The Modern Data Environment
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprecedented Opportunity</td>
<td>1</td>
</tr>
<tr>
<td>Fundamental Elements of Modern Data Environment</td>
<td>1</td>
</tr>
<tr>
<td>Roadmap to Modernization</td>
<td>6</td>
</tr>
<tr>
<td>Summary and Conclusion</td>
<td>6</td>
</tr>
</tbody>
</table>
Unprecedented Opportunity

Traditional enterprise data environments typically include data warehouse (DW) and business intelligence (BI) methods and tools as common features. These methods and tools were introduced several decades ago to provide reporting and analytical capabilities without negatively impacting the performance of core transaction processing systems, and most served that purpose very well.

Enterprises today, however, are facing growing pressures to enable faster user access, exploration of more and different types of data sources, and greater business control over analytics content and workflow. These pressures are stretching the limits of traditional data warehouses and business intelligence tools. As a result, disruptive technologies are emerging that provide greater self-service capabilities and easier, faster data acquisition and preparation for analysis.

At the nexus of the growing pressure for end user control over the analytic workflow and the emergence of powerful disruptive technologies lies an unprecedented opportunity for enterprises to embrace and transition to a modern data environment. Seizing this opportunity allows enterprise to overcome traditional DW and BI limitations, enhance enterprise performance, and accelerate achievement of desirable business outcomes.

This white paper describes the fundamental elements of a modern data environment and identifies the key steps in a roadmap for successful transformation from the current to the target state.

Fundamental Elements of Modern Data Environment

Successful transformation to a modern data environment requires much more than just inserting new technologies into an existing data platform. It demands the establishment of guiding principles, roles and responsibilities, governance processes, and critical capabilities of a modern data environment platform.

The following paragraphs describe the salient characteristics of each of these fundamental elements:

**Principles**

Enterprises should embrace three key principles to guide and underpin a successful transition to a modern data environment

- **Business-Led, IT Enabled Governance (IT, Data, and Analytics Governance)** – A key guiding principle for the modern data environment is that IT is an enabler of business-led decisions regarding data sources, analytics priorities, analytics content, and reporting needs.
Self-Service Mindset – Related to the first principle, embracing a self-service mindset is fundamental to the modern data environment. That is, within the established governance rules, business users and data stakeholders should be enabled to execute the full analytics life cycle without requiring specialized technical skills or extensive assistance from information technology professionals. This means that data access, preparation, analysis, visualization, reporting, and dissemination should all be enabled through a data architecture offering extensive self-service capabilities.

Bi-modal Implementation – Transitioning to and sustaining a modern data environment demands agility and flexibility provided by a bi-modal implementation strategy, as illustrated in Figure 2. Gartner recommends bi-modal IT as a best practice, executing a Mode 1 linear approach to protect the core value of any IT service and a Mode 2 nonlinear approach to exploit disruption and embrace risk to fail early and often and evolve quickly based on lessons learned to accommodate innovations.

Transitioning to a modern data environment requires both types of approaches:

- Mode 1 to maintain core functionality provided by traditional DW and BI capabilities, such as ETL processes foundational to existing standard reports
- Mode 2 to explore viability and value of exploiting new and emerging data sources and self-service data integration, for example.

---

**Figure 1, Bi-modal Implementation**

---

![Topographic Map](https://via.placeholder.com/150)
Roles and Responsibilities

Traditional data environments typically involve two primary types of roles and responsibilities: IT and Business. The IT organization, usually under the leadership of a CIO, drives the tool selection process and the implementation of data warehouse and business intelligence technologies and solutions in the form of data marts and data warehouses and the generation of reports and dashboards from the data in those data stores for use by business stakeholders. Business users are involved to the extent that they articulate reporting requirements, and, in some cases, serve as data stewards who designate authoritative data sources and validate ETL logical and rules, for example:

Enterprises with modern data environments engage business leaders as drivers and decision authorities who set priorities for analytics requirements, sources, and content, IT leaders as enablers, and engage additional governance roles and responsibilities actively in the design, implementation, and evolution of the data environment, as described below:

▶ Enterprise Data Stakeholders – any end user or steward of enterprise data analytics products.
▶ End Users – Any authorized recipient of a data analytics product – report, data visualization, dashboard, etc. (internal or external to the organization).
▶ Data Stewards – Business or IT stakeholders with designated authority to define data concepts, specify authorized data sources, assign data access rules and data quality standards.
▶ Analytics Stewards – Business stakeholders with designated authority to establish priorities for analytics initiatives tied to business needs and to execute the full advanced analytics life cycle from data source evaluation and data discovery and exploration through analytics development and insight sharing.
▶ IT Stewards – IT stakeholders with designated authority to establish IT priorities and make decisions regarding and supporting the implementation of analytics solution architectures, tools, and platforms that enable capabilities required by data stewards and analytics stewards needs.

Governance Processes

Three types of governance are essential to a modern data environment: IT, Data, and Analytics Governance. Each type of governance is distinct, with a specific domain of accountability (i.e., IT, data, and analytics); but all are related and should be inter-connected via established touchpoints:

▶ IT Platform Governance – is the accountability, authority, and processes to select, implement, and manage the enabling modern data environment platform – that is, the reference architecture, the specific set of technologies, and the deployment model (i.e., on-premise, cloud, hosted, or hybrid.) Platform Governance guides tool identification, evaluation, selection, replacement, and integration activities.
▶ Data Governance is the accountability, authority, and processes – to manage data assets in the best interest of the overall enterprise (including data quality, data security, etc.) Data Provenance/Lineage, Data Quality Assurance, Data Security, Acquisition/Preparation, Analysis.
▶ Analytics Governance – is the accountability, authority, and processes to determine business priorities for and to develop and control the analytic workflow and access to analytic content (i.e., analytical algorithms, reports, data visualizations, and dashboards.) Analytics Governance must ensure bi-modal analytics implementation of analytics solutions with Mode 2 enabling Fail Early, Fail Often, and Evolve Quickly strategies.
Critical Capabilities of Modern Data Environment Platform

According to Gartner, IT should establish “a self-contained architecture that enables nontechnical users to autonomously execute full-spectrum analytic workflows from data access, ingestion and preparation to interactive analysis and the collaborative sharing of insights.”

- **Reference Architecture** – Figure 2 depicts a proven Reference Architecture for implementing a modern data environment platform.

Critical Capabilities - The critical capabilities of the modern data environment platform can be organized into primarily four categories: Data Sources, Data Services, Analytical Services, and Presentation Services.

- **Data Sources** – The modern data environment platform extracts data from structured data sources that include operational, transactional, external, and reference data. The platform can ingest structured data but typically ingests unstructured and semi-structured data from high velocity, volume, and variety sources like social media, log files, sensors, machine data, and data created by the Internet of Things (IOT). The modern data environment can ingest data from any data source in any format in batch mode or in near real time mode through streaming. It can access and import data from any RDBMS (Oracle, SQL Server, PostgreSQL, and MySQL), Excel, and CSV or tab-delimited text files. In addition, it can ingest and store text, sound, images, and any other data source in efficient formats.
• **Data Services** – As the reference architecture in Figure 1 depicts, in modern data environments the relational Enterprise Data Warehouse (EDW) is no longer the only authoritative data store and foundation for analytics. This means that the data flow and data integration need not be centralized or physical. For example, virtual data integration using modern technologies based on Apache Spark capabilities enables just-in-time data integration and exploits the power of the cloud and reduces the strain on internal infrastructures. The modern data environment platform extracts and transforms data into target data structures and loads them into enterprise data warehouses, data marts, and relational databases using ETL and data management tools. Big Data Integration services use streams and bulk data transfers to land data into the “Data Lake” built on the Hadoop Distributed File System (HDFS). We define the semantics of this data when needed using NoSQL and Hive meta-data. Commonly called “Schema-on-Read”, this approach allows us to read new formats without prior understanding of structure, reduces the need for heavy data management, and allows us to define new meanings for the same data many times, and many different ways. Furthermore, commodity hardware economically stores HDFS data at a rate of $1K per Terabyte. This is far less than the $35K to $100K per Terabyte costs in a typical Data warehouse and allows us to store far more data for far longer which can result in far richer analytics. Note the interconnected services between the Data Lake and the DW. These pathways allow data to migrate from the Data Lake to the data warehouse after it has been parsed, cleansed, and aggregated and its analytical value has been assessed. Our solution provides a range of data services, some require advanced scripting skills, while others enable end-users to find, transform, enrich, and share data sets with others. Tools in this class of self-service data integration include Talend, Informatica, pgAdmin, and SQL Developer. New classes of Big Data self-service tools include Tamr and Trifacta – easy-to-use against HDFS structures. The modern data platform integrates these tools to enable end users to explore and enrich the available data.

• **Analytics Services** – These services transform data into insights using Descriptive, Diagnostic, Predictive, and Prescriptive analytics, as well as Big Data analytics for Discovery, Machine Learning, Text Analysis, and Search. With these Analytics Services, end-users can find relationships and build descriptive and predictive models, to inform leadership and suggest proactive actions. We make a variety of technologies available, such as R, to allow end-users to perform statistical analysis to derive cause-and-effect relationships. The enabling tools available include conventional statistical methods (covariance, correlation, factor analysis, principal component analysis, regression, and multivariate analysis) and machine learning methods (classifiers, decision trees, random forests, neural networks, support vector machines, and hybrid models), as well as data visualization approaches that leverage human pattern-recognition abilities.

• **Presentation Services** – End-users interact with a wide-array of analytics products through web browsers, thick clients, and mobile devices. These presentation products include reports and dashboards, data visualization, ad hoc and custom reports, OLAP, ad hoc searching, data discovery, predictions and outlier discovery, and generate alerts and workflows. Our solution presents data from analyses in easy-to-read, understand, and drill formats using various visualization tools and techniques. In the development of visualizations, end-users apply their understanding of the business objectives to define user stories. They use these user stories to develop storyboards for mapping the visualization of the analytics using powerful self-service tools like Tableau.

In a modern data environment, the enterprise deploys these critical capabilities following an agile approach that delivers “slivers” of high value capability quickly and in rapid succession (“early and often.”)
Roadmap to Modernization

The details of each enterprise’s roadmap will vary based on needs and specific enterprises, but the following are the four essential steps that are common to all successful roadmaps:

- Embrace Guiding Principles
- Establish Governance Framework
- Implement Modern Data Platform
- Innovate with Mode 2 Design and IT Practices

Summary and Conclusion

Enterprises today have an unprecedented opportunity to embrace and transition to a modern data environment that will enable them to overcome traditional DW and BI limitations and significantly enhance enterprise performance to accelerate achievement of desirable business outcomes. To seize this opportunity, enterprises should embrace the guiding principles, establish the three-pronged governance framework that establishes accountability and empowers people, exploit emerging technologies that provide the critical capabilities of the modern data environment platform, and supplement existing data capabilities with advanced analytics following agile development practices that yield high value quickly at low risk.