Managing Data in the Cloud

By: [Maria Villar](#)

Your company's move to a "cloud-first" IT environment presents both challenges and opportunities. To succeed, chief data officers (CDOs) will need to reevaluate many aspects of their data strategy to simplify and refine existing data processes and practices.

Why? Managing data brings an opportunity to simplify and reengineer existing processes and practices. Successfully moving to a cloud data platform requires a plan for mapping, transforming, and cleaning your data, both at rest and in motion.

Managing master data in the cloud

Some companies choose to run one or more business software applications or business processes from a cloud software solution provider, such as software as a service (SaaS), platform as a service (PaaS), or business process as a service (BPaaS). Popular examples include CRM, HR applications, and supply chain applications.

Cloud software solutions involve standard business processes and infrastructure. Whether to use an underlying cloud infrastructure that is private, hybrid, or public is a decision the CIO team must make. Performance and responsiveness will be key in this decision because delighting business users can quickly fail if they have to wait. And wait.

Some CIOs choose to retain some systems on premise while moving workloads to the cloud for responsiveness and performance.

Planning your enterprise data strategy

1. Organization and governance
Scope and strategy
Metrics and standards
Accountable owners
Executive sponsors

2. Ongoing data maintenance
Always on, proactive data quality
Low-cost centers with service-level agreements
Automated workflows and data quality

Capabilities

3. Data process
Create, update, delete
Simple, automated
Intelligent, friendly
Following data quality standards

4. Tools and systems
Tools for all capabilities
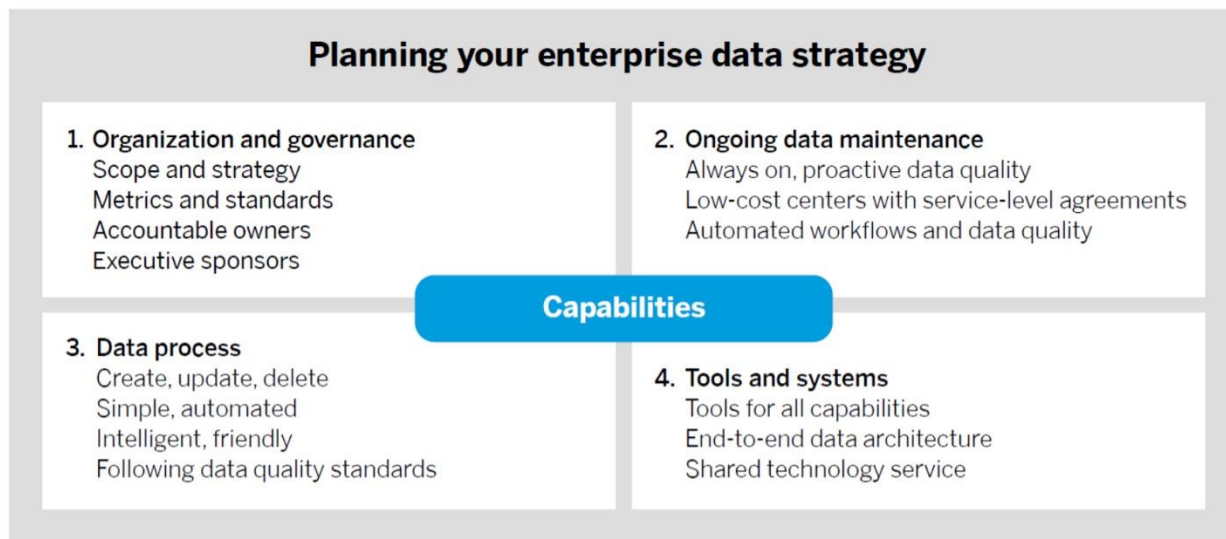End-to-end data architecture
Shared technology service

Figure 1: Planning your enterprise data strategy

Decisions about where that data lives and how it is replicated or federated have a profound impact on managing master data. For that reason, CDOs should be involved in the analysis and resulting decisions.

When business processes are run in a cloud software solution, master data will be created and updated there. The solution also generates new analytical insights about the master data. The CDO should evaluate this strategy and capabilities. CDOs must decide which critical master data fields will be managed (and where they are created, maintained, and replicated) as a significant part of their strategy, and these decisions must be made early in their cloud software solution deployment.

Will the cloud software application's data repository be the authoritative source? Or will you maintain a separate authoritative master data system? How will the two stay in sync? If similar data fields are maintained in both the cloud software application and the trusted source, which is to be the leading system for creating and updating the field

Data quality management

Typically, the primary responsibility and operation for data quality management will still be with the CDO's data management team and not the cloud software solution provider. To assist the data quality efforts, the cloud software solution providers may provide data quality management services or add-ons.

A data quality checklist includes the following tasks.

Define

Define the accountable roles, responsibilities, and metrics for ongoing data quality maintenance between the cloud software solution provider and the company's data management team.

Establish

Establish data archive schedules and rules. While cloud storage is economical, accumulating too much data is risky. Therefore, establish an archive-and-delete schedule to be performed automatically by the cloud provider.

Define or reuse

Define or reuse other data governance rules, standards, and policies for the master data fields that reside in the cloud.

Evaluate

Evaluate the data quality services provided by the cloud software solution provider.

Integrate

Integrate the results in the overall CDO data quality dashboard if you use cloud software data quality monitoring or reporting services.

Data processes and lifecycle management

When deciding which cloud software or Master Data Management (MDM) solution and repository to use as the leading system and trusted source, you will need to consider which data lifecycle processes will be executed in the cloud. Most cloud software solutions can execute these tasks as part of their standard functionality, but it will be up to the CDO team to decide if that capability is used. If you do not use the out-of-the-box cloud capability, then a central create-and-update data process and software application (such as an MDM system) must be developed by the CDO.

Important steps to this end are to:

Decide

Decide which master data lifecycle tasks will be executed in the cloud software solution, such as create, update, delete, and archive.

Automate

Automate the field entry process to speed data entry and maintain consistent data quality. Use known, trusted sources or look-ups where possible.

Determine

For those master data processes not executed in the cloud, determine how the cloud master data process will be disabled to make sure it is not inadvertently used.

Establish rules

Create or reuse business rules and standards for creating, updating, and archiving master data fields in the cloud software solutions.

Ensure accuracy

Ensure data is current and accurate in cases where customers, suppliers, and employees can maintain their own contact data in the cloud. Self-service does not guarantee high data quality accuracy or timely updates.

Data liability and risks

While cloud software environments often have stronger security and data protection than what is found on premise, there will still be risks the CDO must evaluate. In addition to defining the data lifecycle tasks to be performed by the cloud software provider, the CDO must also understand and clearly document the liability risk for data exposure and data risk mitigation obligations including:

- Privacy compliance responsibility

- Data physical location assessment

- Security and data transfer risk management (encryption and anonymization)

- A contingency plan to switch vendors if needed

- Identity management

- Data ethics

Managing cloud data warehouses

In the cloud data management scenario, a data repository is hosted in a cloud infrastructure environment, such as a data lake or a data warehouse as a service (DWaaS). New, fast cloud warehouses are rising in popularity and are replacing in-house warehouses for "fast path" business-led initiatives. A modern cloud data storage architecture offers both technical and business benefits. From a technical perspective, the cloud offers elasticity, efficiency, addressability, and query optimization. These translate into business value.

Improvements in total cost of ownership (TCO)

DWaaS pricing charges only for resources used. This means no forecasting or paying for long-term needs or up-front costs.

Increased flexibility and scalability

With the cloud, an individual department can acquire nearly unlimited computing power and storage with just a few clicks.

Integrating new technologies

Machine learning, artificial intelligence, and analysis tools can easily be integrated.

Data quality management

There are no inherent data quality benefits with a cloud data warehouse versus on-premise solutions; the CDO still must proactively manage data quality. However, the cloud infrastructure and cloud solution provider can execute data quality tasks, as defined by the CDO team, often at scale with automation and data services.

Tasks include:

Develop

Develop an automated, proactive archive-and-delete calendar and rules for "end of purpose" checks.

Create

Create data quality rules and tasks for duplicate matching, elimination, and best-record consolidation.

Automate

Automation and real-time testing are even more vital. Most cloud providers offer data quality management add-on services or tools.

Provide feedback

Provide feedback to third-party data providers on quality issues.

Analytics

CDOs should consider the advantages of analytics in the cloud. Benefits include speed of innovation and operation, greater cost savings, secure access to strategic insights, real-time analysis of data regardless of its location (on premise or in the cloud), increased mobility, and the potential for more accurate and timely forecasting. These make cloud analytics a high-impact driver of innovation and value.

Additional cloud data warehouse capabilities

Machine learning **Artificial intelligence** **Analytical tools**

Figure 2: Additional capabilities

Cloud analytics enable more insight-driven decision-making but also become more complex. Before you implement cloud analytics, you should first address some key questions. Executives, managers, analysts, and front-line workers all make decisions. Some decisions are strategic, some are operational or tactical. How do you support the diverse decision-making needs of your entire organization? Decision-making is a multi-faceted process involving many steps such as making an incremental

investment, launching a new product, or hiring a new employee. How do you ensure that the complete process is supported with no disconnect between all these activities?

Cloud analytics is the most effective way to access cloud data. But what if much of your data remains on premise? What hybrid architecture will best ensure timely access to all your data?

Data liability and risks

Certainly, data liability and risks exist in the cloud, similar to on-premise data. Often the data protection offered by your cloud provider will be superior to what exists in the internal infrastructure.

The CDO team, in partnership with the IT security team and chief privacy officer, should review the cloud provider's encryption, masking, and anonymization methods to ensure they comply with your company's security and privacy standards. This is especially important for protecting personally identifiable information (PII) and data in motion, either from IoT devices or real-time data streams. CDOs must understand how the data will be used by the business community. Remember that one of the fast-growing risk types is "data ethics" risks caused by bias in the data and the analytic algorithms. Best-of-breed organizations implement data ethics guidelines and develop an operational framework to manage these risks.

Managing third-party data providers in the cloud

In managing your third-party data sources in the cloud, what should you consider?

Acquiring, storing, and using third-party data sources have been requirements for CDOs since the 1980s, especially in the financial services and telecommunication industries.

What have we learned over time?

- Finding suitable third-party data providers is time-consuming and often requires using unconventional (and perhaps outdated) sources.

- Often the same or similar data is purchased separately by different business units.

- The process to integrate third-party data into the company's internal infrastructure is often lengthy and complicated, not to mention ensuring the data is used legally across the corporate ecosystem.

- Third-party metadata is often missing or incomplete, requiring internal data teams to perform metadata remediation manually.

Yet, there is an ever-growing need for more and different third-party data. This is driven by new technology, business trends, and the emerging need of your company to predict when to expect "the next normal," for example, a post-pandemic environment.

The growing trend is for third-party data to be provided as add-on features to SaaS, PaaS, and BPaaS cloud solutions, such as the AWS Data Exchange, a data exchange network, or a data marketplace solution. Also growing are data marketplace solutions such as Datarade, which offers buyers thousands of data source options, as well as data-sharing exchanges such as CDQ, which offers data-sharing and collaboration services.

An important best practice for managing third party data is to establish a Head of Acquisition role in the CDO organization. This leader establishes the process for sourcing, acquisition, and use of third-party data in the firm working in partnership with the institution's procurement organization.