. These jobs can be defined using DataOS declarative YAMLs and deployed using DataOS Cluster APIs. This empowers users to create a single source of truth, reduces complexity, and automates data deployment pipelines.

DataOS provides data observability out-of-the-box. Dedicated log datasets contain the details of who, what, where, and when for every activity or event that occurs in the DataOS environment. These log data sets can be consumed and analyzed as needed for all auditing purposes. DataOS also provides a flexible, modern data governance engine for role-based and tag-based access controls. This empowers teams to set up conditional access controls like the ability to access a data set on a certain network, or during a certain period of time. Policy deployment across the entire data pipeline can be automated without frequent reconfiguration and re-engineering. DataOS is a single product that activates and operationalizes data in real-time and gets data engineering, data governance, and data observability right.



Get Started

In Data 3.0, organizations must make data the connective tissue between different domains, entities, and systems within an organization. DataOS, the modern data fabric, truly enables this connectivity.

Learn more about DataOS and how it can work with your current architecture to take your data insights to new heights. Email us at hi@tmdc.io or visit us at themoderndatacompany.com →

About DataOS®

DataOS is an operating system that consists of a set of primitives, services and modules that are interoperable and composable. These building blocks enable organizations to compose various data architectures and dramatically reduce integrations. Enterprises can have the same data-driven decision-making experience akin to data-first tech companies in days and weeks instead of months and years.

About The Modern Data Company

Founded in 2018, The Modern Data Company began with the realization that enterprise-wide data access has been siloed. Data engineers and database administrators have been the longstanding data gatekeepers who funneled data to analysts and data scientists. We aim to change that by freeing enterprises to make better data driven decisions by democratizing access to data. When all employees, irrespective of their technical skills or background, can easily explore and analyze enterprise data, then both productivity and market expansion are realized at a faster pace.

TheModern DataCompany

The Evolution of Data 3.0

The Modern Data Company
306 Cambridge Ave
Palo Alto, CA 94306
TheModernDataCompany.com
info@TMDC.IO