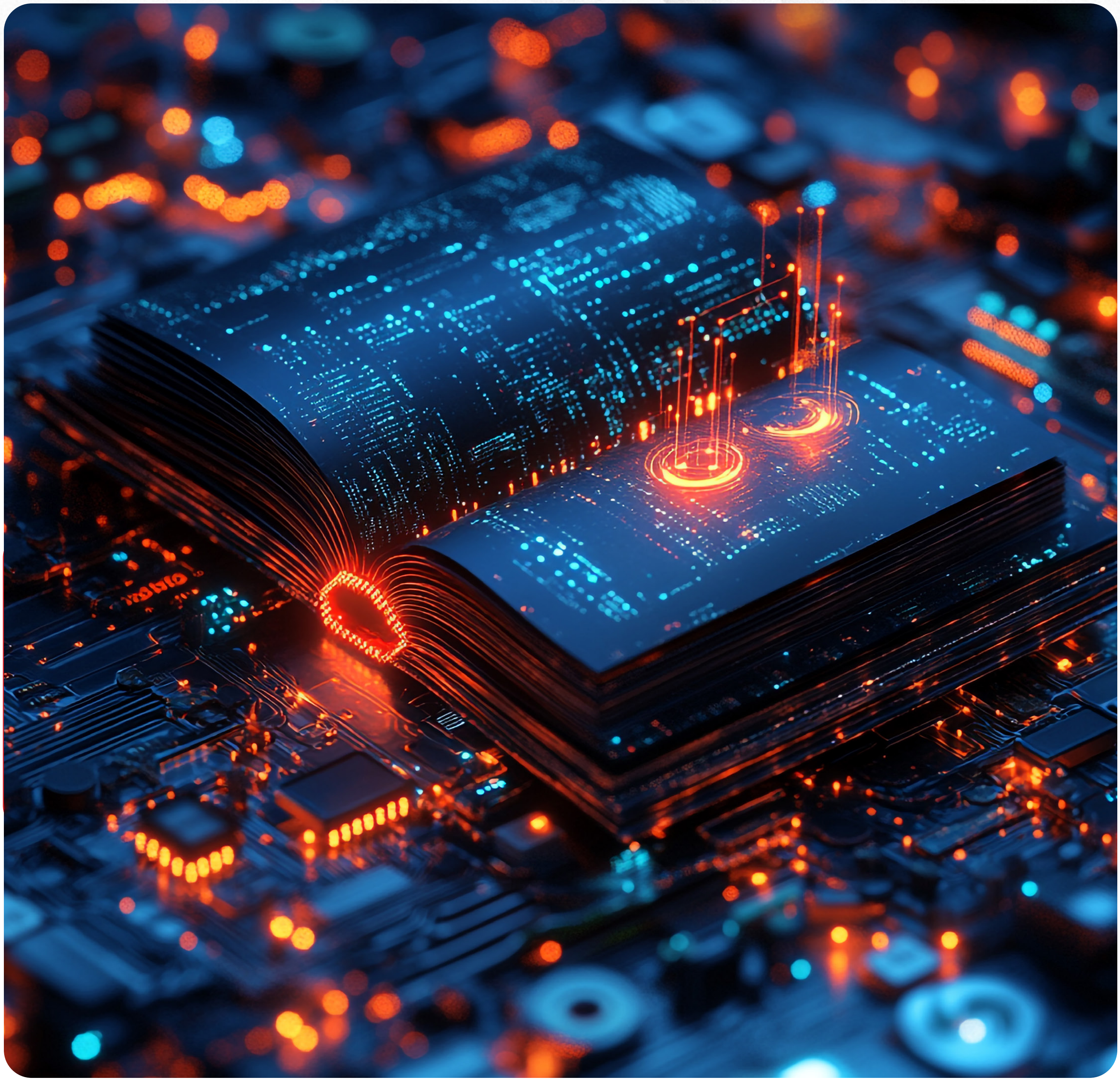
White Paper

Modern Data Platform Vision



In today's world, data is the key to decision-making for nearly every organization. The rise of digital transformation, cloud computing, and advanced analytics has brought forth a new era of data-driven decision-making.

As data continues to grow in volume, however, traditional data management systems are struggling to keep pace. Organizations are increasingly recognizing the need for a robust approach to data management that can simplify the collection, storage, processing, and analysis of large volumes of data.

A Modern Data Platform addresses these challenges by providing a unified, scalable, and flexible infrastructure designed to handle diverse data types and support advanced analytics capabilities. With the integration of technologies like cloud storage, big data processing frameworks, and artificial intelligence (AI), Modern Data Platforms can help organizations gain control of their data, unlock unprecedented insights, and drive innovation.

A Modern Data platform implementation is focused on outcomes. Successful results depend on both technology that is well-suited to the purpose and an approach that supports the intended workflows and mission needs. Therefore, organizations seek ing to improve their business outcomes with a decentralized multicloud platform should begin by asking themselves the following questions:

1. Who in our organization is missing the right data when they need it?
2. Where is this data already located, and why is it inaccessible? Is it:
   - Siloed in a cloud platform or data center?
   - Fragmented across many data stores?
   - Isolated by geography or unreliable infrastructure?
   - Locked in by a proprietary data format?
3. What is the most efficient way to continuously collect, store, and transform our data in real-time and at scale?
4. What tools can be used to connect data that cannot be moved or migrated?
5. Which new or existing processes can be augmented or replaced by AI to improve efficiency, reduce error, extract insights, and guarantee the reliability of our data streams?
6. What tools can we use to derive the most insight from our data given our organization's objectives and needs?

The answers to these questions should drive decision-making around vendor selection, architectural design patterns, and development and integration plans.

Modern Data Platform Design Considerations

Multicloud Capabilities

The increasing importance of performance and capability-to-cost ratios in today's world highlights the importance of having a modern data platform with the capability to extend into multiple clouds. This allows enterprises to choose between the services being offered among multiple providers at the lowest cost while also avoiding the pitfalls of potential vendor lock.

Secure, dedicated connections between cloud platforms enables previously siloed data to be connected and shared more easily. Organizations with data assets fragmented across cloud platforms should prioritize unifying their existing data through reliable, real-time data streams. The availability of dedicated, multicloud connections provides the advantage of secure, direct connections between cloud endpoints, rather than needing to route data streams through on-premises data centers or over the unsecured internet.

Furthermore, organizations looking to deploy new cloud platforms or expand their existing cloud footprint should strongly consider the availability of multicloud capabilities in selecting a cloud vendor or SaaS platform. The ability to seamlessly integrate a new platform with an existing data estate should be treated by all modern organizations as a design consideration of critical importance.

In either of these situations, the critical technical capability to connect and work with data across cloud platforms must be applied in combination with careful planning and deep understanding of the data and its uses, both current and planned. Notably, data semantics and formats are key; while a modern platform can make data connected and accessible, for applications and analytics to make use of it together requires compatible formats and data fields whose meanings are well understood. Without these factors, organizations can find themselves developing solutions that repeat the stovepipes that previously came from hosting data on disparate clouds. Additionally, security considerations may require revisiting in an integrated multicloud context, where access is not limited by access to a specific cloud environment.

Hybrid Deployments

Nearly as important as multicloud connectivity is a platform's support for hybrid deployments – that is, deployments split across cloud and on-premises data centers, sometimes across multiple cloud realms or geographic regions.

Many organizations choose (or are required) to keep certain data and applications within their private data centers for a wide variety of reasons, such as to comply with regulatory policies, to retain the best performance for their systems when cloud services cannot offer suitable advantages, or to leverage investments already made into existing on-premises technology.

Organizations with such needs should ensure that their cloud platform offers full support for hybrid deployments. A fully supported hybrid deployment means that an organization's cloud environment should function as a transparent extension of their on-premises network, rather than an isolated "data enclave."

There are many types of hybrid deployments and each organization's needs may vary considerably. Some broad capabilities to prioritize are:

- Seamless integration between on-premises technology and cloud services across multiple cloud regions or geographic locations.
- Dedicated network connections between on-premises and cloud environments for secure, reliable, and highly available connectivity.
- Ability to leverage existing on-premises or third-party security solutions within the cloud, including the ability to import and manage client-generated keys.
- Cloud service functionality that extends into the data center, or the ability to deploy a private cloud within an organization's data center.
- Fault-independent data centers that meet an organization's regulatory requirements and offer full isolation from other tenants, with the option to deploy in dedicated realms.

The specific priorities among these and other capabilities depend greatly on the organization's hybrid needs, both in the types of data and processing required and the reasons for using a hybrid solution. Organizations should define requirements for isolated use of the data, for integrated use of the data, and for seamless support of their full organizational workflow, as a basis for analyzing the priority hybrid capabilities and the right implementation of those capabilities.

Core to Edge Endpoints

Every day, sensors, autonomous platforms, and IoT devices dispersed across the globe produce massive amounts of mission-critical data for organizations to collect. The information provided by this data is often critical to the decision-making of modern operations teams, whose success may entirely hinge upon the reliability and agility of their real-time tactical data.

However, the sheer complexity and scope of independently collecting, storing, processing, and analyzing these large datasets, especially in remote and austere environments, poses several formidable challenges:

1. Real-time data must travel vast distances from its collection point to be processed and analyzed before it becomes available to key decision-makers, creating significant delays that potentially jeopardize an organization's mission.
2. Unreliable infrastructure or austere conditions that destabilize or sever public utilities and data connections, preventing mission-critical data streams from reaching the analytical systems upon which they depend.
3. Data streams broadcasting over interruptible connections that are difficult to resync or lack protections for validating data integrity, creating unreliable data sets with missing or duplicate information.

4. Required computer hardware is not designed to withstand harsh conditions or lacks the portability to easily deploy it in remote locations.

Organizations should leverage tactical edge devices to overcome these obstacles by bringing their required computing capabilities (collection, storage, processing, analysis) closer to their data's generation point. A ruggedized, mobile, and portable tactical edge device loaded with cloud computing capabilities can help resolve many of the issues involved in edge data processing. For instance, it could:

- Collect, store, process, and analyze data close to its source, eliminating the need to transmit large volumes of data over unreliable networks.
- Allow operations teams to analyze data and act on it immediately, thereby empowering their success even while fully isolated from core systems.
- Create a decentralized computing structure that can function independently of its connection to core systems, while also sharing data when connections are available.

Additionally, these disconnected environments should be further enhanced by a data fabric solution capable of withstanding spontaneous interruptions to connectivity. This should include controls for automatically retrying connections, automatic resyncing once connectivity resumes, and built-in protections for validating the integrity of interrupted streams (through trusted event ledgers, for example).

With the right implementation of capabilities on the edge device and integration with the full fabric for broader or longer-term requirements, a platform that fully incorporates high-capability tactical edge devices can effectively address the challenges presented in effectively realizing the value of vast quantities of distributed, real-time data collection.

Data Fabric

For more than 35 years, monolithic architectures were the dominant paradigm for enterprise data systems. However, the rapid adoption of cloud technology since 2015 has led to the development and maturity of mission-critical systems separated across physical locations, data stores, and networks.

Organizations are now contending with fragmented data footprints that are hindering or outright obstructing their ability to carry out their missions. Organizations may be unable or unwilling to move their data out of these dispersed locations due to cost constraints, migration complexity, vendor lock-in, or the unavailability of mission-critical applications (or other crucial capabilities) on different platforms.

Furthermore, polyglot data, poor latency, disrupted communications, networking fees, and inconsistent sources of truth make connecting this data (sometimes across hundreds of kilometers) very difficult, creating expensive and inaccessible data silos.

As such, many organizations worldwide now face serious issues making data accessible to those who need it most, including key decision-makers and strategists.

In such cases, data from these distributed sources must be integrated and unified by a reliable data fabric. As opposed to the physical network infrastructure that routes packets between digital devices, data fabric solutions operate at the data layer to connect heterogenous data sources without physically moving data from its original location.

Today, data fabric is a heavily researched topic covering a wide set of technologies. The consensus is that no single tool encompasses the full breadth of a data fabric. Industry capabilities continue to expand rapidly to address the widespread problem of data availability, and there are many potential solutions available depending on an organization’s needs. An implementation should focus on the areas that are most critical for the organization, such as data cataloging, seamless access, and consistent security control.

Real-Time Data Pipelines

Continuous, real-time data pipelines enable the rapid and continuous flow of information between systems, providing immediate access to data as it is generated or updated. Data pipelines have become utterly vital for most modern organizations, who rely on them for automatic, reliable, real-time communication among their mission-critical systems.

By automating the flow of data between systems in real-time, data pipelines guarantee the reliability of replicated data across an organization's infrastructure. This ensures that analysts and decision-makers work with the most up-to-date information at all times.

Change Data Capture (CDC) is a critical technique within real-time data replication that involves identifying, capturing, and replicating changes made to source data. When data is updated in one system, CDC replicates those changes to other systems in real-time. Instead of processing entire datasets, CDC only extracts and replicates modified data, making it highly efficient for large datasets.

CDC systems rely on event ledgers to track data changes and guarantee data consistency sources and targets. This also enables detailed auditing of data changes, which may be essential for an organization's regulatory compliance and data governance.

Organizations need to consider several factors when determining how to build their data pipelines:

- What kinds of dissimilar or proprietary endpoints (databases, applications, SaaS platforms, etc.) need to be connected across your entire infrastructure?
- What kinds of entryways do these endpoints provide to permit the flow of data in and out of them?
- Are there native connectors to each different endpoint available, and if not, can one be manually configured?
- What is the “shape” of the data being captured, and will the pipeline need to transform this data mid-flight in order to replicate it between dissimilar systems?

- Will the pipeline transform data automatically, or will it require manual configuration?
- What types of data analytics and enrichment occur as part of the pipelines, and what is treated as fully separate from the pipeline, contained within the endpoints?
- Will the pipeline need to mask data mid-flight in order to comply with data privacy or clearance regulations?
- Is the pipeline capable of combining multiple sources into a single target, or is it able to enrich the replication streams with data from external sources?

Batch Processing

Many organizations still rely heavily on batch processes to move data among systems (i.e. exporting static datasets into files for transportation, transformation, and ingestion into target systems). Oftentimes these processes are driven by humans and must be repeatedly carried out on a frequent basis to ensure the continuous reliability of target systems.

Organizations should seek to replace these processes with real-time data pipelines as soon as possible. The risks that human-driven batch processes pose to an organization's continuous operations cannot be overstated. These processes are highly error-prone due to their repetitive nature and the painstaking attention to detail sometimes demanded by the data transformation process.

Likewise, the complexity of these manually executed procedures often turns them into highly specialized tasks. This creates serious problems when specialists leave an organization without handing over their procedural knowledge to a dedicated replacement.

Batch processing procedures should be fully captured and documented to ensure that automation addresses all aspects, including edge cases and specific situations that call for explicit human judgment. When these batch processes cannot be replaced by real-time pipelines, they should be automated either by code or available third-party tools. Many tools allow organizations to not only design automated batch pipelines, but to specify event-driven triggers for instantaneous, real-time processing.

Open-Source Streaming

Many organizations are increasingly relying on open-source streaming solutions to connect their distributed platforms. Some widely adopted solutions include Kafka, Spark Streaming, and Pulsar.

These open-source solutions offer a wide range of capabilities for building real-time data pipelines, each with its own unique strengths and use cases. Developers and organizations should choose the most suitable technology based on their specific requirements.

Organizations concerned with the scalability and performance of their open-source pipelines might consider a compatible cloud-native service. Several platforms provide commercial, enterprise-grade cloud services built on top of open-source systems, such as Confluence Kafka.

These cloud services typically provide additional features, tools, and support not provided by their open-source versions, making them easier to use and manage in production environments. They are also typically designed for production use cases and may include enterprise-grade SLAs and the ability to quickly scale to enterprise-class workloads. Because these services are built on top of open-source systems, they are natively compatible with their non-cloud versions, allowing seamless integration between the two.

Regardless of the tool, streaming pipelines create the potential to immediately absorb and respond to events at scale, or to overwhelm receiving systems. Solution designs should reflect latency requirements, complexity of required processing, requirements and available scale for storage and retention of resulting data, and the requirements for data disposition.

Data Lakes

The exponential growth of data generation in the digital age has presented numerous challenges for organizations, particularly regarding data storage and management. Many organizations are struggling to keep up with the demands of merely collecting and storing their data, let alone capitalizing on it in a meaningful way.

Traditional data storage methods, such as data warehouses, are becoming increasingly inefficient and expensive to maintain as data volumes continue to expand. This is primarily due to their relational structures and limited scalability, which often result in high costs for storing unstructured data.

Data lakes have emerged as a powerful solution to address these modern data storage challenges. A data lake is a repository that allows storing vast amounts of structured, semi-structured, and unstructured data in its native format.

By providing a flexible and scalable storage architecture, data lakes offer a more cost-effective approach to data management. As such, data lakes are typically considered essential components of many modern data platforms, which must scale to the demands of large-scale data ingestion and storage.

Furthermore, data lakes can handle diverse data types without the need for upfront data transformation, reducing the time and resources required for data collection. This flexibility enables organizations to quickly ingest and store large volumes of data at a much lower cost, ensuring that valuable information is not discarded due to storage constraints.

Data lake approaches can also be combined with data warehouse techniques to form data lakehouses, bringing together the best of the flexible, unstructured data access of the data lake and the reliability and performance associated with a data warehouse.

Cloud storage is typically leveraged for nearly all data lake solutions due to its universally low cost and high scalability. While nearly all digital platforms offer cloud storage options at affordable pricing, the estimated cost of networking fees should never be overlooked when selecting a data lake location.

Data ingress fees mean that organizations will not only be charged for the storage of their ever-growing data, but for its continuous ingestion as well. This can impose considerable costs on organizations producing or collecting terabytes or petabytes of data on a daily basis (e.g., regulatory log archiving, real-time event collection and analysis, continual IoT data streams, etc.)

Likewise, data egress fees pose a significant obstacle for organizations struggling to connect their siloed data to other platforms. The impact these egress fees have on “trapping” data in a single location is often quite remarkable, making it difficult for organizations to de-silo data due to cost constraints.

Other storage capabilities that organizations may want to consider are:

- The ability for other clouds, applications, or platforms to interact with it directly or natively query its data in-place.
- The availability and pricing of storage tiering options (i.e. hot vs. cold storage) and the requirements tied to each (e.g. access frequency, performance, etc.)
- Self-service configurability for auto-versioning or auto-tiering.
- Native integrations with other services on its cloud or platform.

With the right capabilities in place, organizations can implement a data lake that is fully tailored to their specific data characteristics and access and processing needs, ranging from managing storage and record retention with automation and financial efficiency through seamless access to heterogeneous data from disparate sources.

Converged Data

Specialized databases, often referred to as "single-purpose" databases, are designed to store and manage only one type of data, such as relational data, time-series data, document-based data, or graph data.

This has led to the creation of data silos dispersed across multiple clouds and data centers. This fragmentation can lead to inconsistent data, errors, and a lack of transparency. Each additional data store also demands individual maintenance and security, growing an organization’s operational overhead with each new data store provisioned.

One way to consolidate polyglot data and reduce the overhead of single-purpose databases is by processing everything from a single location. This is the concept behind a converged database, which is capable of natively storing and processing all modern data types in a single database.

A converged database greatly mitigates many problems associated with fragmented, siloed data:

- Eliminates the need to integrate multiple sources of data.
- Decreases operational overhead from provisioning, configuring, tuning, patching, backing up, and securing multiple individual databases.
- Improves security by reducing potential attack points for intruders (e.g., fewer privileged credentials to secure).

Furthermore, a converged database pairs symbiotically with a polyglot data lake by being able to process any type of data that might land there. This enables large volumes of continuous data to be rapidly collected and stored in elastically growing, cost-effective cloud storage while also making it immediately available for processing in a converged data warehouse, such as a data lakehouse, without any additional steps.

Organizations dealing frequently with polyglot data should include converged databases in their data modernization strategies wherever possible. A key consideration in this strategy is which data sets to rehost and which data sets to leave in place and connect to from the converged database. This choice affects both the transition process and the maintenance of the data moving forward. Ideally, a converged database should enable both strategies, so that data architects can make the right determination for each data set based on current and anticipated needs.

Converged datasets not only simplify operations and significantly reduce overhead but also enable faster analytics at scale and provide the greatest advantages for AI augmentation. Ideally, teams focused on these types of sophisticated processing should engage directly in the implementation of a converged database, for a seamless transition from managing the data to gaining new insights from the data.

AI Enablement

Modern Data Platform design is increasingly focused on incorporating Artificial Intelligence (AI) and Machine Learning (ML) capabilities to unlock the full potential of data. Modern Data Platform pairs well with AI/ML for a number of reasons:

- A Modern Data Platform's ability to process and store large volumes of data efficiently provides the extensive datasets often required for AI model training.
- AI services such as image recognition, document understanding, and sentiment analysis can be quickly deployed on top of vector data natively residing in converged databases.
- Polyglot data pipelines can be used to transport vector data to other data stores, enabling AI across an organization's entire data infrastructure.

Although generative AI is a very recent and emerging technology trend, there are already a number of ways organizations can integrate this capability within their existing technology stacks, once they identify various operational procedures in their organization for enhancement or acceleration with generative AI. For example, implementations can:

- Provide open-ended chat as an addition or companion to existing applications, using generative AI in combination with organization-specific data, to provide information specifically relevant to the organization's user base.
- Create task-specific prompts or prompt templates and integrate their use with existing applications, providing recommendations or results without prompting by individual end users.

With either approach, organizations can achieve rapid value from a combination of the data platform approaches described here and the availability of generative AI models hosted on a variety of platforms. Organizations can:

- Take advantage of the extensive datasets provided by data lakes or converged databases to begin generating vectors for their private business data to prepare for potential AI use cases.
- Leverage existing solutions with the capability to call out (via REST) to AI services such as OpenGPT. This can provide a highly flexible way to immediately integrate AI capabilities into existing deployments without needing to heavily modify them first.

Analytics

Real-time analytics is often a desired outcome of many modern data platforms. There are many analytics solutions available today that provide various benefits to organizations. For Modern Data Platform architectures, organizations should prioritize the following features:

- Scalability to accommodate future growth in data volume and user demands. A flexible architecture allows for easy adaptation to changing business requirements and the incorporation of new technologies.
- Self-service capabilities for building interactive dashboards, generating reports, and visualizing data to help users interpret data and make informed decisions.
- Support for a range of analytics tools and techniques, including descriptive, diagnostic, predictive, and prescriptive analytics. Support for advanced analytics, machine learning, and AI is essential for extracting valuable insights.
- Ability to handle the scale and variety of an organization's data. This includes the ability to process structured, semi-structured, and unstructured data from various sources, such as databases, IoT devices, social media, and more.
- Support for real-time analytics driven by continuous, reliable data streams.

Many of these capabilities, such as the ability to handle polyglot data, are enhanced by the underlying components supporting the analytics solution (e.g., converged databases optimize an analytics platforms capability to visualize polyglot data).

This highlights one of the most important advantages of Modern Data Platform – namely, that the advantages provided by its individual components enhance the benefits of the overall architecture. Implementing these components in ways that align with the character of the organization's data and operational tempo empowers data analysts and data scientists to derive the full value from incoming and existing data sources.

Unlock Your Data with MANTECH and Oracle

The U.S. Government—including the Department of Defense, Intelligence Community, and Federal Civilian agencies—is confronting a complex and rapidly evolving landscape marked by rising global tensions, data overload and information management, aging infrastructure and technical debt, and growing budget and resource constraints. Government agencies need solutions that drive cost efficiency, reduce time- to- value, and reduce risk.

To help solve these challenges, MANTECH and Oracle are collaborating to offer solutions at an outstanding price-performance ratio, powered by data and embedded AI. Oracle Cloud and integrated Modern Data Platform solutions, can help today's defense, intelligence and civilian agencies can achieve decision advantage, reduce technical debt, and safeguard data against adversaries in a zero-trust environment. This allows the Government to leverage proven, commercial technology and an open, software-defined approach to address mission challenges. As discussed in this paper, there are many considerations in designing, deploying and operating a modern data platform, and unlocking the value of your data with AI and Analytics. As a trusted, multi-cloud national security company, MANTECH brings the necessary technology and mission expertise together to ensure effective outcomes.

Why do Federal agencies choose MANTECH and Oracle?

Trusted partners

Both MANTECH and Oracle have a long and proven history of working closely with the government defense, intelligence and civilian community. Today, 1,000+ public sector organizations and 100% of federal cabinet agencies build, modernize, and innovate using Oracle technology.

MANTECH has deep expertise in managing the government's critical data of all types, from low-level foundations through advanced analytics and insights with custom machine learning models and generative AI. 80% of MANTECH 's employees hold security clearances, and 45% are veterans. Together with Oracle Cloud and Modern Data Platform solutions, MANTECH offers clients the latest analytics and generative AI innovations to help leverage all data structures to gain mission advantage.

Committed to cloud innovation

There’s a reason why Oracle is the fastest-growing cloud: Oracle Cloud was designed to withstand the rigorous demands of critical enterprise workloads using a modern architecture. This means that you get a better-engineered, second-generation cloud designed to provide the best price-performance ratio, more flexibility, and improved security.

Protecting your data

MANTECH’s implementations offer a multi-layered security approach. They include security practices and procedures that protect data and infrastructure with internal controls, governance, and oversight. Defense, intelligence and civilian agencies can safely use your own data for everything from securing command at the tactical edge to turning satellite imagery into actionable intelligence.

MANTECH™
ALWAYS ADVANCING

ORACLE | Partner

For more information, please contact **AI@mantech.com**