research

CHECKLIST REPORT 2020

Data Onboarding for Cloud Data Lakes

Five key considerations for data ingestion to avoid a data swamp

By David Loshin

Sponsored by:

Infoworks

tdwi | TRANSFORMING DATA WITH INTELLIGENCE™

JULY 2020

TDWI CHECKLIST REPORT

Data Onboarding for Cloud Data Lakes

Five key considerations for data ingestion to avoid a data swamp

By David Loshin

TABLE OF CONTENTS

tdwi

Transforming Data With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

Congratulations! Your senior management has approved the plan to migrate to the cloud, and you are now tasked with making it happen. The objectives are clear: reduce reliance on large-scale capital acquisitions of on-premises hardware and software, migrate both the applications and the data to cloud-based platforms, and make sure that the business runs continuously without interruption while executing the migration. Perhaps that might prove to be easier said than done.

The process of cloud migration and modernization is appealing at the CIO or CDO level, but executing this type of digital transformation requires attention to many discrete details at the operational level. Whether you are thinking about the details of data migration, the life cycle associated with application operation and its dependence on data, or even the proliferation of a wider variety of data source types, the process of data onboarding itself poses some degree of complexity.

This checklist explains five ways to support data onboarding and simplify cloud data migration and modernization.

Growing data volumes will overburden manual attempts at data ingestion, so plan for data onboarding that encompasses the full life cycle of data ingestion, synchronization, pipeline orchestration, and governance. Data awareness is critical to proper planning, and we suggest crawling the data to accumulate intelligence about the data landscape. Identifying data owners and engaging them in a dialog about roles and responsibilities will clarify accountability for data usability. Finally, simplify the operational processes by employing products and technologies that leverage data intelligence for automating data ingestion and data synchronization.

Together, these recommendations will not only reduce the complexity of data ingestion, they can also establish good data governance that raises the quality and utility of enterprise data.

1 PLAN DATA ONBOARDING, NOT JUST DATA INGESTION

Because much of the practical experience in preparing acquired data for reporting and analysis is associated with the monolithic data warehouse, a common misperception is that the methods for data extraction, transformation, ingestion, and loading can be easily adapted to a modernized reporting and analytics environment in the cloud.

Although data ingestion may seem relevant for a cloud-based environment, the cloud's flexibility expands the breadth of data sources consumable by numerous simultaneously executing reporting and analytics applications. Cloud-based applications may consume flat files that are migrated to cloud storage, but they might also consume a variety of data sets from a virtual data lake, including streaming data sources, cloud-based database management systems, or virtualized data assets accessed via data federation.

That means that not only are there many types of data assets that can be acquired, there are also numerous ways that acquired data assets will be consumed. Traditional approaches to data ingestion are generally limited to data loading, data transformation, and then storage—either in a data warehouse or some other repository. Data onboarding is different: it is a holistic process encompassing traditional data ingestion along with automated processes for ensuring data asset utility for downstream data consumers across the complete cloud data management life cycle.

Data onboarding combines ingestion with continual data synchronization, lineage management, data pipeline orchestration, and integrated operational stewardship and governance—capabilities that the organization may need to become skilled with rapidly.

Therefore, assemble a plan for data onboarding by formalizing practical steps for assessment and integration:

- Inventory the applications that consume data in the cloud
- Devise a process for assessing the applications' specific data requirements (desired data assets, update frequency, data validation expectations, etc.)
- For each application, maintain a list of desired data assets
- Determine the types of data assets and transformations to be applied
- Document the consumers' expectations for data quality
- Document the consumers' expectations for synchronization
- Employ the right technologies to manage ingestion, synchronization, and governance

2 CRAWL THE DATA

One objective of onboarding data to the cloud is increasing data democracy—simplifying data access for a broader array of data consumers by enabling them to search for, isolate, and select the right data assets for their respective use cases. However, even in a governed environment, the organization's data landscape is distributed and diverse.

It is rare that an organization has mapped that landscape, let alone catalogued, classified, and organized the data assets that exist across its environment. Wholesale data migration to the cloud not only risks creating a "data dump," it also makes the process of enabling data consumers more complex.

The best way to address this is to survey the existing data landscape as part of your data onboarding process. Automated data crawling scans the data landscape to discover knowledge about the data assets, profiles the data to infer metadata, infers the structure and content of the data assets, and documents and catalogs the discovered data intelligence. Crawling the data will inform data onboarding in several ways:

- **STRUCTURE ASSESSMENT** attempts to categorize whether the data asset is structured, semistructured, or unstructured
- **PROFILING** scans values in a structured or semistructured data asset and performs statistical analysis of the values associated with each data attribute
- **METADATA DISCOVERY** infers the data type and data size, as well as data element concepts and names (in cases where there is existing knowledge about named data elements and their corresponding values)
- **DATA PATTERN IDENTIFICATION** infers a semantic data type based on collections of patterns (such as telephone numbers or Social Security numbers)
- **SCHEMA DISCOVERY** allows the data stewards to determine the schema structures of data assets
- **DATA CLASSIFICATION** sheds light on the content of the data asset and the degrees of sensitivity requiring special attention and protection
- **CATALOGING** collects the accumulated data intelligence into a searchable data catalog

Data crawling provides insight into the data landscape to inform the data architects and data stewards managing onboarding across the data life cycle.

3 IDENTIFY OWNERS AND ESTABLISH GOVERNANCE

There are at least two different models for transitioning to a cloud data lake.

In one model, the data lake is a repository for selected data sets that have been deliberately produced and curated for the purposes of data sharing. For example, an organization may maintain a customer database associated with sales and marketing applications (either on premises or in the cloud). If there is a desire to publish this customer data so that more data consumers can use that information, a data set can be extracted, validated, and prepared for data onboarding.

A different model transitions the operational data set to the cloud along with its associated applications. In other words, the data set in the data lake becomes the one and only version of the data. In this case, there is no data production process; instead, the data asset is seated within the cloud data lake.

In both these cases, enabling access to the data asset requires delineating core data governance expectations and establishing accountability for the data asset in the data lake, answering questions about:

- **OWNERSHIP:** Who is designated as the owner of the data set?
- **METADATA MANAGEMENT:** Even in an environment relying on automated data crawling and metadata inferencing, what are the data owner's responsibilities for reviewing and validating discovered metadata and addressing gaps in what is known about the data asset?
- **STEWARDSHIP:** What methods are in place to ensure that data consumer expectations (for quality, timeliness, accuracy, completeness, etc.) are being met?
- **COLLABORATION:** In the event that any data consumers find issues with a data asset or have issues that need to be resolved, what framework is in place for communicating issues to the data owner?
- **PERFORMANCE:** What are the service level expectations for addressing reported issues?

Data onboarding informs data governance practices for ensuring the quality and usability of enterprise data assets. By identifying data owners and supplementing that information with discovered knowledge about metadata, data lineage, and ongoing auditing of catalog search and access, an organization can improve its capability to support enterprisewide data accountability.

4 INGEST THE DATA

Prior to operationalizing the process of bringing data into the cloud data lake, people often underestimate the challenges of data ingestion, especially as the number and volume of data assets increase. There are two key challenges to consider when developing a data ingestion plan:

1. Ensuring ingestion performance meets service levels for data availability, utility, and accessibility
2. Synchronizing both the structure and the content of data lake replicas

The first issue reflects the growing number and size of data assets and, correspondingly, a misunderstanding of the network bandwidth and computational power required to scale up to massive data volumes. It would not be unusual today to consider ingesting huge data assets with orders of magnitude more rows and columns than your enterprise has previously handled. An underperforming process will not only create bottlenecks that impede the availability of the ingested data set, it will also increase the computational load that ultimately impacts other systems.

Focus on the performance aspects of data ingestion and the need to ensure that service levels are met. Leverage the increasingly pervasive availability of commodity-based configurable computing resources (either on premises or in the cloud) in two ways.

First, take advantage of distribution, both in storage and processing. Today's massive-scale environments are ingesting data from many distributed sources. Your ingestion should balance accessibility and data streaming from the distributed sources to ensure an effective, timely schedule without negatively impacting operational system performance. Second, use parallelization to accelerate ingestion by spreading the work across a broad array of computing resources.

Attempting to manually program parallel execution for each data asset to be ingested will rapidly overwhelm your development team. Instead, look for tools that will automate parallelization of data ingestion, which will both simplify and speed data ingestion as well as:

- Manage data source connection pools to get the appropriate balance
- Modulate data bandwidth pipelines to reduce the impact of access latency
- Provide fault tolerance, error handling, and process rescheduling
- Take advantage of fast connectors where available

5 SYNCHRONIZE THE DATA

Automated processes for parallelizing data ingestion address performance issues but do not account for two data synchronization challenges: maintaining data freshness without overwhelming the system's performance and automatically accommodating changes to source schemas without disrupting data updates. This implies a need for automated processes for incrementally refreshing data assets.

To support data freshness, a typical approach has been using change data capture (CDC), which monitors database logs to identify and isolate changes that can be propagated automatically to the target data asset. Alternatively, when database logs are not available, one can query the source to determine what has changed since the prior refresh.

Yet these queries don't necessarily tell the whole story—certain modifications (such as deleted records) are not detectable. Ensuring a sound process for producing queries that detect modifications and reconciling the detected changes requires careful oversight and programming, especially as organizations transition from malleable relational database management systems (RDBMS) to immutable big data stores such as HDFS files or cloud-based object stores.

Source data changes are not the only challenge. Developers charged with data synchronization may not be aware of changes to the source schema, but these changes can wreak havoc on the quality and fidelity of the target environment when attempting to integrate data with structural inconsistencies.

Automated data synchronization can address these two fundamental challenges and reduce the burden on the developers and the DataOps teams. It provides a trustworthy process for maintaining data freshness by:

- Combining log-based CDC with query-based CDC depending on what is most effective for each data source
- Continuously surveying the source data landscape and identifying and assessing structural metadata that can be used to detect changing dimensions in the source systems
- Surveying the target for changes and rapidly applying data refreshes
- Maintaining versions/history of incremental updates and continuously monitoring for schema changes and slowly changing dimensions
- Automatically "healing" data pipelines potentially impacted by schema changes

AFTERWORD

Over the past 30 years, the evolution of segregated reporting and analytics systems such as decision support systems, data marts, and data warehouses has created a need for data integration services. However, as data volumes explode to include a broad swath of different types of data sources, the complexity of data ingestion will far outpace practitioners' abilities to manually manage the process.

As more organizations migrate their information environments to the cloud, it is necessary to adopt alternate approaches to overcome the challenges of ingesting, integrating, and synchronizing data along with the necessary orchestration tactics.

Automated data onboarding is a way to overcome these challenges. Surveying the data landscape and crawling the data will provide the needed data intelligence to plan and manage an automated data onboarding framework.

Look for tools that support automated crawling, ingestion, and synchronization to increase enterprisewide data awareness. Automated data onboarding simplifies the data ingestion and synchronization processes, thereby reducing the effort for implementation and shortening time to value.

ABOUT OUR SPONSOR

Infoworks

Infoworks offers a comprehensive and automated enterprise data operations and orchestration (EDO2) system. It is built to automate and accelerate deployment and orchestration of analytics projects at scale in cloud, hybrid, multicloud, and on-premises-based environments. Through deep automation and a code-free environment, Infoworks empowers organizations to rapidly consolidate and organize enterprise data, create analytics workflows, and deploy projects to production within days—dramatically increasing business agility and accelerating time to value. Infoworks counts some of the world's largest financial, retail, technology, healthcare, oil and gas, and manufacturing companies as its customers.

To learn more, please visit infoworks.io.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

ABOUT THE AUTHOR

David Loshin, president of Knowledge Integrity, Inc., (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices, as the author of numerous books and papers on data management, including *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* and *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is a frequent invited speaker at conferences, web seminars, and sponsored web sites and channels including TechTarget and The Bloor Group. David is also the program director for the Master of Information Management program at the University of Maryland's College of Information Studies.

David can be reached at loshin@knowledge-integrity.com.

ABOUT TDWI RESEARCH

TDWI Research provides industry-leading research and advice for data and analytics professionals worldwide. TDWI Research focuses on modern data management, analytics, and data science approaches and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of business and technical challenges surrounding the deployment and use of data and analytics. TDWI Research offers in-depth research reports, commentary, assessments, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.