› White Paper

The 5 Essential Components of a Data Strategy

sas
THE POWER TO KNOW.

Contents

Data Strategy: What Problem Does It Solve? 1

Data: Past and Present 1

The Business Without a Data Strategy 2

Data Strategy Defined 3

The 5 Components of a Data Strategy 3

- Identify 4
- Store 4
- Provision 6
- Integrate 6
- Govern 7

Defining a Data Strategy Is Key 8

The Power of a Data Strategy 8

Learn More 9

Evan Levy, Vice President of Data Management Programs at SAS, is an acknowledged speaker and writer in the areas of enterprise data strategy, data management and systems integration. With business experiencing exponential growth in data volumes, sources and systems, Levy advises clients on strategies to address business challenges using their existing data and technology assets coupled with new and creative methods and practices.

Despite heavy, long-term investments in data management, data problems at many organizations continue to grow. One reason is that data has traditionally been perceived as just one aspect of a technology project; it has not been treated as a corporate asset. Consequently, the belief was that traditional application and database planning efforts were sufficient to address ongoing data issues.

As our corporate data stores have grown in both size and subject area diversity, it has become clear that a strategy to address data is necessary. Yet some still struggle with the idea that corporate data needs a comprehensive strategy.

There's no shortage of blue-sky thinking when it comes to organizations' strategic plans and road maps. To many, such efforts are just a novelty. Indeed, organizations' strategic plans often generate very few tangible results for organizations – only lots of meetings and documentation. A successful plan, on the other hand, will identify realistic goals along with a road map that provides clear guidance on how to best get the job done.

Let's see how this played out in real life at one organization that set out to develop a data strategy.

Data Strategy: What Problem Does It Solve?

Some time ago I was leading a consulting team at a large bank that was developing a data strategy. From the start, the project champion had found it hard to get his VP to understand the need for and importance of a data strategy. Why?

The bank was already successful. Its revenue and costs were well-managed, and the individual business units and technology groups were good at delivering against their commitments. To the bank's credit, it wasn't complacent. Management was always looking for ways to increase staff members' productivity and reduce ongoing costs. There were all kinds of metrics and key performance indicators (KPIs) to measure IT performance, business benefits and total cost of ownership. The idea of building yet another road map to address a problem that wasn't well-understood met with pushback.

The VP gave his explanation along with some questions:

> "We've got dozens of projects going on at any given time. We're very good at managing our storage needs, our application systems, the analytical platforms, software costs and individual project budgets. Every project identifies staff and resource costs, and we don't ever move forward without the business covering the costs.
>
> **Why do we need a data strategy? What problem are you solving?"**

With the bank doing so many things right, he needed to understand why and how a data strategy would make a difference. To answer these questions, we first have to consider how data was created and used in the past compared to how it's created and used today.

Data: Past and Present

Once upon a time, data was perceived as a byproduct of a business activity or process. It had little value after the process was completed. While there might have been one or two other applications that needed to access the content for follow-up (e.g., customer service, special reports, audits, etc.), these were usually one-off activities.

Today, business is very different. The value of data is accepted; the results of reporting and analytics have made data the secret sauce of many new business initiatives. It's common for application data to be shared with as many as 10 other systems.

While the value of data has evolved tremendously over the past 20 years – and business users recognize it – few companies have adjusted their approaches to capturing, sharing and managing corporate data assets. Their behavior reflects an outdated, underlying belief that data is simply an application byproduct.

Organizations need to create data strategies that match today's realities. To build such a comprehensive data strategy, they need to account for current business and technology commitments while also addressing new goals and objectives.

The Business Without a Data Strategy

Thinking back to our story, the bank executive's concerns were not hard to understand. He spent lots of time wading through project proposals that his devoted staff was incredibly emotional about. In many instances, his team's project proposals were about delivering perfection – turning something that already worked into something faster, stronger or better. The executive understood the world of finite budgets and resources where any new approved project would ultimately take funding and resources away from another request. His mantra was well-known:

> "Tell me why your idea is more important than the items already on the priority list."

We were prepared for this discussion.

We weren't challenging the premise or value of any individual project. The problem was the approach that each individual project and activity took. Each activity addressed data needs independently from one another without any awareness of the overlapping efforts and costs.

- Most projects required access to the same data content. Unfortunately, there was no coordination to prevent overlapping (and wasted) work.
- There was no data sharing, no data reuse, or any economies-of-scale activities to simplify or reduce the cost of data movement and development.
- Business users accessed common data across separate applications. Data value names and formatting varied across applications.
- Users found inconsistencies across reports because source data wasn't documented, and it varied across individual reports.

The result was duplicate data, processing overlaps and little awareness that individual projects were replicating work. There wasn't anything in place to support communicating, collaborating or sharing data methods and practices across projects and systems.

The problem: Every project at the bank addressed data issues as one-off, built-from-scratch activities.

Case Study: The Bank's Data Challenges

The bank's IT team had 17 projects underway (new applications, application enhancements, new reports, etc.).

- Each project required access to customer data, and each had overlapping tasks and resources.
- Every project included a source data inventory and analysis activity because there was no way to know where specific data resided.
- New data extracts (subsets of the application's data copied for use by other systems) had to be built because IT had no way of determining if the data was already available.
- No two teams shared their source extract data. Each had their own copies to support their integration and database build activities (which tied up storage for this transient content).
- Each team's integration logic was custom built and individually maintained, because the logic and rules weren't identified or documented to be shared.

The business staff – dependent on its own operational and reporting efforts – had experienced other challenges:

- Marketing had to continually update its campaign system to adjust to frequent (and uncommunicated) changes occurring to the layouts of the extracts it received.
- Sales managers always had questions about KPI reports with customer details because titles and labels varied across reports (even though they contained common data).
- Business unit users often built their own reports instead of using the standard reports from finance, because there was no way to determine the origin of standard report data.
- The data warehousing team had to continually chase data problems because data issues weren't managed like other business support activities.

Data Strategy Defined

The concepts of standards, collaboration and reuse are well-understood across organizations within most companies. Most development teams are well-educated about system architecture, development methods, requirements gathering, testing and even code reusability. Most business teams can recite the concepts of business requirements, business process definition and results measurement. Unfortunately, the notion of applying these concepts to data to support improved accuracy, access, sharing and reuse is still foreign to most organizations.

> A data strategy is a plan designed to improve all of the ways you acquire, store, manage, share and use data.

The idea behind developing a data strategy is to make sure all data resources are positioned in such a way that they can be used, shared and moved easily and efficiently. Data is no longer a byproduct of business processing – it's a critical asset that enables processing and decision making. A data strategy helps by ensuring that data is managed and used like an asset. It provides a common set of goals and objectives across projects to ensure data is used both effectively and efficiently. A data strategy establishes common methods, practices and processes to manage, manipulate and share data across the enterprise in a repeatable manner.

While most companies have multiple data management initiatives underway (metadata, master data management, data governance, data migration, modernization, data integration, data quality, etc.), most efforts are focused on point solutions that address specific project or organizational needs. A data strategy establishes a road map for aligning these activities across each data management discipline in such a way that they complement and build on one another to deliver greater benefits.

The 5 Components of a Data Strategy

Historically, IT organizations have defined data strategy with a focus on storage. They've built comprehensive plans for sizing and managing their platforms and they've developed sophisticated methods for handling data retention. While this is certainly important, it actually addresses the tactical aspects of content storage – it's not planning for how to **improve all of the ways you acquire, store, manage, share and use data.**

A data strategy must address data storage, but it must also take into account the way data is identified, accessed, shared, understood and used. To be successful, a data strategy has to include each of the different disciplines within data management. Only then will it address all of the issues related to making data accessible and usable so that it can support today's multitude of processing and decision-making activities.

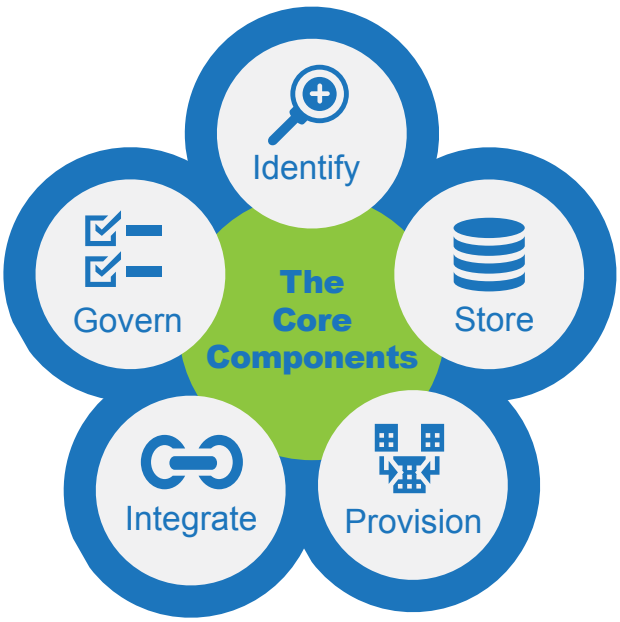

Figure 1: The five core components of a data strategy.

There are five core components of a data strategy that work together as building blocks to comprehensively support data management across an organization: identify, store, provision, integrate and govern.

Identify

Identify data and understand its meaning regardless of structure, origin or location

One of the most basic constructs for using and sharing data within a company is establishing a means to identify and represent the content. Whether it's structured or unstructured content, manipulating and processing data isn't feasible unless the data value has a name, a defined format and value representation (even unstructured data has these details). Establishing consistent data element naming and value conventions is core to using and sharing data. These details should be independent of how the data is stored (in a database, file, etc.) or the physical system where it resides.

It's also important to have a means of referencing and accessing metadata associated with your data (definition, origin, location, domain values, etc.). In much the same way that having an accurate card catalog supports an individual's success in using a library to retrieve a book, successful data usage depends on the existence of metadata (to help retrieve specific data elements). Consolidating business terminology and meaning into a business data glossary is a common means to addressing part of the challenge.

Libraries have card catalogs because it's impractical to remember the location of every book. Metadata is critical for business data usage because it's impossible to know the location and meaning of all of the company's business data – thousands of data elements across numerous data sources. Without data identification details, you would be forced to undertake a data inventory and analysis effort every time you wanted to include new data in your processing or analysis activities.

Without a data glossary and metadata (i.e., the "data card catalog"), companies are likely to ignore some of their most prized data assets because they won't know they exist. If data is truly a corporate asset, a data strategy has to ensure that all of the data can be identified.

Store

Persist data in a structure and location that supports easy, shared access and processing

Data storage is one of the basic capabilities in a company's technology portfolio – yet it is a complex discipline. Most IT organizations have mature methods for identifying and managing the storage needs of individual application systems; each system receives sufficient storage to support its own processing and storage requirements. Whether dealing with transactional processing applications, analytical systems or even general purpose data storage (files, email, pictures, etc.), most organizations use sophisticated methods to plan capacity and allocate storage to the various systems. Unfortunately, this approach only reflects a "data creation" perspective. It does not encompass data sharing and usage.

The gap in this approach is that there's rarely a plan for efficiently managing the storage required to share and move data between systems. The reason is simple; the most visible data sharing in the IT world is transactional in nature. Transactional details between applications are moved and shared to complete a specific business process. Bulk data sharing isn't well-understood and is often perceived as a one-off or infrequent occurrence.

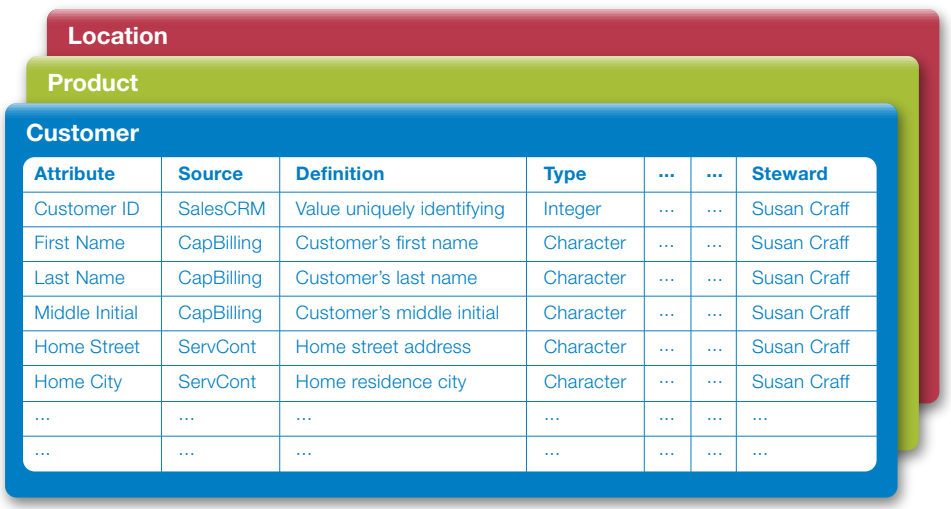

Figure 2: A data card catalog.

With the popularity of big data, the growth of business analytics and increased information sharing between companies, it's much more common to share large volumes (or bulk) data. Most of this shared content falls into two categories: internally created data (customer details, purchase details, etc.) and externally created content (cloud applications, third-party data, syndicated content, etc.). The lack of a centrally managed data sharing process typically forces all systems to manage this space individually, so everyone creates their own copy of the source.

> Forbes magazine[1] identified a medical research facility generating 100 terabytes of data that was ultimately copied and retained by 18 different teams and required more than 10 petabytes of storage.

As organizations have evolved and data assets have grown, it has become clear that storing all data in a single location isn't feasible. It's not that we can't build a system large enough to hold the content. The problem is that the size and distributed nature of our organizations – and the diversity of our data sources – makes loading data into a single platform impractical. Everyone doesn't need access to all of the company's data; they need access to specific data to support their individual needs.

The key is to make sure there's a practical means of storing all the data that's created in a way that allows it to be easily accessed and shared. You don't have to store all the data in one place; you need to store the data once and provide a way for people to find and access it.

We know that once data is created, it will be shared with numerous other systems; it's critical to address storage efficiently, in a way that simplifies access. A good data strategy will ensure that any data created is available for future access without requiring everyone to create their own copies.

[1] forbes.com/sites/ciocentral/2012/07/05/best-practices-for-managing-big-data

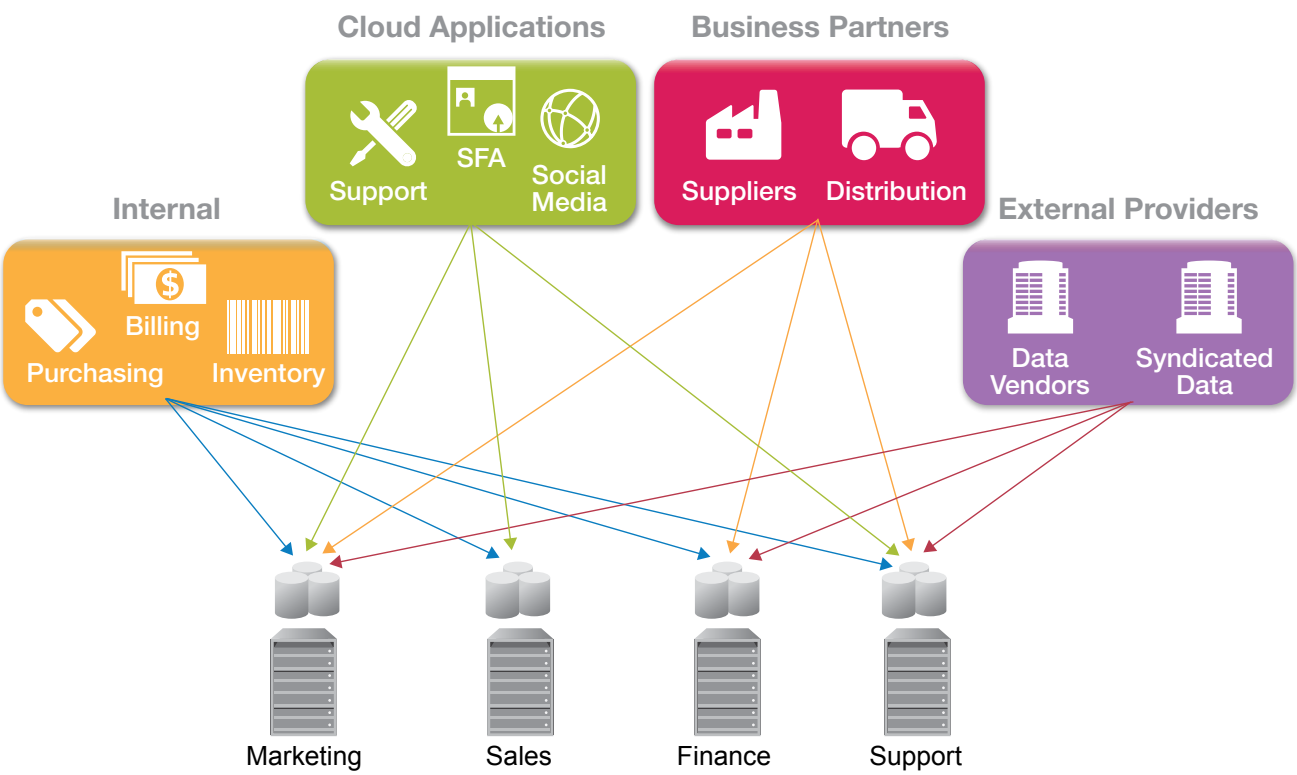

Figure 3: Each system creating its own data copies causes a fourfold increase in storage and processing.

Provision

Package data so it can be reused and shared, and provide rules and access guidelines for the data

In the early days of IT, most application systems were built as individual, independent data processing engines that contained all of the data necessary to perform their defined duties. There was little or no thought given to sharing data across applications. Data was organized and stored for the convenience of the application that collected, created and stored the content.

When the occasional request for data came up, an application developer created an extract by either dumping that data into a file or building a one-off program to support another application's request. The developer didn't think about ongoing data provisioning needs, or data reuse or sharing. At that time, data sharing was infrequent. Today, data sharing is definitely not a specialized need or an infrequent occurrence – data is often used by 10 other systems to support additional business processes and decision making.

But most application systems were not designed to share data. The logic and rules required to decode data for use by others is rarely documented or even known outside of the application development team. Most IT organizations don't provide budget or staff resources to address nontransactional data sharing. Instead, it's handled as a courtesy or convenience – and often addressed as a personal favor between staff members.

When data is shared, it's usually packaged at the convenience of the application developer, not the data user. Such an approach might have been acceptable in years past, when just a few systems and a couple of teams needed access. But it's completely impractical in today's world where IT manages dozens of systems that rely on data from multiple sources to support individual business processes. Packaging and sharing data at the convenience of a single source developer – instead of the individuals managing 10 downstream systems that require the data – is ridiculous. And expecting individuals to learn the idiosyncrasies of dozens of source application systems just so they can use the data is an incredible waste of time.

Sharing data is no longer a specialized technical capability to be addressed by application architects and programmers. It has become a production business need. Businesses are dependent on data being shared and distributed to support both operational and analytical needs. Sharing data can't be managed as a courtesy; the method for packaging and sharing data can't be treated as a one-off need.

If a company's data is truly a corporate asset, then all data must be packaged and prepared for sharing. To treat data as an asset instead of a burden of doing business, a data strategy has to address data provisioning as a standard business process.

Integrate

Move and combine data residing in disparate systems and provide a unified, consistent data view

It's no secret that data integration is one of the more costly IT activities; nearly 40 percent of the cost of new development is consumed by data integration activities. Integration isn't just about traditional data extract, transform and load processes associated with data warehousing; it includes all data (structured, semistructured, unstructured, etc.) and its movement across and between all systems (operational and analytical). The challenge of data integration is to match data across multiple sources without having to use an explicit key or unique identifier.

Developers spend enormous time building logic to match and link values across a multitude of sources. Unfortunately, as each

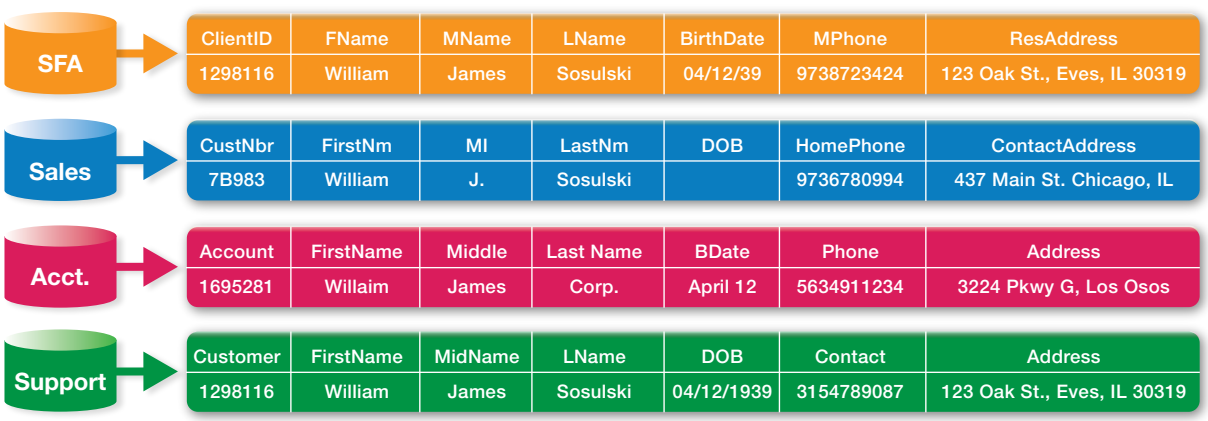

Figure 4: Customer details stored and referenced differently in each operational application.

new development team requires access to individual data sources, they each reconstruct or reinvent the logic needed to link values across the same data sources. The tragedy of data integration is that this rework happens with each new project because what was learned in the past is never captured for reuse.

At most companies, data integration isn't a centralized function. Each system or application team addresses its own needs. While most IT organizations have consolidated data integration activities associated with their own data warehouse into a single team, data integration development still often occurs in other application areas. With the existence of cloud-based applications and business-based technology teams, data integration development is often spread across multiple organizations at a company. Developing code to identify and match records across these individual sources can be quite complex, particularly when some systems require data from 20 or more sources.

What complicates matters even more is that most development teams operate in silos with little awareness of what other teams are doing. The lack of data collaboration tools and methods often prevents teams from realizing available code that they could potentially reuse. Integration should ensure that data is distilled and merged into resultant data sets in a consistent and repeatable method for all consuming systems.

While most organizations have initiatives to address code reuse and collaboration for application development, it's time to identify data development as a discipline that requires the same rigor and methodology. If all the different teams develop and manage their own integration logic, the likelihood of creating (or integrating) data consistently across projects will continue to decrease as the quantities of data sources and volumes spiral. If data is truly a corporate asset, and if accuracy and consistency are critical, a data strategy must include integration as a core component.

Govern

Establish, manage and communicate information policies and mechanisms for effective data usage

Since data is still often perceived as a byproduct of application processing, few organizations have fully developed the methods and processes needed to manage data outside the context of an application and across the enterprise. While many have begun investing in data governance initiatives, many are still in the infancy stage of their respective initiatives.

Most data governance initiatives start by addressing specific tactical issues (e.g., data accuracy, business rule definition or terminology standards) and are confined to specific organizations or project efforts. As governance awareness grows, and as data sharing and usage issues gain visibility, governance initiatives often broaden in scope. As those initiatives expand, organizations may establish a set of information policies, rules and methods to ensure uniform data usage, manipulation and management.

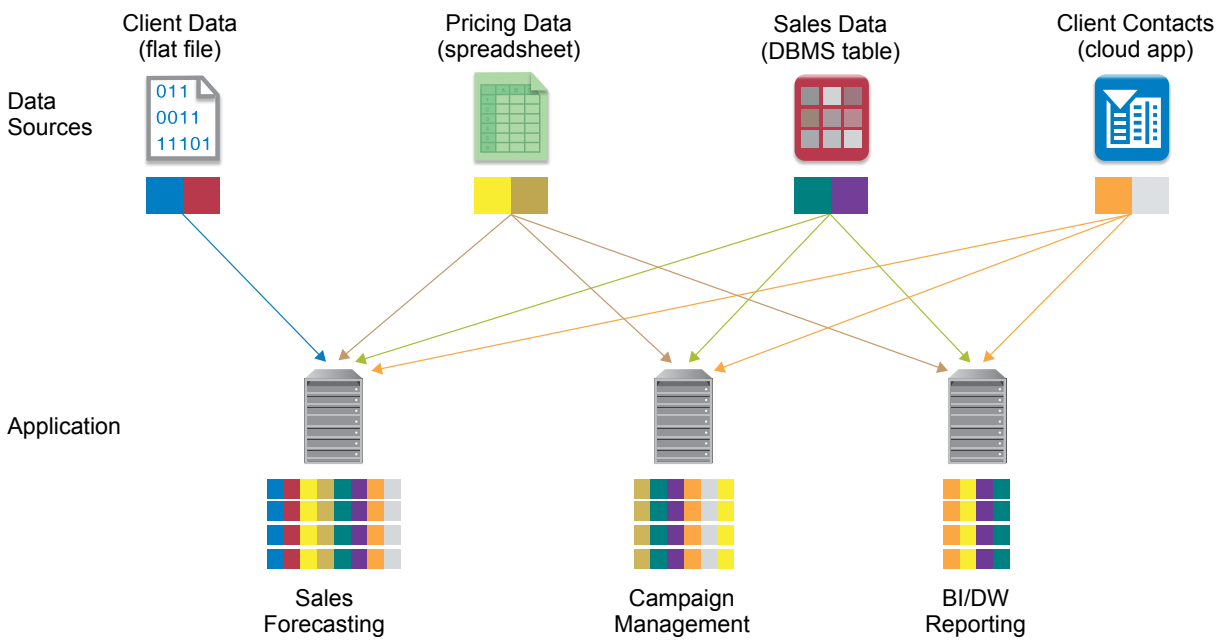

Figure 5: Each data source contains unique data (colored boxes). Since each application creates its own integration logic, the data values may differ across each application.

But all too often data governance is perceived as a rigor specific only to users and the analytics environment. In fact, data governance applies to all applications, systems and staff members. The biggest challenge with data governance is adoption – because data governance is an overarching set of information policies and rules that everyone must respect and follow.

The reason for establishing a strong governance process is to ensure that once data is decoupled from the application that created it, the rules and details of the data are known and respected by all other data constituents. The role governance plays within an overall data strategy is to ensure that data is managed consistently across the company.

Whether it is for determining security details, data correction logic, data naming standards or even establishing new data rules, effective data governance makes sure data is consistently managed, manipulated and accessed. Decisions about how data is processed, manipulated or shared aren't made by an individual developer; they're established by the rules and polices of data governance.

The purpose of data governance isn't to limit data access or insert a harsh, unusable level of rigor that interferes with usage. Its premise is simply to ensure that data becomes easier to access, use and share. The rigor introduced by a data governance effort shouldn't be overwhelming or burdensome. While data governance may initially affect developers' productivity (because of the new processes and work activities), the benefits to downstream data constituents and dramatic improvements in productivity should more than counteract the initial impact.

It should be no surprise that a data strategy has to include data governance. It's simply impractical to move forward – without an integrated governance effort – in establishing a plan and road map to address all the ways you capture, store, manage and use information. Data governance provides the necessary rigor over the data content as changes occur to the technology, processing and methodology areas associated with the data strategy effort.

Defining a Data Strategy Is Key

Nearly every new application or report requires access to other corporate information. And in most instances, the only practical method for developers to determine the existence of that data and identify the best potential source is through conversations, meetings and tribal knowledge. But as source applications increase and cloud-based applications grow, the resulting number of systems creating data has expanded way beyond the knowledge of any individual. There are simply too many systems, sources and data for anyone to track and manage it all. Use of the company's data assets shouldn't rely on word of mouth or tribal knowledge.

While most companies have invested millions of dollars to improve data management, most activities are point solutions addressing individual problems and issues. Few people are aware of the impact a single investment may have in strengthening or (unfortunately) weakening other projects or data initiatives. The challenge most organizations have is realizing that data access and usage stretch across every organization and skill level at their company.

The risk of investing in a point solution is that its focused nature prevents it from addressing issues that cross organizational and project boundaries – and data issues by nature are not specific to a single application or organization. Efforts to deliver new data and/or analytics to a business won't succeed unless all of the other data-related components have been addressed: identify, store, provision, integrate and govern.

The Power of a Data Strategy

Returning to our banking story: Once we had reviewed the different data strategy components with the bank executive, he began to realize that many of the projects under his direction weren't aligned to share and grow the company's data assets. He acknowledged that while there was considerable project rigor for systems and application activities, data hadn't received the level of attention that we were describing. His company offered little in the way of project methods, or even tooling, that supported data sharing and reuse. He was interested in moving forward, but wanted to make sure his efforts wouldn't be perceived as blue-sky activities. He wanted realistic goals with measurable deliverables. He explained:

> "We struggle with new strategic initiatives. They often fail because the goals are too high level and success is never well-defined, or they become 'boil the ocean' programs that become too costly or complex. Data is a big issue at this company. How do we move forward without making the same mistakes of the past? How do you undertake a new strategic initiative as a small, value-based endeavor?"

The strength of the data strategy components is that they help you identify focused, tangible goals within each individual discipline area. Every company has a unique combination of skills and a different set of strengths and weaknesses. Moving forward with a data strategy starts with identifying the strengths and weaknesses that exist within your data environment (within each component area) – and identifying an achievable and measurable set of goals that will improve data access and sharing. The components' purpose is not to identify every potential activity within a data strategy; the components offer visibility into the different disciplines that contribute to a data strategy.

A data strategy initiative isn't a once-and-done effort; by its very nature, a strategy is a long-term set of goals. It's common to identify a multiyear set of goals and identify a shorter-term set of delivery milestones (e.g., quarterly or yearly). This allows the strategy to undergo review and measurement on an ongoing basis to prevent the types of challenges the bank executive mentioned. The components provide a means of categorizing activities and identifying shorter-term deliverables.

Most companies have already invested in data management activities across the different component areas; unfortunately, the different areas are not typically coordinated or aligned with one another. The bank's data management challenges illustrate how the lack of a data strategy (and aligned activities) can cause significant tribulations for data access and usage. A data strategy gives visibility into the relationship each of the components (or disciplines) have with one another. If you don't coordinate the different component activities, you risk delivering a series of point solutions that can't work together.

The idea behind a data strategy isn't to build a perfect world that can address any unforeseen data need. The power of a data strategy is that it positions you to deliver the best possible solution as your organization's needs grow and evolve. When new requirements arise and gaps become visible, the component framework provides a method for identifying the changes needed across your company's various data management capability and technology areas. Your data strategy is a road map and means for addressing both existing and future data management needs.

Learn More

Find out how SAS® Data Management can help you build a successful data strategy by visiting SAS Data Management Consulting.

To contact your local SAS office, please visit: sas.com/offices

108109_S149228.0316