INFOWORKS WHITE PAPER

ACCELERATING DATA VALIDATION AT SCALE

Enterprises are focused on dramatically increasing and accelerating their ability to leverage all their data assets to drive business value and become more competitive. Leveraging data is a Board-level strategy focused on improving overall operating efficiency, elevating customer experience, and gaining insights to develop new products and solutions. Legacy data platforms do not provide the scale, capability, or agility that these strategies demand, so companies are trying to accelerate the migration of their data platforms to the cloud. But simply lifting and shifting their old data to the cloud only moves the old challenges to a new infrastructure, you must migrate and modernize simultaneously to achieve transformative results. The Infoworks Platform was built from the ground up to provide the automation and scale necessary to successfully migrate and modernize, and to capture the value of the cloud.

Massive volumes of data, metadata and workloads must be successfully migrated, and it is crucial that this be achieved rapidly, without errors. Traditional approaches to data validation – hand-coding, open-source point tools or stitched-together solutions – require too much time, end up being costly, are prone to error, and have proven ineffective at scale.

Infoworks has pioneered a unique approach to data validation: integrated, end-to-end automation that validates data at every hop, at scale, and uses a multi-layered validation approach that accelerates data validation by up to 8-10x over traditional approaches.

This white paper outlines the challenges to data validation, why traditional approaches no longer work, and how Infoworks enables and accelerates data validation automatically.

A RECENT HARVARD BUSINESS REVIEW STUDY FOUND THAT 67% OF ENTERPRISES ARE ACCELERATING THEIR MIGRATION TO THE CLOUD.

ENSURING PETABYTES OF DATA ARE RAPIDLY, AND ACCURATELY MIGRATED IS A MASSIVE UNDERTAKING THAT'S UNACHIEVABLE WITH TRADITIONAL SOLUTIONS.

BY INTEGRATING AND AUTOMATING THE PROCESS END-TO-END, INFOWORKS ACCELERATES DATA VALIDATION UP TO 8-10X, ENABLING BETTER, FASTER, LOWER COST CLOUD MIGRATION.

The Challenges with Data Validation

Enterprises are faced with migrating massive volumes of data from on-premises to the cloud, ensuring the migrated data is identical to the source data; at the same time accommodating time and resource constraints while ensuring business continuity. Legacy approaches are not able to scale to meet the requirements, and introduce critical challenges to success including:

- Significant time and resource requirements
  Validating large data sets when migrating to the cloud requires significant time and specialized resources and must be seamless to avoid being disruptive to the business.
- High computational costs
  Significant compute is required to validate large data sets, increasing computational cost and investment to ensure sufficient horsepower and scale.
- Inability to perform in hybrid environments
  Data validation must seamlessly operate in today's hybrid multi-cloud environments.
- Reliance on manual solutions
  Relying on manual solutions to validate massive volumes of data leaves enterprises prone to error, repetitions, and interrupted data validation processes over the course of migration.

Traditional Approaches to Data Validation

Traditionally, enterprises have relied on hand-coded, point tools, or stitched-together solutions for data validation, however with petabytes of data and 1000's of data tables being migrated to the cloud, new approaches must be considered to accommodate cost, time, resource, and scalability requirements:

- Hand-Coded Solutions
  Specialized resources are required to write scripts for the validation process using different scripting languages (Scala, Python, Java, etc.). While hand-coding is simple it is extremely time and cost-consuming, prone to human error, and inefficient for high volume, very complex processes.
- Open-Source Point Tools
  Various open-source tools (Data Validation Tool, Data-Diff, etc.) can be used but they cannot fulfill the requirements for data validation, not to mention require specialized technical resources for enhancements and integration, and generally do not meet enterprise security thresholds.
- Enterprise Tools
  Enterprise data integration and data validation tools are fit for purpose - generally for select data sources or targets. These tools also require specific infrastructure and must be stitched-together for various steps throughout migration and typically are higher cost to deploy.

The Infoworks Solution

Infoworks delivers automated, end-to-end data validation built on the enterprise capabilities of the Infoworks platform.

Infoworks automatically validates data at every hop, at scale, ensuring data that is migrated is identical to source data. Using a unique layered data validation approach, and applying column-hashing versus row-hashing, data validation is up to 8-10x faster than other approaches. Even the largest volumes of data tables can be validated, and accelerated, with Infoworks APIs.

Data is continuously validated throughout the entirety of migration from Hadoop or Enterprise Data Warehouses, seamlessly:

Infoworks Replicator (www.infoworks.io/infoworks-replicator/), designed for Hadoop migration to the cloud, automatically validates Hadoop data and metadata as its replicating to the cloud. By nature, Hadoop clusters have unique challenges for migration, synchronization, and validation; Infoworks solves these challenges by enabling rapid migration of Hadoop data and metadata to the cloud.

Infoworks Platform (www.infoworks.io/products/platform/) migrates data, metadata, and workloads from EDW's to the cloud, incorporating automated data validation from on-premises to the cloud, or once in the cloud when migrating from one cloud to another.

Infoworks provides an efficient, automated, scalable solution to simplify and accelerate data validation.

Efficient

- Memory-based columnar processing vs core compute row-based processing reduces compute time and costs
- Unified, end-to-end solution vs. multiple point tools
- Single compute platform validation approach reduces redundancy and solution fragmentation

Automated

- Automated data validation eliminates the need to re-migrate data
- Automated historical and incremental data validation
- Automated multi-layered data validation
- Data is continuously validated at every hop: while data is being moved the cloud and once data is in the cloud

Scalable

- API mass generation of data validation pipeline
- Enterprise-grade capabilities including orchestration, fault tolerance, security, encryption
- Parallel processing natively on target platforms
- Universal solution for both Hadoop and Enterprise Data Warehouse (EDW) data sets

The Infoworks Difference

The Infoworks approach automatically validates data throughout the course of migration:

- Data validation as data is being moved to the cloud
  - File-based checksum validation for Hadoop data replication
  - Aggregate query-based validation for non-Hadoop/EDW data sources.
- Data validation once data is in the cloud
  - Column checksum validations
  - Duplicate check validations
  - Row count check and column aggregations (max, min, avg, sum)
  - Data profiling validations
  - Row-hash comparison (optional)
  - SQL Exploration (for custom queries to run)

Infoworks data validation is supported across hybrid multi-cloud, data lake and data warehouse environments; the example below illustrates the data validation process within a GCP environment.

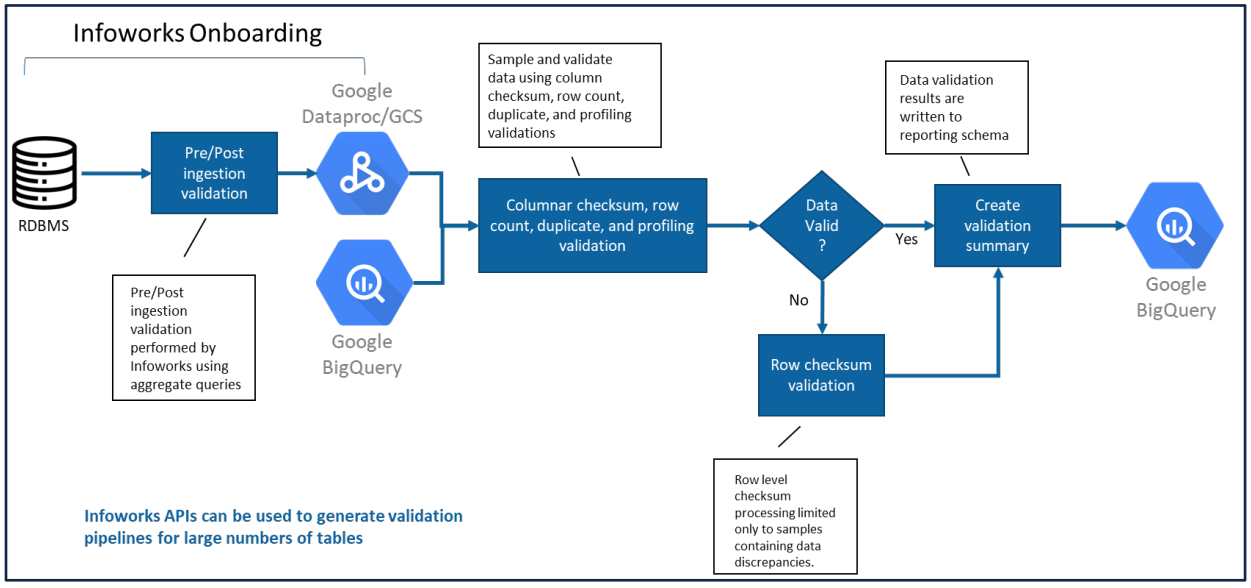


Figure 1: Data validation process in GCP environment.

Infoworks accelerates data validation by up to 8-10x over traditional approaches.

Layered Data Validation

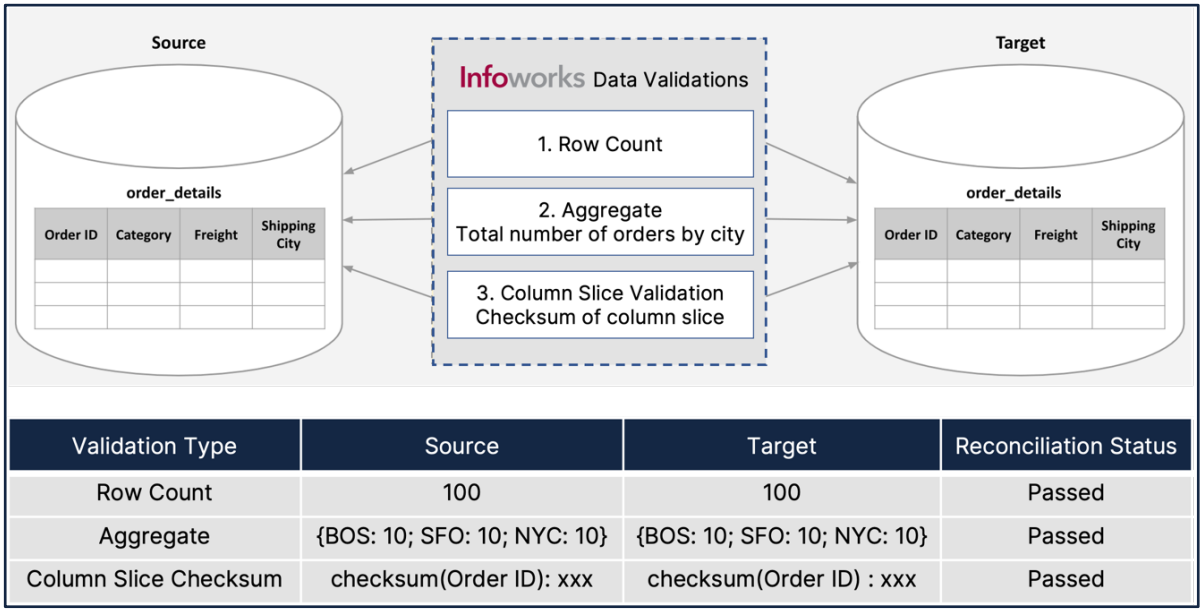
Infoworks uniquely applies a layered data validation approach to validate migrated data at scale, increasing performance by up to 8-10x over traditional approaches:

- The layered validation approach progressively processes each layer of validation, comparing results on the source and target data.
- If any layer of validation fails, metadata is generated for reporting, and an alert is raised to notify users of the data validation error.
- The Infoworks approach takes advantage of column-store functionality to improve performance by up to 8-10x over row-level approaches.

Validation Layer	Validation Layer Description	Validation Type
1	Infoworks automatically calculates checksums at the file and partition level to guarantee that the data in the cloud matches the data on premise.	Standard
2	Row-counts are compared on source and target files/tables.	
3	Aggregated column values for user specified columns (with or without window function) are compared on the source and target (~10x faster than row wise checksum).	Advanced
4	Column slice validation compares column level checksums for user specified columns on source and target.	

Figure 2: Layered data validation process

The below example illustrates layered validation with table 'order_details' being migrated from on-premises to the cloud. The layered data validation process progresses through validation of row counts, column aggregates, and column slice validation based on column checksums.



Validation Type	Source	Target	Reconciliation Status
Row Count	100	100	Passed
Aggregate	{BOS: 10; SFO: 10; NYC: 10}	{BOS: 10; SFO: 10; NYC: 10}	Passed
Column Slice Checksum	checksum(Order ID): xxx	checksum(Order ID) : xxx	Passed

Figure 3: Example of layered data validation

Accelerating Data Validation at Scale with Automated Mass Pipeline Generation

Infoworks API's accelerate data validation at scale by automating the generation of data validation pipelines in bulk for large numbers of tables.

Infoworks automated mass pipeline generation uses REST API endpoints to generate visual, no-code data validation pipelines in bulk and leverage this approach at very large scale in production (100's of 1000's of tables). Dedicated compute infrastructure and elastic scalability provide performance acceleration. Built-in job orchestration provides flexibility in addressing validation exceptions.

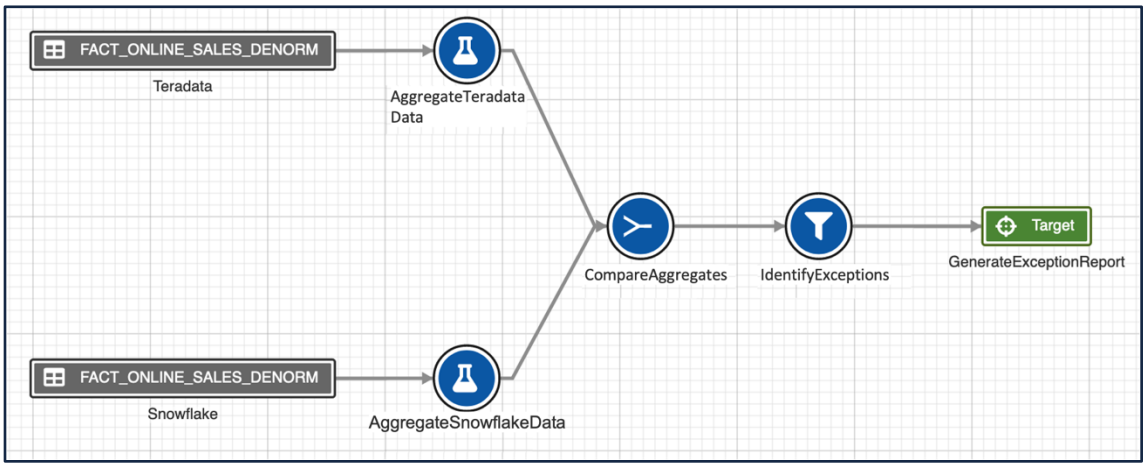


Figure 4: Data validation pipeline in Infoworks

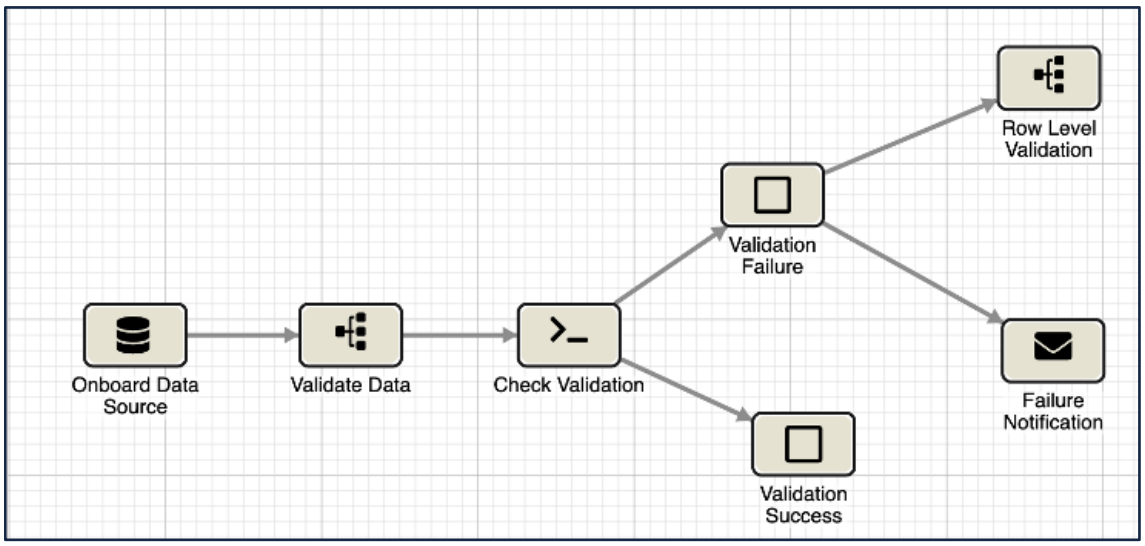


Figure 5: Data validation workflow in Infoworks

The Impact of Infoworks Data Validation

Case Study: Enterprise Digital Transformation at Leading Healthcare Solutions and Services Company

Background

A Fortune 50 Healthcare Solutions company required a solution to accelerate their cloud migration. The challenge: migrating their massive volumes of data to the cloud, ensuring the data quality and integrity was validated at every step throughout the course of the migration. A universal approach to validation that was able to expand to petabyte scale was crucial to their success.

Challenges

- Computational cost and time introduced barriers to compare datasets across environments at scale.
- Existing tools, frameworks, and hand-coded solutions introduced too much overhead and were difficult to maintain.
- Validating petabytes of data at scale
- Protecting data privacy and security
- Running data validation across heterogeneous platforms created platform-specific validation requirements.

Solution

Infoworks data validation optimized cost, time, and quality, enabling data validation in a fraction of the time of legacy approaches:

- Unified, end-to-end solution vs. multiple point tools
- Automated historical and incremental data validation including file-based checksums for Hadoop replication, spot checks, column checksums, duplicate checks, row counts, aggregates, data profiling and row checksums.
- Automated error detection and reporting capabilities ensured data validation at every hop.
- Accelerated large-scale validations, leveraging Infoworks APIs to automate validation pipeline creation for large numbers of tables.

Conclusion

Data validation is one of the most important factors that dictate the success of the data migration process; it is critical to the success of cloud migration. However, traditional approaches to data validation are expensive, resource-intensive, prone to error, or unable to scale and support today's hybrid environments. Enterprises are migrating petabytes of data, 100's of 1000's of data tables. Automation is essential to ensuring timely, accurate, scalable data validation.

Infoworks is the only solution that provides automated, end-to-end data validation. Applying innovative approaches including layered data validation and automated mass pipeline generation,

Infoworks accelerates data validation by up to 8-10x over traditional approaches, and overcomes the challenges historically inherent to data validation, providing:

- Data validation at scale in a fraction of the time required of traditional approaches; no specialized resources required, zero business disruption.
- Optimized performance and scalability that avoid additional computational costs incurred using other approaches.
- Seamless integration into any cloud environment, supporting today's hybrid multi-cloud, cloud data lake and data warehouse environments.
- End-to-end automation that eliminates the risks common to manual approaches, ensuring massive volumes of data are accurately and continuously validated.

About Infoworks

Better, faster, lower-cost cloud migration
Migrate to the cloud and modernize your cloud data operations 3x faster at 1/3 the cost. Deploy new analytics use cases 4x faster.

Infoworks pioneered end-to-end automation to migrate to the cloud and modernize cloud data operations better, faster, and at lower cost. Some of the world's largest enterprises rely on Infoworks to realize the true value of their most strategic asset - their data.

AUTHORS

Roy Effendie
Senior Director, Customer Solutions
Infoworks

Vineet Jain
Solution Architect
Infoworks

LEARN MORE QUICK LINKS

Infoworks Replicator Overview

Infoworks Platform Overview

FOR MORE INFORMATION

www.infoworks.io

email: info@infoworks.io