Data Strategy for the Ultimate AI Lab

Guidehouse

Overview — Building the Ultimate AI Lab

As public and private sector leaders develop data, cloud, and Artificial Intelligence (AI) / Machine Learning (ML) strategies, the need for a unified framework has never been more apparent. These strategies naturally overlap, presenting challenges spanning strategy development, implementation, continuous delivery, and operations.

Curated, tracked, and governed data served through cloud native infrastructure fuels impactful and high value AI/ML solutions. Organizations and their respective data, cloud, and AI divisions need to coalesce around a unified vision, strategy, and execution plan. Additionally, they must bring to bear rigorous engineering practices, coupled with a scientific approach to continuously deliver solutions that shorten the **time to customer value (TTCV)** and build a sustained competitive advantage.

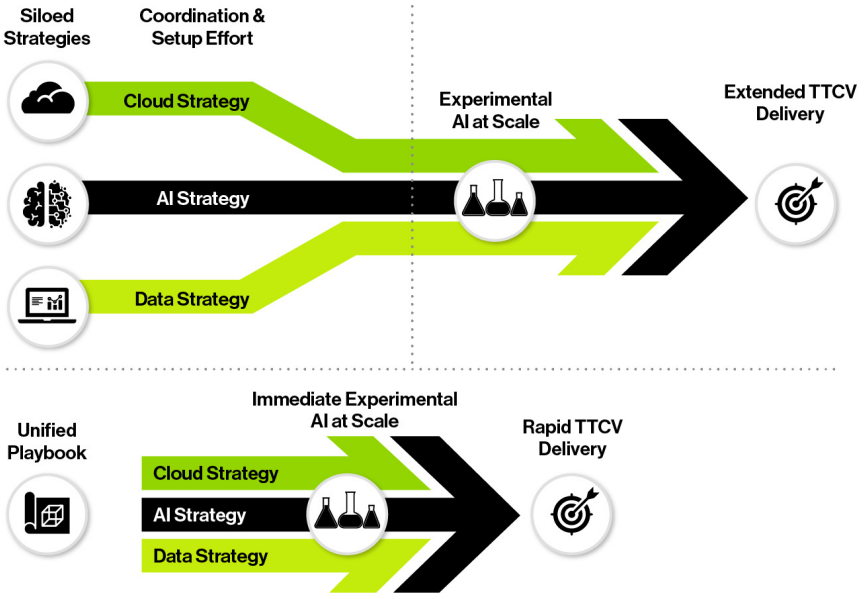


Figure 1: Harmonizing AI, Cloud, and Data Strategies expedites TTCV

This white paper explores some of the challenges and necessary considerations when developing, maintaining, and governing Data solutions. While setting a data strategy is essential to any organization, it is also crucial to consider how the data will be used to generate insight and value; an integrated platform strategy enables AI teams to develop an Ultimate Lab where data is discoverable and computational resources are scalable and secure. Accordingly, this series also includes the **Cloud Strategy for the Ultimate AI Lab** and **AI Strategy for the Ultimate AI Lab** white papers, which explore the implementation of each respective strategy, along with a culminating business case for strategy integration: **The Ultimate AI Lab.**

Leveraging its IT strategy, cloud operations, AI, and data governance expertise, Guidehouse provides a unified framework for addressing each of these strategic domains driven by business goals. Our approach harmonizes these interdependencies to develop an organizational strategy focused on delivering and shortening TTCV.

Data Strategy

A data strategy is critical to aligning data-related needs and planned usage to priority enterprise goals. It is developed from an understanding of the requirements inherent to mission and business priorities. The strategy's goals and objectives serve as the framework for subsequent implementation of plans and actions.

Organizations adopt a data strategy for a variety of reasons:

- To manage diverse risk elements associated with acquisition, storage, and transmission of data;
- To empower teams to realize the benefits of their data in the value stream; and
- To make their data discoverable, actionable, and secure for the stakeholders who need it.

Figure 2. The DAMA-DMBOK2 Guide Knowledge Area Wheel

Data management is a critical complement and supporting vehicle to the overarching data strategy.

Data management consists of an actionable plan for governing data quality, integrity, access, and security. "*The Data Management Association (DAMA) International Data Management Body of Knowledge 2.0*" (DMBOK2) outlines 10 core knowledge areas as the framework for an effective data management plan, which is depicted in Figure 2.

Often, data strategies incorporate AI-specific goals. As AI requires a fair amount of high quality, well-documented data, Guidehouse recommends the establishment of a functioning data governance and management program that can feed into subsequent cloud and AI strategies and initiatives. As expected, this is not without its challenges:

- Organizations may accumulate data that falls under different regulatory and security requirements. Therefore, they must find storage and access solutions that are compliant with the various regulations; and
- Data can accumulate at extremely high rates and organizations need to have the infrastructure in place to ensure systems are not overwhelmed.

While no two organizations are the same, there are common themes and considerations all organizations need to address in order to navigate these challenges and develop a successful data strategy. In addition to a team of experienced, highly qualified data, cloud, and AI practitioners, organizations need a well-defined change management approach to guide stakeholders and employees through the adoption of their redesigned data management and governance practices. As there are numerous components that go into a data strategy tailored to the unique requirements and priorities of an organization, stakeholders and business leaders need the opportunity to explore their avenues for growth and take calculated, incremental steps towards their envisioned outcomes.

Data Governance

SET THE VISION

As with any project, it is critical to have an established vision and corresponding goals and objectives to guide and align data-related efforts. Understanding where you aim to go sets the directionality of short- and medium-term efforts you choose to invest in. This vision must have buy-in across all levels of the organization so that individuals can understand the value in how their work aligns and contributes to the vision.

Having a designated individual held accountable for setting, driving, and achieving this vision is also highly beneficial to the success of a data governance program. In many organizations, this individual is the chief data officer, who works with other C-suite leaders to realize the vision.

WHO IS RESPONSIBLE?

A cornerstone of enterprise data strategy is to establish clear roles and responsibilities surrounding data. These roles, which include data producers, stewards, custodians, champions, consumers, and more, should be documented as part of the data management plan, which will help provide a clear framework for data governance. When teams do not have clarity about the ownership of data upon discovery, their obligations about the data are not discoverable, and the data is not truly actionable. Identifying specific individuals to assume data governance roles both across an organization's business units and from operational to strategic levels will ensure accountability. Organizations often charter governance councils and committees to guide strategic direction of the governance program, approve and enact standards and policies, support stewardship, and promote best practices. These governance groups support a clear data management structure that will help enable a successful data management practice.

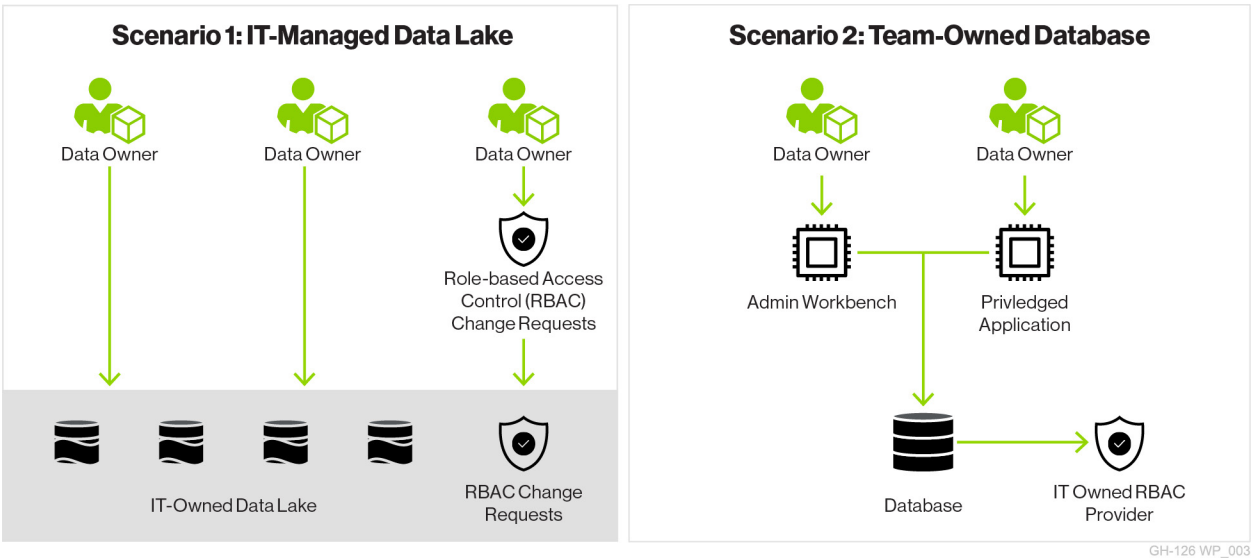


Figure 3. Examples of Data Ownership and Data Stakeholder Access and Control Patterns.

DRIVING DATA QUALITY AND ACCESSIBILITY

Regular maintenance of data management artifacts, such as data catalogs, dictionaries, inventories, and processes, will aid in successful data governance. At the infancy of implementing data strategy and governance, developing a data maturity model and quality scorecard will provide a framework to measure progress against as data standards and policies are enacted and implemented. Further, developing enterprise data models—such as conceptual, logical, and physical data models—will illuminate discrepancies and serve as critical references for resolving existing data issues and mitigating future conflicts when integrating new systems and technologies. Data models can be leveraged as a guide for implementing architectures, databases, and services that enable proper data accessibility.

LEVERAGING AAA

Except in very simple data landscapes, an effective data strategy makes assumptions about the Authentication, Authorization, and Auditing (AAA) that the data architecture can leverage. These three principles control and coordinate access to a defined cloud-based environment and its application servers.[1] The uses for these AAA resources in data strategy are diverse:

- Creating RBAC boundaries to support compliance, operational, or security requirements;
- Recording the identity of individuals and services that access sensitive data; and
- Enabling "self-service," as with customer-facing applications that use or create data.

COMPLIANT DATA ARCHITECTURE

Data architects are often responsible for designing systems that meet a variety of regulatory requirements, such as...

- Physical or software-defined isolation of data;
- Restricted transport of the data over political or geographic boundaries;
- Specific encryption methodologies or other security measures; and
- Restrictions (e.g., International Traffic in Arms Regulation) related to the systems administrators working where the data is stored.

To make matters more difficult, the regulatory environment can be complex and change quickly. Effective data strategy empowers teams and data architects by creating clear guidance about compliance in data storage, processing, and transmission systems. Ideally, data architecture should align to and support a maintained enterprise data model that facilitates the data governance vision and its objectives.

DISASTER RECOVERY

Disaster recovery solutions represent a significant investment in risk mitigation and should be considered essential components of any mission-critical system. Data architects often express disaster recovery objectives in terms of specific, well-understood metrics, such as...

- **Recovery Point Objective (RPO):** The maximum time that a system can roll back for "point in time" recovery, such as to a backup in a database; and
- **Recovery Time Objective (RTO):** The maximum downtime that the system can experience before it "fails over" to a redundancy or recovers.

A complete data strategy does not need to fully define these metrics for the various service categories. Instead, it should give teams clear guidance about how to apply these metrics in a manner that supports the goals of the business.[2]

Architecting Data Solutions

STORAGE

Enterprise storage solutions are a market unto themselves, and each product in the market has its own unique value proposition. The job of an effective data management strategy may, in some circumstances, involve prescribing specific foundational storage solutions. However, the primary purpose of addressing storage in an enterprise data strategy is to describe how the organization will improve value creation as part of managing storage solutions. For example, an enterprise data strategy can define how a legacy storage system will be phased out, a roadmap of migrations, upgrades, and technology adoption, or a plan for managing costs and risks associated with data storage.

DATA LAKE

For the purposes of developing an enterprise data architecture, data lakes support a "store everything" strategy. Organizations that anticipate retaining large quantities of raw or unstructured data are the archetypal users of data lakes. Because object storage solutions offer extremely scalable storage at very low cost, they have become popular for data lake use cases.

Governing data lakes presents unique challenges because only a limited subset of the data can reliably be used as part of the access patterns that data engineers will use to query the lake. This often results in strategies such as treating all data from a particular source as covered by a regulatory requirement, even if only a specific subset is likely covered.

Later, when the raw data is "enriched" and available in a different part of the data architecture, an exclusion rule can be applied to identify the subset.[3]

DATA WAREHOUSE

Data warehouses enable (and sometimes obligate) data producers within the business to make their "data marts" visible across the enterprise. Properly executed, a data warehouse forms the basis for business intelligence at scale. Whereas data lakes require technology leaders to strategize about the opacity of raw, unstructured data, data warehouses present the opposite problem:

- Defining structured, relational data;
- Data quality, integrity, and provenance concerns; and
- Latency of the data.

Technology leaders must carefully consider how a data warehouse can support their strategic objectives. Enterprise data strategy is supporting the success of enterprise data architects when data warehouses deliver actionable intelligence across operations and business value streams.[4]

DATA GRAVITY

Data lakes and warehouses are common sources of a phenomenon known as data gravity. It is the thought that as data accumulates in volume, it seems to also develop qualities similar to density and mass in that it can assert influence over other services and applications like a gravitation pull. The "heavier" the data becomes, the more difficult it is to move. Therefore, services and applications are moved to where the data is stored with the intent of increasing throughput and decreasing latency. While the scalability of an enterprise storage solution should be a major consideration for managing the dynamics of data gravity, the solution should also include a careful examination of the secondary services and applications needed to drive value from the data.[5]

Alignment to Scale

SMALL DATA

Many organizations have value streams that are associated with small, domain-driven data sets. Effective data strategy in these organizations must prioritize for these specific domain requirements before more general "big data" concerns.[6] Supporting the priorities of these value streams takes careful consideration of factors, such as...

- Managing complexity for secure, scalable data architectures;
- Prioritizing data fidelity; and
- Balancing value creation against expansion.

BIG DATA

Organizations often only begin to think about enterprise data strategy when they begin running into "big data problems"—when their systems begin exceeding the scales that they were designed to handle. In order to be successful with big data, organizations need to build a culture that enables scaling. They may develop a recruiting program that enables them to compete for the limited talent pool that has experience building (and using) scalable big data systems and/or develop training programs to upskill their current teams. Data architecture focuses deeply on the limitations of each component that the data touches.[7]

Integration with AI and Cloud Strategy

CLOUD GRAVITY

The selection of services offered between cloud platforms can strongly influence the infrastructure that data systems are built on; this is known as Cloud Gravity. Major cloud providers are competitive, and in many cases, the feature profiles of data systems have significant overlap. Organizations that are looking to maintain a "cloud-agnostic" operational profile should track and manage their cloud centers of gravity.

AI ARTIFACTS

AI projects may call for robust artifact storage, such as for snapshots of inputs, outputs, and metadata for each operation of an AI program. These artifacts can support operations, but they may also be used for compliance. In addition to these snapshots, AI artifacts may consist of the AI models themselves, which can consist of very large data objects. Aligning data strategy with AI strategy to support the use of AI artifact systems can clear the way for collaboration between data operations teams and data science programs.

Conclusion

Developing AI solutions at scale requires an AI strategy that directly incorporates data management and cloud infrastructure solutions tailored to their organization's needs. Integrating data, cloud, and AI strategies enable organizations to fully realize their investment value in each of these domains while collectively acting as a force multiplier.

An effective data strategy ensures that stakeholders across an organization can exploit a variety of data assets to achieve key business objectives. It is fundamentally dependent on having the right infrastructure to enable security, durability, and control. Integrating cloud and data strategies enables data owners to create the correct roles and access patterns within the organization's data infrastructure such that access needs are balanced with regulatory compliance. Additionally, when data is discoverable, secure, and accessible, AI and business teams can rapidly realize and benefit from its value. Unifying data, cloud, and AI strategies maximize return on investment while minimizing TTCV, providing organizations a competitive edge.

Guidehouse brings unique, combined expertise in applied sciences, life sciences, AI, cloud, data governance, IT strategy, and change management. This diverse but tightly coupled set of capabilities and experience uniquely position us to design and build the Ultimate AI Lab, ensuring fertile grounds for ML algorithmic experiments to discover the most promising models and guaranteeing maximum value for our customers.

1. Intelligence Based Authentication-Authorization and Auditing for Secured Data Storage (2015) | G. Ranjith, J. Vijayachandra, P. Sagarika and B. Parthusha
2. Disaster Recovery Planning: A Strategy for Data Security (2000) | Steve M. Hawkins, David C. Yen and David C. Chou
3. Data Lake Management: Challenges and Opportunities (2019) | Fatemah Nargesian, Erkang Zhu, Renee J. Miller, Ken Q. Pu and Patricia C. Arocena
4. Key Organizational Factors in Data Warehouse Architecture Selection (2010) | Thilini Ariyachandra and Hugh Watson
5. Personal Data Lake with Data Gravity Pull (2015) | Hassan Alrehamy and Coral Walker
6. Why We Need a Small Data Paradigm (2019) | Eric B. Hekler, Predrag Klasnja, Guillaume Chevance, Natalie M. Golaszewski, Dana Lewis and Ida Sim
7. Big Data Challenges (2015) | T Nasser and RS Tariq

Guidehouse

Contact Information

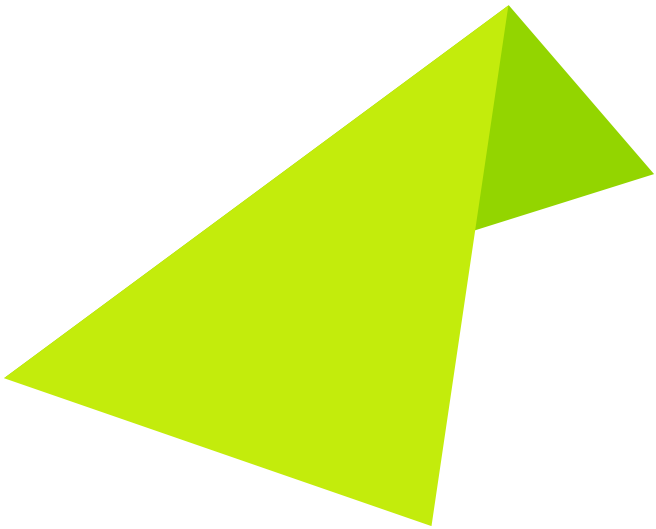
Bassel Haidar
Associate Director, Advanced Analytics and Intelligent Automation
bhaidar@guidehouse.com

Charles Landau
Managing Consultant, Advanced Analytics and Intelligent Automation
clandau@guidehouse.com

Stephanie Ling
Managing Consultant, Advanced Analytics and Intelligent Automation
sling@guidehouse.com

Acknowledgments

The authors would like to acknowledge the following people for their contributions: Alex Gromadzki, Kate Pokrass, Minsoo Kim, Devan Visvalingam, and Stephen Williams



About Guidehouse

Guidehouse is a leading global provider of consulting services to the public and commercial markets with broad capabilities in management, technology, and risk consulting. We help clients address their toughest challenges and navigate significant regulatory pressures with a focus on transformational change, business resiliency, and technology-driven innovation. Across a range of advisory, consulting, outsourcing, and digital services, we create scalable, innovative solutions that prepare our clients for future growth and success. Guidehouse is a Veritas Capital portfolio company, led by seasoned professionals with proven and diverse expertise in traditional and emerging technologies, markets, and agenda-setting issues driving national and global economies. For more information, please visit: **www.guidehouse.com.**