DEFINING A DATA STRATEGY
Introduction

Data, and the ability to strategically deploy IT systems to manage that data, is changing at an unprecedented rate. Data is exploding, arriving from previously untapped sources with a volume and velocity that require IT architectures to smartly evolve with it.

Solutions for managing data are evolving along with the data itself. A dizzying array of new technologies and solutions for storing and processing data are available for IT organizations to leverage. The ability to smartly deploy IT systems and processes that turn the data deluge into a competitive advantage starts with an understanding of the data itself.

This paper explores the environmental forces changing the face of data and provides a framework for developing a data strategy to help organizations deal with both the disruption and the opportunities offered by this changing world.
It wasn't long ago that data was something that organizations collected and used operationally before locking away forever on farms of spinning disks and tape archives. Business data was not fast, fat, or deep in most organizations. That world has changed.

Data today is not simply customer transactions, accounting, and other traditional enterprise applications. Modern data relates to every aspect of an organization's business. Data continues to expand in exciting dimensions, including:

- Sensor data from the Internet of Things (IoT), both in raw form as well as data that has been processed at the edge. IDC recently predicted that the worldwide install base of IoT endpoints will exceed 82 billion devices by 2025.¹
- Video data from connected cameras, security systems, and in-house media streaming.
- Aggregated and raw log data from infrastructure spread across the enterprise.
- Raw web tracking and customer analytics data.
- Customer engagement data from social media, customer chat logs, and raw audio from voice conversations.
- Machine learning and artificial intelligence systems for image recognition, voice processing, and pattern recognition from business or IT log data.
- Extended data from partners feeding into CRM and predictive analytics.

The evolving world of data brings with it new opportunities for competitive differentiation for companies who take the time to understand how to leverage their data. Even with those opportunities, many organizations do not leverage their own data at all.²

In a recent article in the Harvard Business Review it was stated that “less than half of an organization’s structured data is actively used in making decisions – and less than 1% of its unstructured data is analyzed or used at all”.²
That is a surprising set of statistics given the current ability to apply analytics and artificial intelligence to leverage insights from data. As never before, IT organizations have the ability to implement systems and processes that deliver new insights for business stakeholders, enable new efficiencies across the organization, and create compelling new digital experiences for their customers.

Understanding and leveraging that data can be the secret to an enterprise’s success and will be the difference between the disruptors and the disrupted. Data has become strategic. Organizations that treat their data as a strategic asset are insight-driven organizations.

At the same time, this new world also brings forward new challenges for IT organizations. Thankfully, the technology industry has evolved along with this data deluge. As the ability to generate and process data has evolved, so have the technologies, data architectures, and best practices that can tame the data deluge and leverage it into new opportunities.

Becoming an insight-driven organization requires deliberate action and conscious effort on the part of both the business’s leadership and its IT organization. It requires understanding that data that the organization collects, and the capabilities that exist to exploit that data for competitive advantage.

Becoming an insight-driven organization requires a data strategy.
Every enterprise has a long-range plan, but most do not have a strategic data strategy that maps to that long-range plan. A data strategy looks holistically at the needs of a business mapped against the gamut of data that may be available. This mapping leads naturally toward a roadmap of extracting value from data, and how IT architecture must evolve to support that vision. It sounds like a simple process, but the complexities of modern data processing and storage are mind boggling.

A data strategy comprehends not just data, but the rapidly changing range of targeted IT solutions for storing and analyzing that data. It’s no longer as simple as plugging in a server and a storage array and loading up an application. There is rapid disruption in the data management space, with traditional storage arrays being supplemented with new architectures such as converged and hyper-converged infrastructure, various flavors of software defined storage, and cloud and hybrid-cloud storage.

A data strategy is composed of four basic elements:

1. **Goals and Objectives**: A clear understanding of the value for each data element for any analytics needed by the organization.

2. **Data Inventory**: A comprehensive understanding of the attributes and access needs of data being processed and stored.

3. **Data Architecture**: A Dmapping of data to the compute resources and applications that will ultimately consume and process an organization’s data.

4. **Data Protection**: A data retention and protection strategy, defining how various data must be maintained long-term to satisfy both corporate needs and governmental regulations.

---

### Goals and Objectives

The first and most critical step in defining a data strategy is simply articulating an understanding of what it is that the data strategy is supposed to accomplish. While many of the specifics of this activity will be driven by the needs of the business, there are some commonalities.

The goals and objectives of a data strategy should include elements of the following:

- **Ensure alignment between the long-term strategic IT plan and the needs of the business.**
- **Understand the totality of the data collected by the enterprise, and the systems that process that data.**
- **Document gaps between business goals and the data collected to address those goals, along with a plan to close those gaps.**
- **Identify untapped opportunities within the data and document a plan to exploit those opportunities for competitive advantage.** An example of this might be the application of new machine learning algorithms against historic transaction data to identify non-obvious purchasing patterns.
- **Alignment of IT resources against the resulting data strategy.**

The needs of every organization will be different, but it is important that the organization define what it is that is being accomplished.
A critical step in building a data strategy is getting a handle on the nature, character, and needs of the data. This sounds simple, but even a modest sized organization will have data spread across multiple sources, each with multiple needs.

IT organizations all strive for well-defined “data lakes,” where structured and unstructured data co-exist. In reality, most organizations have some amount of “data swamps,” where pockets of data are spread across the enterprise, with haphazard consistency between them. A data inventory will help identify the swamps and move the organization towards a coherent data lake.

The first step in taking data inventory is the simple act of understanding exactly what data is collected. Enumerate the known data sources, listing out the applications (if any) that utilize that data. The size of the organization will determine the complexity of this step – smaller organizations may have one or two resources to target, while larger enterprises may have hundreds.

Once the data is listed, the attributes need to be determined. These attributes may span both business and technology realms and will likely differ across organizations. Some common types of questions that need to be answered include:

- Who is the data custodian? This is the person responsible for the business needs of the data.
- Who is allowed to access the data both within and outside the organization? This will determine required access control lists and other business processes for data protection and auditing.
- Where is the data located geographically? This may be an unnecessary question for an organization completely contained within one country’s boundaries. For others, it is a critical question, as local regulations may govern data retention and other related policies.
- Where is the data located within the data center? Locality of data relative to both the applications that consume that data, and the applications that populate the data onto a storage device, has direct implications to the performance of those applications.
- What applications access the data, and at what frequency? This determines the best placement of the data, both in terms of proximity to the applications that consume it, and the best underlying storage technology to house it. Data targeted to frequent high-speed random access from a machine learning application has very different needs than server logs that will only be accessed periodically.
- Does the data have any special needs? Examples of these special needs might be encryption either at rest or in transit, or whether the data has specific retention needs. These are the types of attributes that are likely defined by regulations governing the type of data in the geography that the data is being stored.

This is not a comprehensive set of questions, but it is representative of the types of cataloging that should be done as part of a data inventory. The actual needs will vary based on the organization doing the inventory.

As an organization takes on a project of this type, it's imperative that stakeholders across the business are engaged and agree on what should be cataloged as part of the inventory. There should be a clear rationale for everything collected, and it should be mapped to a direct action, decision, or policy affecting that data.
Data Architecture

While a data inventory describes an organization’s data, it is really the applications that process the data that transform raw bits into business insights.

A data strategy includes the critical step of mapping the needs of the applications against the characteristics of the data. The questions that an organization asks in this phase will be very dependent upon type of applications that are deployed and the goals of processing the data.

It is nearly impossible to think about data architecture without also thinking about the underlying storage architecture. After all, the needs of data processing and analytics applications directly impacts how and where data is stored.

Example elements to think about while assessing the needs of the application may include the following:

- Is data shared between applications?
- Is there overlap in data being served or generated between applications? If so, does there need to be a rationalization exercise to either remove redundancies, or otherwise ensure replicated data remains consistent?
- How fast does the data need to be served to the application? Will the application stall while waiting on data? If so, does it materially impact the ability to achieve the goals of the processing?
- Does the application migrate between servers, or even between on premise data center and a public/private cloud? If so, does the data set migrate with it? Does it need to?
- Does the application better consume data from traditional files or objects?
- Is the data structured or unstructured? What are the implications of that to both storage and processing?

It may turn out, for example, that a streaming media system will operate better in a server with an SSD disk dedicated to do caching. Or perhaps a CRM system can tolerate the long latencies that would be offered by a remotely located storage array, while a machine learning system might require a direct-attached all-flash array.

It’s impossible to enumerate all of the possibilities in this paper, but it is a critical component of the data strategy that the data access and availability needs of every business application be understood. This drives data placement and selection criteria for the associated storage devices.

We recommend that the application vendor, or in-house specialist, be engaged to define the parameters needed for crafting and determining the needs of the applications.
Data Protection

Every IT organization has a responsibility to protect data. Data protection includes elements of data retention and availability.

Data must be maintained in compliance with the retention policy but more importantly, data must be available and able to withstand failures within the infrastructure. The data retention and availability requirements will determine long-term storage decisions.

Data retention may be governed by regulation, business needs, or both. Retention may be required on data that is in active use, or data may simply be archived and rarely accessed. Answering these questions will determine the storage hierarchy required to support the policy.

Data that is simply archived, for example, may do well on very slow archival media such as tape backup or remote and slow hard-disk drive storage systems. Customer transaction data, on the other hand, which may be retained for years, may also regularly feed analytics systems and thus need to be continuously available.

Understanding both the usage and retention needs of data enables a data strategy that marries the appropriate underlying storage technologies and architectures to the needs of the business.

During the data inventory, the availability needs of the data should be noted. This will drive disaster recovery and availability choices. Example of availability needs are:

- Full availability is required, driving the need for fail-over capabilities, where both applications and data survive the total failure of a site.
- Tolerance of periods of outage, but where the data must be tolerant of failure and easily recoverable.
- Tolerate an outage long enough to restore data from back-ups.

The important point to take away is that there is no “one size fits all” solution for data availability. The needs of the applications will drive decisions about the mechanisms that need to be employed to ensure that data is appropriately available. Most modest sized organizations will have multiple solutions, each mapping to a different set of applications.

Define these needs as part of a strategy. This will facilitate the decisions needed when choosing storage technologies to house the data for various applications.
Data Strategy is a living concern. Strategies are not just ‘write once’. Effective strategies must evolve with both a business’s needs and the technologies available to support those needs.

Data strategies should be revisited whenever new applications are deployed. They should also be revisited when there is opportunity for significant data center or architecture changes. This might occur when new storage arrays are deployed, or when an organization is looking at deploying HCI to consolidate disparate platforms.

The point is that the data strategy should be reviewed periodically. Ensure that it remains alive and relevant as the business and available technologies evolve.

Build an Advisory Panel

The technology choices available to an IT organization when implementing a data strategy can be challenging. Storage and compute architectures are evolving at a faster pace today than they have in a generation. Business critical systems are gaining new capabilities to take advantage of evolving storage and compute functionality.

To supplement the knowledge of an IT organization, it’s important to supplement internal expertise with that of industry. Engage industry leaders when crafting a data strategy. It is the application developers, system vendors, and technology providers who have a continuing understanding of both current technologies and issues, as well as a grasp on emerging technologies that might impact the future needs of an enterprise.

Create an advisory panel of industry experts, including:

- Experts in the applications that are critical to the execution of the business. This may be in-house expertise, or experts provided by the application vendors. This doesn’t need to include every application in the data center, but those that are truly impactful to the business.
- Representatives of industry-leading infrastructure providers. This may be as simple as engaging the existing vendors of an organization’s server, networking, and storage solutions. It could also include competitors to existing providers, in order to provide broader perspective.
- Core technology providers who are supplying the industry with the enabling components and sub-systems that drive the capabilities of the data center. Including the technology providers in the conversation will ensure the broadest possible perspective on current and future trends in data storage and processing.
Building a comprehensive data strategy needs to be a key element of any enterprise’s plan for success, but it’s not something that an organization should try and do alone. Build an industry advisory panel. Include technologists from your current partners who provide the key capabilities in enterprise software for your industry today. Engage OEM partners, such as Hewlett Packard Enterprise*, Cisco Systems*, or Dell EMC*, who are all focused on delivering solutions.

It’s also important to engage broader thinkers, who are both powering today’s solutions as well as thinking about the near-term evolution of data storage and compute. In the enterprise data center world, that’s a pretty short list. Partners such as Intel can provide insight into the technologies that exist today and the direction in which those technologies are evolving.

Data has become strategic. Managing and processing data has become complex and confusing. Tomorrow’s successful companies will have a data strategy that lives and evolves side-by-side with their long-term plan. Much like the long-term plan is created with the input of customer advisory panels and other external inputs, a data strategy needs to be informed by the IT leaders supporting your industry.