Building Blocks of a Modern Data Platform

Authors: Milind Chitgupakar, Mark Ramsey & Baz Khuti

A guide to developing a modern data platform strategy to aid in the acceleration of digital transformation efforts.

modak

Why do we need a Modern Data Platform?

Tomorrow's business and technology leaders are in search of a blueprint that addresses the key data challenges, cultural mindsets, and architectural principles required to remain competitive in a digital-first economy. A future where a "datafied" organization systematically handles data as a business discipline and as a strategic asset, would thrive in their markets, and accelerate continuous innovation. The panacea for Chief Data and Analytics Officers (CDAOs)!

A Modern Data Platform (MDP) from a technology perspective enables organizations to automate the ingestion, curation, and consumption of disparate data sources (internal, external, structured, semi-structured, and unstructured) that continuously feed knowledge and insights. This contrasts with traditional data platforms where integration is primarily focused on very narrow data sets, structured data, to provide information.

Consequently, we are at a technological and organizational paradigm shift from traditional data management tools that feed data warehouses and data marts with pre-defined data schemas and tables to provide business users with information on events that have already occurred. For example, "How have customer revenues changed in the last year compared to the previous year?" To help users create a report, the underlying data structure has been defined by IT to address the specific question, with very narrow data sets and this approach is repeated on a use case by use case basis. IT, as the traditional custodians of applications and systems, owns the data and provides business users with standardized, centralized, and governed data platforms.

Research firm Gartner estimates that:

"Through 2025, 80% of organizations seeking to scale digital business will fail because they do not take a modern approach to data and analytics governance"

Gartner: Predicts 2021: Data and Analytics Strategies to transform Digital Business.

Regulation and compliance in financial services, life sciences, insurance, and other industries reinforce the need for strong centralized data and analytics governance. Over time governance has evolved into semi-federated hub-spoke data management model, with the business managing analytical development and IT providing the data management and infrastructure. Traditional data platforms have been built on-premises, engulfed by thousands of batch ETL jobs, have created a spaghetti data interconnections, and where the "T" (transformation) is tightly coupled with the underlying compute and storage infrastructure.

modak

Research firm Gartner estimates that:
"Through 2026, 90% of data management tools and platforms that fail to support multi-cloud and hybrid capabilities will be set for decommissioning within three years."

Gartner: Predicts 2022: Data Management Solutions Embrace Automation and Unification

What is a Modern Data Platform (MDP)?

A MDP is a fundamentally different approach, architecture and culture to the way data is and has been managed in the past. Modern Data Platforms provide an automated data infrastructure that continuously fuel algorithms that learn and evolve as more data is fed into them. As AI requires more and more data, there is an explosion in the breadth, depth and velocity of data needed; thereby enabling business users to explore a portfolio of use cases simultaneously. This requires interoperability across curated data assets and the creation of data domain assets.

Within a MDP framework the aim is to harness all the full data landscape, internal and external to the organization irrespective of the data format. This convergence of data, consolidated in hybrid or multi-cloud data lakes allows business users to explore, experiment and create data domain assets for a variety of use cases, purposes, and outcomes.

What are the key principles of a Modern Data Platform?

The principles listed below are based on learnings and experiences from some of the largest life science customers in the world operating modern data platforms in production today.

- Full data landscape requires inventorying all data sources and not being selective to solve specific and targeted use cases. A full scan of the data sources requires automated crawling and profiling of data, populating an active metadata repository where data relationships can be defined, visualized, and searched upon.

  Effectively a "**Data Fabric**" is created that acts as a net cast over the data sources to stitch together multiple heterogeneous data sources and types, through automated data pipelines that proliferate an active metadata repository.

- Consolidation of data into cloud enabled infrastructure to provide a "**Data Lake**" – enabling flexibility of deployment on multi-cloud and hybrid cloud infrastructure.

- The application of advanced AI and ML driven techniques to automate the standardization and harmonization of data sets into data domain assets allows the fusion of operational, external, and analytical data sets to create business owned data products with their own lifecycle for consumption through APIs.
  This data democratization can be seen as a "**Data Mesh**" providing federated data governance and access to data assets through self-served data domains/self-service data capabilities.

- Do not limit the approach to a single use case, instead enable a portfolio of use cases to be created and delivered, prioritized to business needs and outcomes.

modak

What are the cultural values needed for success?

As the adage states "Culture Breeds Success". Why do certain teams and firms succeed while others, who are as capable and hardworking struggle to generate outstanding performance? Based on our experiences and work to deliver large-scale MDP initiatives, we see a common pattern of values and behaviors emerging:

- Organizations that truly adopt and embrace data as a strategic asset, where data is not owned by any function but accessible and used across business areas for decision making.
- Leadership by example – senior executives with deep domain knowledge owning and taking accountability for the data value chain, providing subject matter expertise, mentorship and transparency in decision making and outcomes.
- Openminded to try new approaches and stop accepting traditional ways and doctrine of how things have been done and operated in the past.
- Pace – a sense of urgency to move at speed, to learn, adapt and fail fast adopting start-up founders' mentality to focus and deliver value.
- Understanding the **why**? Ensuring team members, stakeholders and external partners understand the business outcomes and value created. The "dots" connecting technology, business and customer values are shown and communicated.

> "It's not about the technology, it's about what's between the ears of individuals and their willingness to change and use the technology".
>
> Dave Clementz, CIO, Chevron Texaco 2004.

Putting all the pieces together - Customer Case Study

The Life Sciences industry is a high-risk venture, with the average time for drug approval being 10 years, approximately $2 billion in investment with a 10% success rate. Within this background, a Top 10 Life Sciences company embarked on an ambitious vision to create a next-generation modern data platform that enables data convergence across their R&D organization. To allow researchers and data scientists who struggled to answer key questions, such as:

- How to accelerate clinical trial research yet reduce the risk of failure?
- How to increase the pipeline for new medicines and reuse data from previous clinical trials?
- And how to share research data, build communities of expertise and increase the productivity of scientists spread across hundreds of labs?

To answer these questions, researchers were challenged to find data across thousands of data silos, having to encode data language and nomenclature encapsulated into proprietary databases that were far removed from the business context.

Solution and Outcomes

The Life Science company identified a powerful combination of Cloud infrastructure, Cloudera CDP, a hybrid, multi-cloud data platform, Modak Data Engineering Studios and Modak Nabu™, a modern data engineering platform.

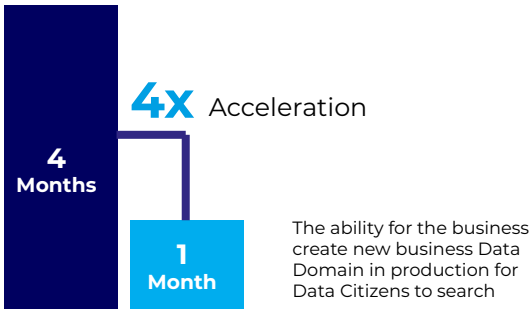
Cloudera's CDP software provided the foundations for the R&D data lake, hosted on AWS Cloud, integrated with Modak's Nabu™ software to automate ingestion pipelines, profiling to feed an active metadata repository and curation services to standardize, normalize and harmonize data domain products for consumption. Empowering thousands of users to visualize data relationships and spend less time finding answers to research questions such as drug adverse events, drug-to-drug interactions, and drug-to-genome associations.

The time to value was measured in the 12 weeks to profile 200+ data sources and ingest 76TB data and the ability for the R&D team to design, develop and deploy 70+ use cases, 8 new data products and a 4x reduction in time to create new data products.

Solution implementation time.

Reduced time to implement enterprise scale Data Fabric and Data Lake

Speed to build new data domain products.



The ability for the business to create new business Data Domain in production for Data Citizens to search

About Modak

Modak is a solutions company that enables enterprises to manage and utilize their data landscape effectively. We provide technology, cloud, and vendor-agnostic software and services to accelerate data migration initiatives. Using machine learning (ML) techniques to transform how structured and unstructured data is prepared, consumed, and shared.

Modak's portfolio of Data Engineering and DataOps studios provides best-in-class delivery services, managed data operations, enterprise data lake, data mesh, augmented data preparation, data quality, and governed data lake solutions.

- **Find out more**
  https://modak.com/contact/
- **Follow us**
  https://www.linkedin.com/company/modak/
  https://www.facebook.com/ModakData/

Modak Nabu™

Modak Nabu™ enables enterprises to automate data ingestion, curation, and consumption processes at a petabyte-scale. Modak Nabu™ empowers tomorrow's smart enterprises to create repeatable and scalable business data domain products that improve the efficiency and effectiveness of business users, data scientists, and BI analysts in finding the appropriate data, at the right time, and in the right context.

modak