modak

Delivering Modern Data Solutions Faster

Author: Sanjeev Mohan

Gone are the days when a buyer would place an order and wait for hours for the goods to be delivered. Or for a letter to arrive by mail. Why then should a business user not expect results from their data at the speed of their decision making? If an organization is launching a new campaign, it is imperative that the data outcomes not only be available when needed but also be trustworthy.

So much emphasis has been put on technology itself that data professionals have lost sight of the original goal - to meet business needs. Most so-called modern data stack conversations start with how a comprehensive architecture comprising a plethora of products will give business what they need. This technology-first approach has led to suboptimal solutions are expensive and take a long time to build. This approach isn't sustainable.

As we pivot to a decentralized approach for developing data outcomes, more responsibility shifts to the business domains that understand their data best. This approach also removes bottlenecks for central IT teams and increases accountability.

Changing our mind-set to become business-outcome first means that we need to understand what the business needs. At a bare minimum, we need the ability to make business decisions at the right time. For this to happen, minimum standards must be maintained, which leads to the following expectations:

- **High quality and accurate data** that can be trusted by the business users.
- **Personalized user experience** with self-service access to data.
- **Complete reliability** from their "always-on" data subsystems infrastructure.
- **Data privacy and security** policies are maintained to meet regulatory compliance requirements.
- **High performance** on analysis of data to meet current and future use cases.
- **Cost** estimates are met and there is a transparency into the value created.

How hard can this be? Is this now a pipe(line) dream or a bridge too far?

In the current approach, IT teams address business needs by cobbling together a complex architecture comprising multiple software products. Next, they integrate these products by building the pipelines that connect various upstream data producers with the increasing number of downstream data consumers who are hungry to access more and more data. Often, data journeys from data sources to data processing, and consumption tools are spread across on-premises and multiple clouds, which further exacerbates complexities.

This IT-centric approach is causing increasing frustrations from the business users who are now leading an effort to modernize their data infrastructure. While IT professionals endlessly debate the pros and cons of bundled vs. unbundled approaches, the business teams question the value, time, cost, and effort taken to deliver their needs. Yet, the IT landscape has been fragmenting at an alarming rate and there are multiple narrow categories of capabilities, each with their own vast number of products and approaches. Confusion persists on which technological approach to pursue and which ones will future-proof their investments.

Organizations lack proper guidance and direction on how to modernize. They are wary of investing in technologies that will be obsolete and lead to an expensive and risky overhaul to meet expanding needs.

While businesses face many challenges, recent developments have started addressing them:

- **Time to value.** Building data pipelines requires a heavy integration overhead as the products involved in the process lack industry standards. The complexity and cost goes up further as many new SaaS data sources themselves mature. For example, they may change schema or features, requiring expensive ongoing maintenance or organizations end-to-end data pipelines. As a result, there is a move to adopt cohesive platforms that pre-integrate some of the basic building blocks.

- **Reliability.** Pipelines involving disparate products lead to a lack of transparency into the health of data when it moves from data sources to targets. This leads to brittle pipelines and a lack of accountability, which is the reason why the data observability category has seen an explosion of product offerings. Data observability introduces proactive monitoring and alerting of anomalies.

- **Quality.** Organizations have built silos of data to overcome the inefficiencies of their data infrastructure, thereby perpetuating poor data quality. The past approach of manually fixing data quality in a downstream system is no longer viable, which is the reason why data mesh and data products are increasingly being deployed. They promote domain ownership by shifting the responsibility for development to the business teams, thereby eliminating bottlenecks of centralized data engineering teams.

- **Skills.** Modern data infrastructures require an army of specialists, when the focus should be on business outcomes. The goal should be to balance automation of non-value-add and repeatable tasks while using human-in-the-loop to maintain context. In addition, new skills are needed, such as product management within the data teams.

If these challenges are not addressed, it leads to reactive data teams, poor developer experiences, and introduces unnecessary risks and costs to organizations. Surely, a better and proactive approach exists.

Best-of-breed vs. Integrated

The pendulum between centralized (bundled or integrated) and decentralized (unbundled or decoupled) constantly swings. An integrated, or a centralized approach has been in use for the last few years but may lead to IT bottlenecks. On the other hand, the best-of-breed option consisting of specialized products may provide the depth of functionality, but at the cost of higher integration overhead. This age-old dilemma is the well-known buy (integrated) vs. build (best-of-breed) debate.

The best-of-breed approach can lead to an unmanageable set of products that don't share metadata and hence require time-consuming integration. This slows down the velocity of delivering data outcomes. In addition, debugging errors also become time-consuming. In theory, point solutions enable upgrades with new, more capable options, but this is predicated upon using products that follow open standards. This is easier said than done. For example, SQL is an ANSI standard, but no two databases implement it exactly the same way. Hence a time--consuming and costly migration effort may be required even when the technologies in use are the same.

Proponents of the integrated approach point out the overhead of managing point solutions even though it is more time consuming than a tool that provides unified business processes and data.

Key takeaway: There is no black-or-while right answer. Organizations' corporate standards and guidelines should determine the right approach. With all things being equal, when it comes to build vs. buy decision, unless the build adds to competitive advantage, your choice should favor the buy option. The buy option typically follows the integrated approach and can aid in faster development, implementation, and onboarding of users, but may also lead to a higher vendor lock-in.

The hybrid approach bundles core functionality, but provides an ability to connect to external tools for additional features, like monitoring, DataOps, etc.

Proprietary vs. Open Platform

Proprietary solutions provide peace of mind but at a cost. They abstract non business value-add activities and focus on providing a superior user experience. Often these solutions appeal to the teams that lack deep technical skills. Their focus is on addressing business outcomes. While they may introduce some lock-in, if they follow open standards then those risks can be minimized.

Open source products provide lower license costs and the users can benefit from the products' community. However, it may leave an enterprise open to unforeseen risks, which has led to the recent popularity of commercially-supported open source software.

Key takeaway: An organization's IT skills maturity should be one of the factors that determine the choice between open source and proprietary. Organizations with mature IT teams can invest in evaluating, integrating, and enhancing the open source products to meet their needs.

Control vs. Managed

Some organizations, especially the ones that are heavily regulated, desire control over their IT assets and usually have enough staff skilled in advanced technologies. In addition, their internal corporate security guidelines dictate where the data resides - in their own security perimeter or the software vendor's. Self-managed IT deployments fall under Infrastructure as a Service (IaaS) or Platform as a Service (PaaS).
Medium to small-companies may prefer managed services, such as PaaS, SaaS or serverless options.

Key takeaway: Modern architectures have too many moving parts to effectively operate and debug faults. The first right of refusal should be towards managed services.

No-low-code vs. Programmatic

A data scientist typically needs access to as much raw data as possible to write code in an IDE using Python, Spark, etc., while a data analyst needs access to curated data and may write SQL statements. On the extreme end of the spectrum, a data executive or even a citizen data scientist may need access to a semantic layer using a no/low-code tool.

The trend in recent years has been towards tools that reduce the coding effort. Even a data scientist may start by using Pytorch and then graduate to AutoML. A visual UI that helps users drag and drop widgets into a canvas and build the pipelines can cater to citizen roles that are not proficient in coding.

Key takeaway: Employ a hybrid approach that supports the varying needs of different personas - from programmatic to low-code to no-code.

This section provides direction for a decision matrix to evaluate architectures that suit the specific needs for an organization. It explores when an integrated approach is preferred over the best-of-breed approach. However, as the key takeaways suggest, a hybrid approach that combines the best of the two options should be the preferred option. The next section explores the building blocks of such an approach.

Introducing Intelligent Data Architecture Platform (IDAP)

The IDAP provides the building blocks of a next generation platform that unifies data and metadata to enable faster development of data products. It centralizes data infrastructure and metadata discovery, while accelerating decentralized development. Metadata use cases, like data quality and observability, are first-class citizens and not an afterthought. This hybrid option is a business-led IT platform.

Intelligent Data Architecture Platform

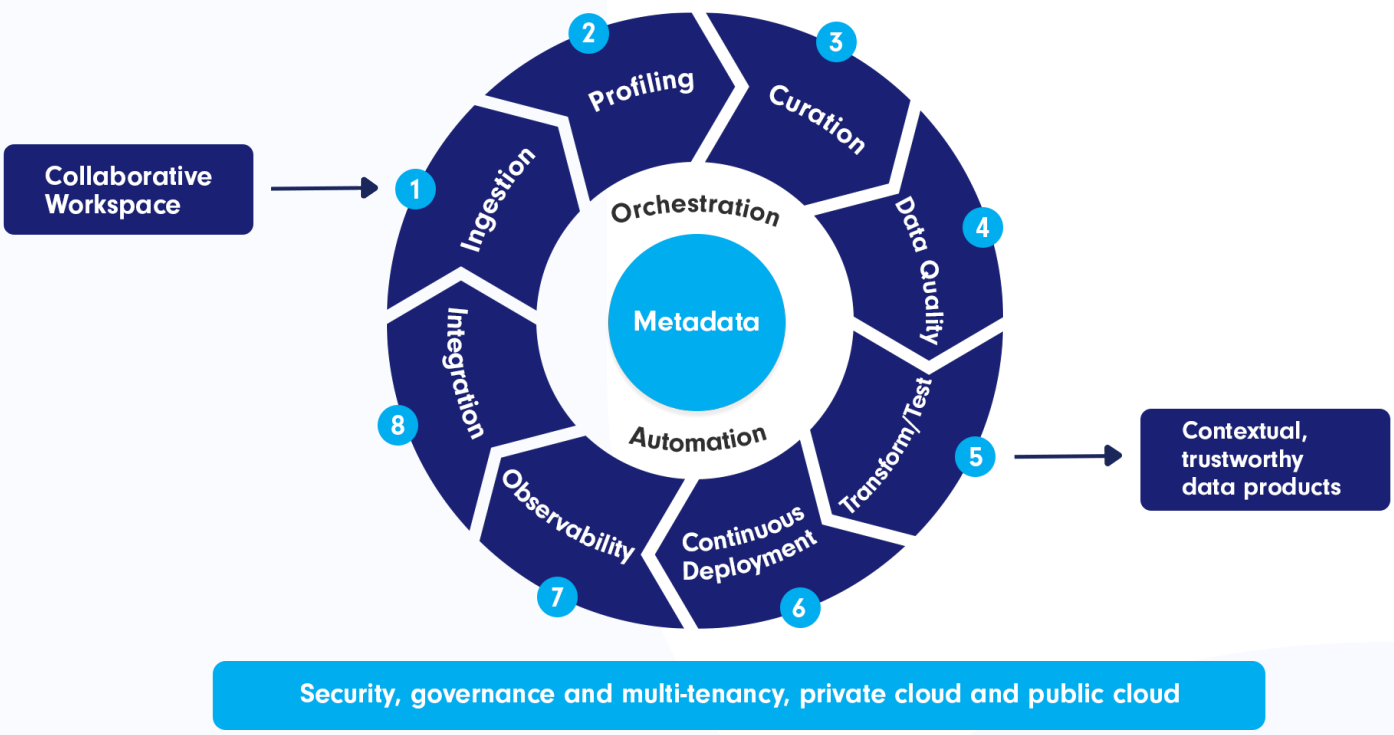

Figure 1. An integrated data architecture platform manages data product lifecycle which speeds up the process of building, managing, and operating data products.

This system is called an intelligent system because machine learning (ML) undergirds the metadata collection and discovery to perform the tasks. As a result, the metadata substrate powers the automaton and orchestration backplane. This unified backend engine enables domain owners to build and manage data products in a collaborative manner. More specifically, the IDAP, lifecycle begins and includes the following:

Ingestion

As the first step in the life cycle, data ingestion is the process of importing or receiving data from various sources into a target system or database for storage, processing, and analysis. It involves extracting data from source systems, transforming it into a usable format, and loading it into the target system. Data ingestion is a critical step in creating a reliable and efficient data pipeline.

Its overarching goal is to reduce multiple ingestion pipelines on the same data sources, as they can slow down operational systems, cause data sprawl, and lead to security risks.

This becomes even more critical as the number of data sources are increasing exponentially. Recent studies show that medium-sized enterprises, on average, leverage 110 SaaS products and large companies now have close to 500. This scale exacerbates data ingestion and leads to a spaghetti of scripts.

Some data is ingested in batch mode using data movement options like secure FTP, and some sources allow real time ingestion using pub/sub mechanisms like Apache Kafka or APIs. The IDAP needs to not only manage varying frequencies on when to ingest the data, but also discover its schema and handle changes, like schema drift.

The IDAP should provide automation and orchestration capabilities and the ability to use cheaper cloud resources to perform ingestion and minimize the overhead on the mission critical source systems. Using change data capture, the ingestion system should ingest only incremental data since the last ingestion run.

Profiling

Next, the ingested data from operational and transaction sources is loaded into a data warehouse or a data lake, where they are integrated and modeled for consumption by downstream systems and data consumers. However, before this data can be used intelligently, it needs to be profiled.

Traditional systems have provided mechanisms to profile ingested data and extract technical metadata, such as column statistics, schema information and basic data quality attributes, like completeness, uniqueness, missing values. This is technical metadata. IDAP, in addition, uses ML to build a knowledge graph, infer relations, and data quality rules. It helps generate operational metadata.

The generated metadata should be stored inside a native or a 3rd party metadata catalog. Many organizations already have data catalogs, hence, IDAP should provide an ability to integrate with the existing products. This makes the metadata discoverable and searchable.

Curation

Data curation involves the selection, organization, and maintenance of data to ensure its accuracy, reliability, and usefulness for analysis and decision-making. It involves activities such as data cleaning, transformation, and enrichment, as well as metadata creation and documentation. Effective data curation is essential to normalize, standardize, and harmonize datasets to deliver successful data-driven projects.

To speed up business-led data product development, the technical metadata comprising technical column names should be converted into business-friendly terms. This is business metadata. In this data curation step, the business metadata is linked to the technical metadata and added to the business glossary.

According to the data mesh principles, data products should adhere to certain principles, including addressability. The business metadata is used to build a semantic layer, so that there is a consistent definition across all the data consumers. The IDAP abstracts changes to the schema from the business by utilizing the semantic layer.

Transformation/Testing

One of IDAP's critical goals is to provide an excellent developer experience because if the developers are highly productive, then the business can achieve all its needs as described earlier. The IDAP does this through various means:

- A **collaborative workspace** is utilized to develop and deploy code. The IDAP borrows best practices from software engineering of agile and lean development, including reusability of the data transformation code.
- The workspace uses a **no/low code transformation engine** to speed up development. This engine may be built-in to the IDAP or integrated with an existing engine.
- **Continuous testing and automation** are other key components of the DevOps philosophy. IDAP applies these principles to data management. This fast maturing discipline is called DataOps .

Application of these best practices, improve reliability and trustworthiness of the data.

Data Quality

The quality of data plays an instrumental role in delivering high-quality data products to business teams. As data volumes continue to grow exponentially, ensuring data quality has become paramount and cannot be an afterthought for organizations.

IDAP provides embedded data quality checks into its data pipelines to address data inaccuracy, duplication, and inconsistency. By offering these data quality checks, IDAP delivers exceptional data products while enhancing the reliability of data for organizations.

Continuous Deployment

The IDAP has two distinct aspects - development and deployment. The deployment aspect also utilizes the DataOps best practices to push the code into production in a governed and secure manner. For example, it uses version control into a Git repository and CI/CD capabilities.

This allows the business users to accelerate experimentation by branching/testing new features without introducing breaking changes into the production pipelines. The new features can be rolled back quickly if needed. The IDAP introduces the much-needed A/B testing capabilities into the development of data products.

IDAP's orchestration engine enables the complete end-to-end pipeline.

Observability

In an integrated approach such as the IDAP, metadata is generated at every stage. This metadata can be used to develop a lineage and for faster time to detect and resolve errors. This capability is provided by the observability aspect of the IDAP.

The IDAP uses ML to detect anomalies and has an alerting and notification engine to escalate critical issues. Traditional systems were rule-based, which led to a large number of notifications that caused "alert fatigue. Modern observability systems are able to proactively determine anomalies to avoid downtime and to handle notifications intelligently to reduce the overload.

Integration

This section has already mentioned integration with data transformation, catalog and version control products. In addition, one of its biggest integrations is with the security infrastructure, including identity and access management. The IDAP uses standard enterprise security protocols.

An open platform helps in faster adoption of the integrated approach.

modak

Case Study

What good is an approach if it doesn't meet, and ideally exceed, business expectations. Here is a case study that looks at how one customer applied Modak's technology.

A leading company in life sciences leveraged Modak to develop an innovative knowledge platform that consolidated data from over **170+ sources** in order to meet their goal of enhancing productivity in R&D so they can develop better therapies for various diseases. The company's primary objective was to break down silos and unlock the potential of all stakeholders in the R&D process. The company created a "Repository of Knowledge" by building an enterprise wide knowledge platform that harmonizes and connects data from multiple sources. Using Modak Nabu™ and other partners, the platform automated data movement, operations, and managed ontologies, creating a mechanism to deploy computational resources on-demand.

This helped achieve key milestones, including reducing deployment time for computational infrastructure to mere minutes and search/access time for datasets to seconds. The platform's agile process for controls and approvals has facilitated broad reuse and eliminated delays in data ingestion for most datasets. Additionally, a clear mechanism for identifying data and preventing duplicated acquisition or storage at the enterprise level has led to significant cost savings. More importantly, the platform has enabled successful identification of new and repurposed drugs, making precision medicine a reality within their business.

The knowledge platform comprising 170+ data sources and over 500 TB of raw data was made immediately accessible to on-demand compute infrastructure. The knowledge graph has over 32M entities and 2 billion relationships. The platform is used by over 150 data scientists and thousands of researchers.

This platform led to a number of benefits, such as:

- **Saved time by delivering domain-driven data products**. The client saved over two weeks of work for every safety assessment.

- **Identified new opportunities via its Knowledge Graph**. Subgraph analysis identified 800k+ drug repurposing or rare disease opportunities. In one case, evidence connecting a rare disease with no known treatments to an FDA approved drug has been proposed as a possible therapeutic under consideration for off-label prescription for a patient.

- **Increased R&D Productivity.** Reduced deployment time for computational infrastructure to minutes, reduced search/access time for datasets to seconds

- **Established Data Governance.** Created an agile process for controls and approvals to enable broad reuse virtually eliminates delays to ingestion for the majority of datasets.

- **Streamlined Data Acquisition and Improved Data Quality.** Developed a clear mechanism to identify data and prevent duplicated acquisition or storage at the enterprise level has enabled significant savings.

Conclusion

The future belongs to organizations that are led by business-outcomes, rather than being technology-focused. These organizations are laser-focused on delivering business value at all times and they have an urgency to transform fast, quickly stand up analytics use cases, and continuously innovate. Oftens, this requires adopting a hybrid approach that integrates the best of centralized infrastructure with domain-driven data products development. This approach helps deliver results faster and aligns well with organizational culture and skills, creating solutions with more value to clients/customers.

Partners that provide an integrated platform save their customers time and money while also delivering trusted business outcomes. The time saving comes from avoiding integration of several technologies. No value proposition resonates more than the one that can measurably make their clients significantly more efficient.
Organizations can easily measure the benefits, such as the ratio of successful projects, deployed use cases, and the frequency of new releases. These factors are a reflection of higher trust in data. Organizations finally have a hope to achieve the aspirational goal of becoming data driven.

These products gain from economies of scale, and like an ML model gets better by retraining itself frequently, so do these cloud-native multi-tenant data frameworks.

About Modak

Modak is a solutions company that enables enterprises to manage and utilize their data landscape effectively. We provide technology, and cloud-agnostic software and services to accelerate data migration initiatives. We use Machine Learning (ML) techniques to transform how structured and unstructured data is prepared, consumed, and shared. Modak's portfolio of Data Engineering Studio provides best-in-class delivery services, managed data operations, enterprise data lake, data mesh, data fabric, 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://medium.com/@modak

USA
21660 W Field Parkway, Deer Park, IL – 60010

USA
312 S, 4th St, Suite 700, Louisville, KY 40202

UAE
Dubai Silicon Oasis (DSO), 341041

INDIA
The Platina,Gachibowli, Hyderabad, 500032

modak