Wrangling big data: Fundamentals of data lifecycle management

How to maintain data integrity across production and archived data
Introduction

Quality data, quality results

Managing the data lifecycle

Benefits across the enterprise

Evaluating data lifecycle management solutions

Resources
Introduction: Big data is a big deal

Everywhere you turn in enterprise IT, you hear the buzz about big data: billboards, commercials and likely your own conference rooms and inboxes. The variety, volume and velocity of data streaming through enterprise systems is on the rise, and so is the amount of discussion about the best way to handle it. **But while the big data phenomenon is giving organizations a broad range of new options for data analysis, it also compounds the business challenges associated with collecting, managing, organizing and protecting data.** Successfully leveraging big data for analytics demands that companies develop strategies to reduce the cost of managing data and reduce the risk involved in organizing and protecting that data.

In addition to these business challenges, enterprises face technical challenges in managing rapidly expanding volumes of all types of data. The sheer number of continually proliferating data sources introduces complexity into data lifecycle management processes. If IT departments cannot manage information appropriately—from the moment it is created to the point when it can be archived or defensibly disposed—they risk violating legal and regulatory compliance requirements.

To efficiently manage data throughout its entire lifecycle, IT leaders must keep three objectives in mind:

1. Data veracity (trustworthiness) is critical for both analytics and regulatory compliance.
2. Both structured and unstructured data must be managed effectively.
3. Data privacy and security must be protected at all times.

**This e-book will explore how these objectives can help you corral big data into a manageable, dependable—and eminently valuable—enterprise resource.**
Quality data, quality results

It’s obvious but important: the better the data, the better the results. When the New Intelligent Enterprise—a joint partnership between the MIT Sloan Management Review and the IBM Institute of Business Value—conducted a survey on analytics, it found that organizations that used analytics for competitive advantage were 2.2 times more likely to substantially outperform their industry peers. Ensuring the accuracy of the data used for analytics is a key factor in enhancing overall organizational productivity.

But achieving that accuracy requires a strong commitment to data lifecycle management—the process of managing business information throughout its existence, regardless of where that information is stored. That means ensuring that large data volumes do not inhibit application performance in production environments, and acquiring the ability to both offload the data into archived systems and maintain accessibility to that archived data for retention and business continuity.
The emergence of big data only reinforces the need for effective data lifecycle management. Big data is more than just information stored in an Apache Hadoop-based framework; it is also the structured data within data warehouses, databases and standalone applications. The size of the data source—whether a single database, an entire data warehouse or a Hadoop framework—doesn’t matter. In the world of big data, all data sources are crucial and must be managed throughout their lifecycles.

Data lifecycle management principles come into play whenever data is moved or changed, such as:

- When new databases and data sources are first created
- Any time governance stewards need to track audits, assess compliance or protect personally identifiable data
- When line-of-business owners need to conduct real-time analysis
- When data is archived, restored or defensibly disposed

But for data lifecycle management to be most effective, it must be addressed when data sources are first created, at the beginning of any deployment, to maintain proper structure as new data is introduced. By applying these basic governance techniques, it becomes easier to add and accommodate new information in a data deluge.
Managing the data lifecycle

The data lifecycle stretches through multiple phases as it is created, used, shared, updated, stored and eventually archived or defensively disposed. Three elements of data lifecycle management play an especially key role in several phases of data’s existence:

1. **Test data management**: During the process of developing new data sources, whether databases or data warehouses, test technicians must automate the creation of realistic, rightsized data sources that mirror the behaviors of existing production databases. To ensure that queries can be run easily and accurately, they must create a subset of actual production data and reproduce actual conditions to help identify defects or problems as early as possible in the testing cycle.

But actual production data contains information that may identify customers—a clear violation of privacy policies. That’s why organizations must mask sensitive information in test environments for compliance and privacy. Applying data masking techniques to the test data means testers use realistic-looking but fictional data; no actual sensitive data is revealed. Leveraging test data management also means that technicians can easily access and refresh test data, speeding the testing and delivery of the new data source.

The entire data lifecycle (shown as the grey circle) benefits from good governance, but management capabilities that focus on the use, share and archive steps have wide-ranging benefits for cost reduction and efficiency gains.
2. Data masking and privacy: The ability to mask data and protect privacy has implications beyond test data, extending into production data. Whether information is stored in Hadoop or a data warehouse, certain facets can be masked from unauthorized users, even though they may be authorized to see the data in aggregate.

For example, a pharmaceutical company that is submitting drug-testing results to the U.S. Food and Drug Administration may mask social security numbers and dates of birth, but not patients’ ages and other demographic information. Masking certain data this way satisfies corporate and industry regulations by removing identifiable information, while still maintaining business context and referential integrity in production environments.

Data lifecycle management: An industry view

Not sure where data lifecycle management fits into your organization? Here is a sample of data-driven applications in various industries that could benefit from more streamlined, accurate data:

- **Energy and utilities**
  - Metering
  - Utility rates and billing

- **Healthcare**
  - Electronic medical records
  - Insurance claims and payment systems

- **Financial services**
  - Mergers and acquisitions
  - Customer service actions and preferences

- **Telecommunications**
  - Call-data records
  - Customer relationship management systems

- **Transportation**
  - Online travel booking and ticket information

- **Insurance**
  - Claims processing and billing
  - Pension and benefits accrual

- **Retail**
  - Purchase records and customer profiles
  - Sales, distribution and inventory management
3. **Archiving**: Effective data lifecycle management includes the intelligence not only to archive data, but archive it based on specific parameters or business rules, such as the age of the data. It can also help storage administrators develop a tiered and automated storage strategy to archive dormant data in a data warehouse, thereby improving overall warehouse performance.

Intelligent data archiving supports data retention needs based on such factors as the age of the data, any legal hold requirements within data sources and defensible disposal of data. In addition, effective data lifecycle management enables IT to query archived data in Hadoop-based systems, providing flexible access to that data. The effect is to reduce the total cost of ownership (TCO) of data sources by intelligently archiving and compressing historical data, while ensuring that data is available for applicable analytical purposes.
Effective data lifecycle management benefits both IT and business stakeholders. Because data lifecycle management solutions provide a consistent view into data sources from a variety of angles, enterprises gain usability and training advantages.

**Test data management**
The ability to create realistic, rightsized test data helps ensure reliable, trustworthy test results that reflect the way an application will actually perform with live data. All teams have access to accurate, appropriately protected data for their work. Developers can test multiple application functionalities and setups, confident that they can quickly create custom data sets for any target scenario. QA staff can validate performance and integrations based on specific test cases. Collaboration is streamlined by the sharing of data sets and reports based on repeatable test data, increasing the opportunity for predictable results.

**Data masking**
Masking data in both development and production environments gives compliance officers and auditors higher levels of certainty regarding security and protection of personally identifiable information and other sensitive data. Customer service staff can see detailed purchase records without being given access to sensitive credit card details. Medical billing teams can access insurance and payment details but be prevented from seeing treatment plans or doctor’s notes. Developers can pull subsets of address information to test location-based apps without seeing the actual names or other identifying characteristics associated with those records.

**Archiving**
Disciplined data archiving benefits any department that stores data—which means almost all of them, from marketing to IT to executive management. HR teams can securely move old employee records into long-term “cold storage,” and save more time by enabling automatic removal and disposal of certain items as they pass retention deadlines. Well-maintained archives help the legal department quickly locate and retrieve items for auditing purposes. Marketing can include archived data in their analytic efforts, supporting strategic decisions with long-term historical evidence and trend evaluation.
Evaluating data lifecycle management solutions

The right data lifecycle management solution for your company will depend greatly on the mix of data types and access requirements in your organization. When evaluating solutions and contemplating deployment strategies, seriously consider these three capabilities:

1. **Compatible and scalable**: The system should support heterogeneous environments across multiple databases, data warehouse environments, systems and platforms, as well as have the capacity to handle big data across production and testing environments. If it works with only some of your infrastructure and some of your data, you’ll get only some of the possible benefits.

2. **Understands the complete data relationship**: Data loses value when it loses context. The most effective data lifecycle management platforms ensure referential integrity when masking data, accurately subset data for test data management and support independent access to archived data.

3. **Maintains accessibility**: Managing and preserving data access based on retention requirements as well as system and employee needs is a vital component of extending data lifecycle management benefits across an enterprise. Leading solutions have the flexibility to create intelligent, targeted policies for data access and archiving that help organizations improve application performance, reduce costs and minimize risks.
Delivering comprehensive data lifecycle management capabilities
With IBM® InfoSphere® Optim™ solutions, enterprises can easily govern data throughout its lifecycle by creating and managing realistic test data, masking data in both development and production, and automating data archiving policy enforcement. InfoSphere Optim software is compatible with multiple databases, data warehouse environments and enterprise applications. Its ability to support multiple access methods to archived data makes it a valuable solution in supporting data retention requirements, as well as supporting the business-critical analytical needs of Hadoop-based systems.

By managing the complete data lifecycle from a single platform, enterprises can make sure their systems and staff members have the highest-quality information for all tasks and decisions. At the same time, developers and IT can secure and protect sensitive data to minimize risk and exposure, both in testing and other data delivery scenarios. And because InfoSphere Optim works with a variety of data sources, its value to the enterprise scales as organizational data needs grow.

InfoSphere Optim in action
Toshiba TEC Europe, a leading subsidiary of the Toshiba Group, needed to proactively manage application data growth to support business expansion and deployment of Oracle E-Business Suite across business units. It implemented InfoSphere Optim to archive historical transactions and reduced batch processing time for 19,000 jobs from 250 hours to only 65 hours—a 75 percent increase in application availability. Plus, it reduced database size by 30 percent, enabling it to manage continued data growth, yet retain access to current and historical transactions. As a result, Toshiba TEC was able to satisfy business unit requirements and support continued business growth.
Resource

To learn more about data lifecycle management for big data and the capabilities available in InfoSphere Optim solutions, please contact your IBM representative or visit:

- Manage the data lifecycle of big data environments
- Solution brief: IBM InfoSphere Optim solutions for data warehouses
- IBM InfoSphere Optim Data Growth Solution demo
- IBM InfoSphere Optim Test Data Management Solution demo