# Developing a data integration and lifecycle management strategy for a hybrid environment

ENTER »

# Table of contents

# Introduction: Imposing order on chaos



Cloud-based data presents a wealth of potential information for organizations seeking to build and maintain competitive advantage in their industries. However, as discussed in "The truth about information governance and the cloud," most organizations will be challenged to reconcile their legacy on-premises data with new third-party cloud-based data. It is within these "hybrid" environments that people will look for insights to make critical decisions.

Hybrid environments generally grow without much advance planning, making the task of managing ever-growing data stores all the more difficult. Additionally, scalable data platforms such as Hadoop offer unparalleled cost benefits and analytical opportunities. However, while Hadoop and Hadoop-based solutions have their advantages when it comes to addressing big data volumes, Hadoop itself is not designed for data integration. Data integration carries its own unique requirements (such as supporting governance, metadata management, data quality and flexible data delivery styles) for success. Yet, there is a way to make sense of the chaos. As always, the first step is understanding the nature of the problem. The primary focus must be on the data itself, rather than the sources of the data and the systems used to manage the data. If you make data and ownership of information derived from the data the top priority, everything else falls into place quickly.

## The four pillars

How can your organization realize the financial benefits of the cloud while ensuring information culled from cloud sources is secure and trustworthy? The answer is governance.

Good hybrid information governance rests on four key priorities for IT and the business:

1. **Broad agreement on what information means,** including metadata on common policies and plain-language rules for the information the business needs and how it will be handled.
2. **Clear agreement on how owned-information assets will be maintained and monitored**—for example, operational data quality rules to master data management in on-premises systems.
3. **Enterprise and departmental-standard practices for securing and protecting strategic information assets,** such as articulating role-based access to information, creating rules governing how information is shared and protecting sensitive information from third parties.
4. **An enterprise data integration strategy** that includes lifecycle management, architecting how data will flow and be assembled into strategic information, and also understanding how that information will be maintained over time.

These components form the foundations of information governance in a hybrid environment. In each case, you need a blend of process, organizational and technical enablers for success. With these pillars in place, your organization will have the flexibility to move with speed and confidence.

**This e-book focuses on the fourth pillar: building and executing a data integration and lifecycle management strategy.**

### Taking ownership of strategic information

Adopting a hybrid environment does not imply you must have your IT strategy completely worked out. In fact, cloud-based aspects of the environment will evolve rapidly in response to business priorities. However, even if only a small percentage of data is flowing in from cloud-based sources, IT does need a plan for data integration and security. IT must help the organization ensure it owns the information created from all data and processing, no matter where that information is located. The hybrid infrastructure and decentralized computing are means to the ultimate end of creating strategic information assets.

# Turn data into information: Defining an integration and lifecycle strategy

You've recognized the importance of owning data that is strategic to your organization. In turn, this creates the need for a clearly articulated data integration and lifecycle management strategy. The strategy will define which data sources are acceptable, what data elements to own, how the owned data will be stored or maintained, and how long the data will survive within the organization.

To make these decisions, start by addressing three specific questions:

1. **Where is the data that my organization needs?**
2. **How will the data be assembled to create the information my organization needs?**
3. **Do I really need to keep all this data?**

Let's consider each of these questions individually.

## Where is the data that my organization needs?

Business leaders are eager to harness the power of data, regardless of where it resides. However, before setting out into the expansive data world, be aware that as data volumes increase, it becomes exponentially more difficult to ensure that source information is trustworthy and accurate. If this trust issue is not addressed, end users may lose confidence in the insights generated from their data, which can result in a failure to act on opportunities or against threats.

Unfortunately, the sheer volume and complexity of data and data sources means that traditional, manual methods of discovering, governing and correcting information are no longer feasible. You need to implement an information integration and governance (IIG) solution to support diverse data applications, data warehouses and data warehouse augmentation initiatives—regardless of whether the information they rely on is on-premises or in the cloud.

As important or potentially important data sources are uncovered, a good data integration strategy will account for myriad data topologies by integrating information at the point of data creation. **A solid IIG program must include automated discovery, profiling and metadata management for diverse data sets to provide context and enable employees to make informed decisions.**

The implication is clear: it's not enough to simply pull data on-premises and prepare it for analysis. Instead, a combination of on-premises and cloud-based integration practices and technologies (such as platform-as-a-service) is necessary in a hybrid world. You must integrate and govern on-premises data with on-premises technologies, and integrate and govern cloud-based data with cloud-based technologies.

Furthermore, whether the data is on-premises or in the cloud, data integration technologies must be agile enough to accommodate a wide variety of data formats and to integrate seamlessly with different data storage technologies, from data marts to Apache Hadoop systems.

## How will the data be assembled to create the information my organization needs?

If data integration technologies are pushed out to data sources, how do you pull the pieces together to create the information assets your organization needs? Data is streaming from a variety of sources on demand and at high velocity, so performance is key.

The rapidly shifting data must be fed quickly to various applications in the system so business leaders can react to changing market conditions as soon as possible. To successfully handle this data, you need an enterprise-class data integration solution that is:

- **Dynamic** to meet your current and future performance requirements
- **Extendable** and partitioned for fast and easy scalability
- **Integrated** with Hadoop or other big data storage technologies

## A massively scalable foundation

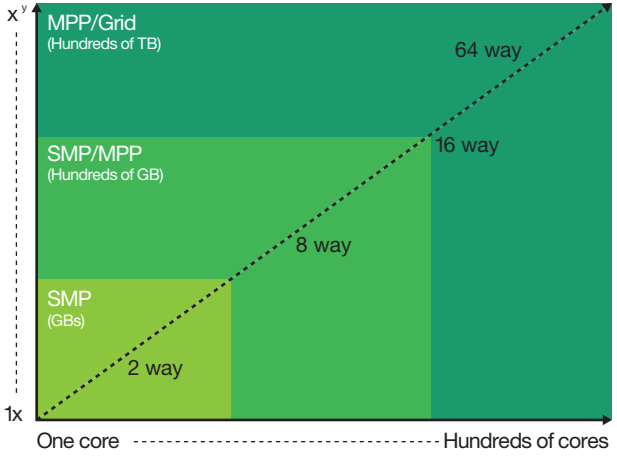
The most critical requirement for processing large hybrid data volumes for data integration is massive data scalability (MDS). MDS provides the ability to process vast quantities of data in parallel, dramatically reducing the amount of time it takes to handle various workloads. Unlike other processing models, MDS systems optimize available hardware resources, allowing the maximum amount of data to be processed per node (see Figure 1).

MDS is important because processing large-scale hybrid data volumes makes it possible to solve many high-value business problems for the first time while ensuring that a hardware platform will yield predictable benefits.

MDS systems have four key things in common:

1. Feature a shared-nothing architecture
2. Are implemented using software dataflow
3. Leverage data partitioning for linear data scalability
4. Use a design isolation environment

**Data scalability across hardware architectures**




- Same functionality
- All architectures
- Linear speedup
- Decision (at runtime) to scale “N way”

*Figure 1.* The essential characteristics needed to support MDS requirements to enable processing of unlimited data volumes.

**Shared-nothing** architected software is designed from the ground up to exploit a shared-nothing, massively parallel processing (MPP) architecture by partitioning data sets across computing nodes and executing a single application with the same application logic executing against each data partition (see Figure 2). This means there is no single point of contention, or processing bottleneck, anywhere in the system. Therefore, there is no upper limitation on data volume, processing throughput, or number of processors and nodes.

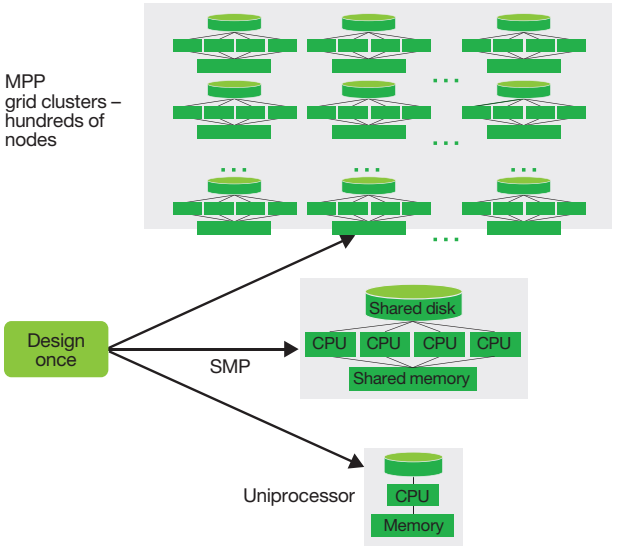



*Figure 2.* An example of a shared-nothing architecture.

**Software dataflow** allows full exploitation of shared-nothing architecture by making it easy to implement and execute data pipelining and data partitioning within a node and across nodes. Software dataflow also hides the complexities of building and tuning parallel applications from users.

Software dataflow is the best architecture for exploiting multi-core processors within a symmetric multiprocessing (SMP) server (application scale-up) and for scaling out to multiple machines (application scale-out)—which will be the more commonly used system architecture in hybrid on-premises/cloud data environments. The architecture:

- Supports pipelined and partitioned parallelism within and across SMP nodes
- Provides a single mechanism for parallelization across all hardware architectures, helping to eliminate complexity
- Reduces the difficulty of building, tuning and executing parallel applications
- Has no upper limit on data volumes, processing throughput and numbers of processing nodes

**Data partitioning** means that data sets are partitioned across separate nodes and a single job executes the same application logic against all partitioned data (see Figure 3). Other approaches, such as task partitioning, cannot deliver linear data scalability as data volumes grow because the amount of data that can be sorted, merged and aggregated is limited to what can be processed on one node. This also means data from various sources can be assigned unique nodes, if necessary or desirable.

Characteristics of systems with data partitioning include:

- Distributes data partitions across nodes
- Executes one job in parallel across nodes
- Enables pipelining and repartitioning between stages and between nodes without landing to disk
- Exploits low-cost grid hardware for big data

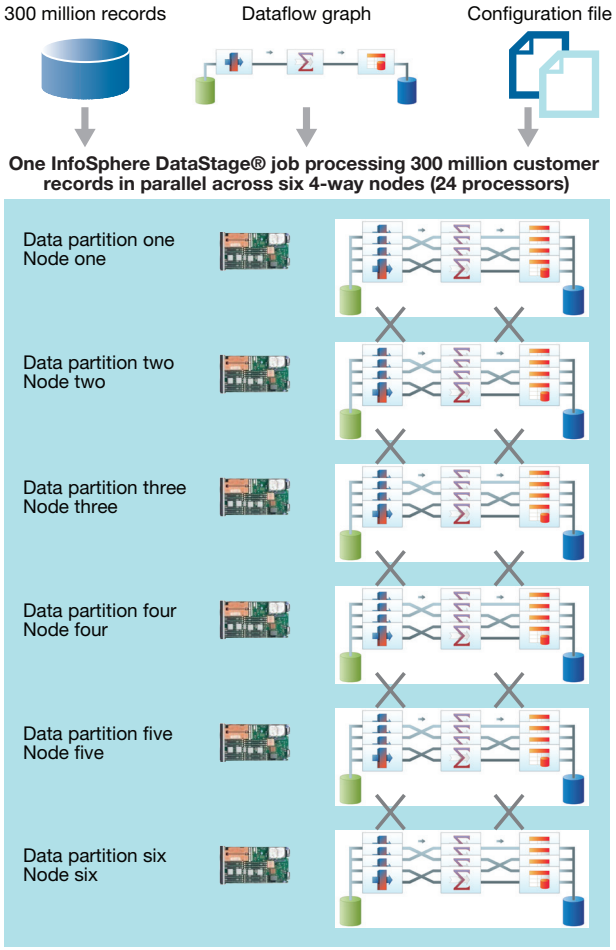



*Figure 3.* Data partitioning architecture.

Finally, design isolation means a developer can design a data processing job once, and use it in any hardware configuration without needing to redesign and retune the job. Characteristics and benefits include:

- Build once and run without modification anywhere
- Reduce complexity with one unified mechanism for parallelizing
- Achieve a clean separation between the development of a job and the expression of parallelization at runtime
- Eliminate the need for performance tuning every time you change hardware architecture
- Add hardware with no data scalability upper limit

All of the most scalable platforms (IBM® Netezza®, IBM PureData™ System, IBM DB2® Database Partitioning Feature, Teradata, Hadoop and IBM InfoSphere® Information Server) have been built from the ground up to support these four characteristics and can seamlessly exploit MPP and commodity grid architectures.

## Leverage data replication for speedy insights

To maximize the amount of insight derived from hybrid data sources, you should employ different data delivery styles depending on the use case. Certain use cases require an up-to-the-minute (or up-to-the-second) view of data to make trusted decisions. In some scenarios, like fraud detection, inventory analysis across channels and real-time operational analytics, basing decisions on data that is a month, a week or even a day old is counterproductive.

You can deliver outstanding customer experiences by collecting and leveraging data faster than the competition and offering greater continuity of services. In recognition of this, data transformation and delivery requirements have broadened to include real-time data transfer based on data replication capabilities, specifically around change data capture.

When it comes to maximizing the performance and scalability of real-time data integration, there are three factors to consider:

1. **The approach used to capture a change at the source or sources:** The most flexible and efficient option for capturing changes at the source is a replication process to capture changes as they're written to a source log. As soon as source data is modified, the mechanism becomes aware of the alteration and forwards the changed data with little to no impact on the source database and application, thereby minimizing the need for large batch windows.
2. **The mechanism used:** Many mechanisms can be used for data replication. When properly implemented, a log-based capture approach often has a lower impact on the source database, resulting in higher overall performance.
3. **Temporary data persistence:** Whether data is temporarily persisted also impacts data replication performance. Ideally, an organization would be able to stream changes without persisting them to increase performance, because data does not need to be written to disk and then accessed by a transformation engine.

Performance isn't the only important factor in data integration solutions, though. Additional considerations include:

- **Solution flexibility:** Flexible mechanisms support multiple platforms, sources and targets, including Hadoop Distributed File System (HDFS) files for large volumes of data. They will also support an equally wide variety of topologies.
- **Impact on existing IT infrastructure and processes:** The preferred data replication solution must be easily integrated into your existing change management processes, by allowing easy automation through scripting or common programming languages like Java.
- **Ease of use:** Huge learning curves have an impact on data integration. Powerful graphical user interfaces that allow easy configuration and monitoring will minimize the time spent deriving real insights from big data.

## Do I really need to keep all this data?

Many analytics users call for a keep-everything approach to data, because you never know what insights might be drawn in the future from data that isn't currently useful. However, this approach can result in significant risk and cost to the business.

Following the "let's keep it in case someone needs it later" mandate, many organizations keep too much historical data, much of which has no value. Opening the doors to excessive storage and retention only exacerbates the situation. At the same time, organizations must ensure the privacy and security of the growing volumes of confidential information. Government and industry regulations require organizations to protect personal information no matter where it lives—even in test and development environments.

Moreover, as users execute queries on hybrid data volumes, slow response times and degraded application performance become major issues. If left unchecked, continued data growth will stretch resources beyond capacity and negatively impact response time for critical queries and reporting processes. These problems can affect production environments and hamper upgrades, migrations and disaster recovery efforts. Implementing intelligent data management of historical, dormant data is essential for avoiding these potentially business-halting issues.

Data growth can drive up infrastructure and operational costs, often consuming most of your data warehousing or data management budget. Rising data volumes require more capacity, and you often must buy additional hardware and spend more money to maintain, monitor and administer the expanding infrastructure. Large data warehouses and complex data environments generally require bigger servers, appliances and testing environments, which can also increase software licensing costs for the database and database tooling—not to mention labor, power and legal costs.

## The role of archiving

The data lifecycle stretches through multiple phases as data is created, used, shared, updated, stored and eventually archived or defensively disposed. Data lifecycle management plays a particularly important role in the archiving phase.

As part of an archiving strategy, retention policies are designed to keep important data elements for reference and for future use while deleting data that is no longer necessary to support an organization's business or legal needs. Effective data lifecycle management includes the intelligence not only to archive data in its full context, which may include information across dozens of data sources, but also to archive it based on specific parameters or business rules, such as its age. Data lifecycle management can also help storage administrators develop a tiered and automated storage strategy to archive dormant data in a data warehouse, thereby improving overall analytical application performance.

Effective data lifecycle management benefits both IT and business stakeholders by:

- **Increasing margin:** Lowered infrastructure and capital costs, improved productivity and reduced application defects during the development lifecycle
- **Reducing risks:** Less application downtime, minimized service and performance disruptions, and adherence to data retention requirements
- **Promoting business agility:** Improved time to market, increased application performance and better application quality

# IBM solutions for data integration and lifecycle management

IBM has been providing solutions that can handle enterprise-class data from multiple sources for decades. The company has long led the way with data integration, management, security and analytics solutions that are known for their reliability, flexibility and scalability.

IBM InfoSphere Information Server is a market-leading information integration platform that helps organizations understand and govern data, create and maintain data quality, and transform and deliver data. Along with InfoSphere Data Replication and InfoSphere Federation Server, InfoSphere Information Server delivers powerful IIG capabilities:

- **Data integration:** Transforms data in any style and delivers it to any system, ensuring faster time to value and reduced risk for IT. This package also includes the InfoSphere Data Click feature, which supports self-service data integration.
- **Data replication:** Supports real-time data replication requirements, and enriches mobile applications, business analytics and big data projects by integrating replicated data.
- **Information governance catalog:** Helps you understand and govern your information, encouraging a standardized approach to discovering IT assets and defining a common business language so you are better able to align business and IT goals.
- **Data quality:** Enables you to analyze, cleanse, monitor and manage data, adding significant value by helping you make better business decisions and improve business process execution.
- **Data federation:** Flexibly delivers data and supports virtual data hub requirements.

InfoSphere IIG capabilities also support the IBM big data platform, which includes tools for visualization and discovery, Hadoop-based analytics, stream computing, data warehousing and text analytics.

IBM InfoSphere Optim™ solutions supplement IIG capabilities with data lifecycle management capabilities that scale to meet enterprise needs. Whether you implement InfoSphere Optim solutions for a single application, data warehouse or any complex data environment, you get a consistent strategy for streamlining data lifecycle management. The unique relationship engine in InfoSphere Optim provides a single point of control to guide data processing activities such as archiving, subsetting and retrieving data.

InfoSphere Optim solutions help you meet IIG and governance requirements as well as address challenges exacerbated by multi-source hybrid data environments. By archiving old data from huge data warehouse environments, you can improve response times and reduce costs by reclaiming valuable storage capacity.

# Next steps: Continuing the cloud governance discussion

Cloud-based data and processing services present too much opportunity for business users to ignore, and IT is charged with maintaining the integrity of internal, on-premises transactional and reporting systems. Charting a governance strategy for a hybrid environment is not something to consider at a future date. It needs to happen now.

This e-book discusses one of the four pillars for successful hybrid environment governance: data integration and lifecycle management in a hybrid environment. **To explore the other pillars, download one or all of the e-books in this series:**

- The truth about information governance and the cloud
- Make sense of your data
- Prepare and maintain your data
- Securing data in the cloud and on the ground

For additional information on IBM governance thought leadership and supporting technologies, visit: **ibm.com**/analytics/us/en/technology/agile/

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing



Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
September 2016



Please Recycle



IMM14177-USEN-01