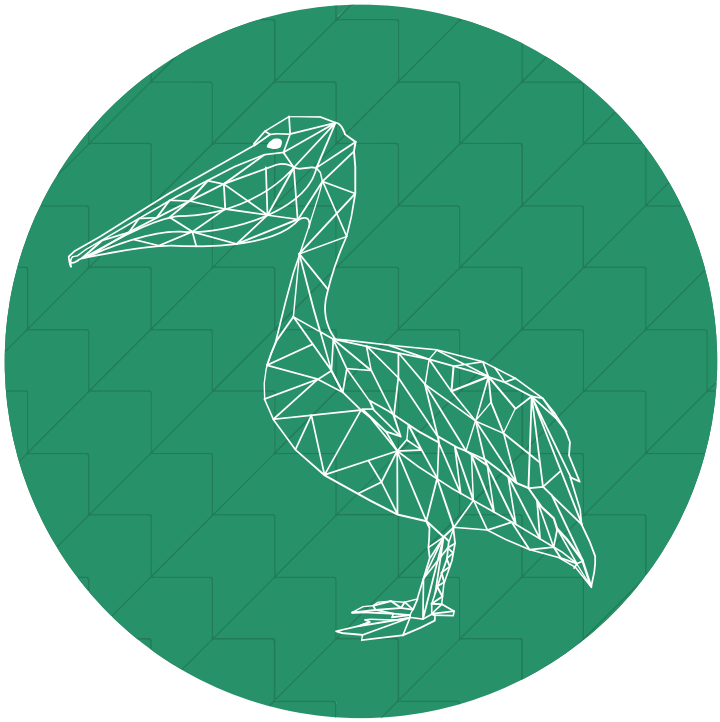
datametica
GIVING DATA WINGS



A QUICK GUIDE TO DATAMETICA'S AUTOMATED DATA VALIDATION TECHNOLOGY

TABLE OF CONTENTS

Data validation or ETL validation testing is a crucial step in data warehouse, database, or data lake migration. It is also an important process for regular check-ups and maintenance of data quality. Validating millions of tables holding billions of records within defined timelines while obtaining the appropriate test coverage is also one of the most complex challenges in a data migration project. According to Gartner, more than 50% of data migration initiatives will run over budget and/or cause some type of business impact due to poor execution. There is always the likelihood of missing data or data corruption during data migration. As a result, it is necessary to test whether the entire dataset is successfully migrated, taking into account both historical and incremental data migration.

The majority of data migration projects go far beyond a simple "lift and shift." All legacy workloads cannot be simply migrated to the new environment as-is; some require additional optimization, while others require complete re-engineering for efficient use of resources in the cloud. The migrated applications must also facilitate all use cases in the new environment, which must be validated against a live dataset. Manually validating the accuracy of migrated code and applications on the destination is an extremely tedious and time-consuming process. Here are some of the complexities at work:

- Different code types—ETL workflows, orchestrator scripts, procedural logic, etc.
- Complex business logic.
- Complex and conditional query logic.
- Platform specific code constructs.
- Platform specific performance hacks.
- Multiple enterprise scenarios and edge cases.

How does Pelican Ease the Complex Process of Data Validation?

Pelican is Datametica's cutting-edge automated data validation and reconciliation tool. It not only eases post migration data validation but also helps businesses in checking and maintaining data quality and data reconciliation. It automatically performs data validation testing or ETL validation testing by evaluating and reconciling petabyte-scale data at the cell level across different source and destination databases with no data migration.

Datametica's Pelican enables businesses to do data quality testing and compare any quantity of data as many times as they desire. It helps to reduce the risks associated with data migration by offering cell-level validation (including table, column, and row-level comparison), delivering 100% accuracy, running both new and old systems concurrently, and minimizing the unit testing associated with a modernization program. During the validation process, Pelican also assures data security.

Why is Pelican Needed?

Pelican automatically runs data validation processes, making the process seamless. It compares datasets and validates the data stored in the legacy system and the cloud with 100% accuracy, thus allowing businesses to confidently decommission their existing data warehouse. It also tackles various data validation obstacles that make the process more complex, time and resource-consuming.

Problem 1: Schema Validation

Checking schema is an important step in the data migration and validation process. Manually verifying a schema with thousands of tables is a very time-consuming operation.

Solution

Pelican performs data type validation and creates column to column mapping. This mapping further helps in highlighting discrepancies between source and destination data. Pelican validates the name of the tables, column order and numbers.

Problem 2: Cell-by-Cell Comparison

This level of testing ensures complete ETL validation of the data, which helps to avoid time-consuming and costly data quality issues that are commonly discovered after data migration.

Solution

Pelican performs cell-by-cell validation and highlights discrepancies between source and destination.

Problem 3: Reconciliation Checks

Running reconciliation checks is an important step in the data validation process. This process makes sure that the data is not corrupted, the date formats are preserved, and the data is fully loaded.

Solution

Pelican runs reconciliation checks and showcases number of duplicate cells, number of mismatches. It also highlights missing or extra rows/columns in source and destination datasets.

Problem 4: NULL Validations

NULL validation is essential to ensure that NULLs are not accidentally replaced with valid data. Also, NULL and blanks appear virtually identical and are difficult to distinguish.

Solution

Pelican identifies NULL cells. It also helps in distinguishing NULL and Blank cells and columns.

Problem 5: Security Validation

Security testing evaluates whether or not all security user roles have been successfully moved to the new cloud database, preventing unauthorized user access to system data.

Solution

Pelican validates the migration of all the security roles that are stored in the database.

Business Opportunity that Pelican Helps Unlock

Pelican has been designed/developed to make the data validation or ETL validation process seamless and time and cost-effective. It is an automated data validation technology that conducts data validation testing automatically by comparing and reconciling huge quantities of data at the cell level across heterogeneous source and destination systems with zero data movement. It performs cell-level validation with 100% accuracy by running both new and old systems concurrently.

Pelican tackles all the major data validation challenges and concludes the process with 100% accuracy and at significantly lower time and cost.

Challenges Faced by Pelican while Validating Data

Pelican very effectively handles the data validation process, but there are a few challenges that this state-of-the-art technology tackles while performing the validation process. Here are the challenges faced by the Pelican.

- **Begin the Validation Journey**

Learning where to start is a crucial step in data validation. Pelican analyzes the sequence of validation and starts the validation process.

- **Size of Datasets**

Validating petabytes of data is not an easy job, but Pelican very efficiently performs this task. It validates the datasets in source as well as destination storage with 100% accuracy, irrespective of the size of the datasets.

- **Execution Window**

As Pelican performs the data validation of the production data while the systems are live, it executes the validation procedure in the small windows that it gets while the data is being used. Pelican performs a complete validation process on the live systems, without hindering the performance of the systems.

Datametica's Automated Data Validation Technology

Datametica proprietary, Pelican has been designed to ease the data validation process. Pelican addresses all the complexities of data validation. Pelican's algorithm is designed to handle and validate petabytes of data with billions of tables with ease, without moving the actual datasets, at lower cost and time, and with 100% accuracy.

- **Pelican - High-Level Logical View**

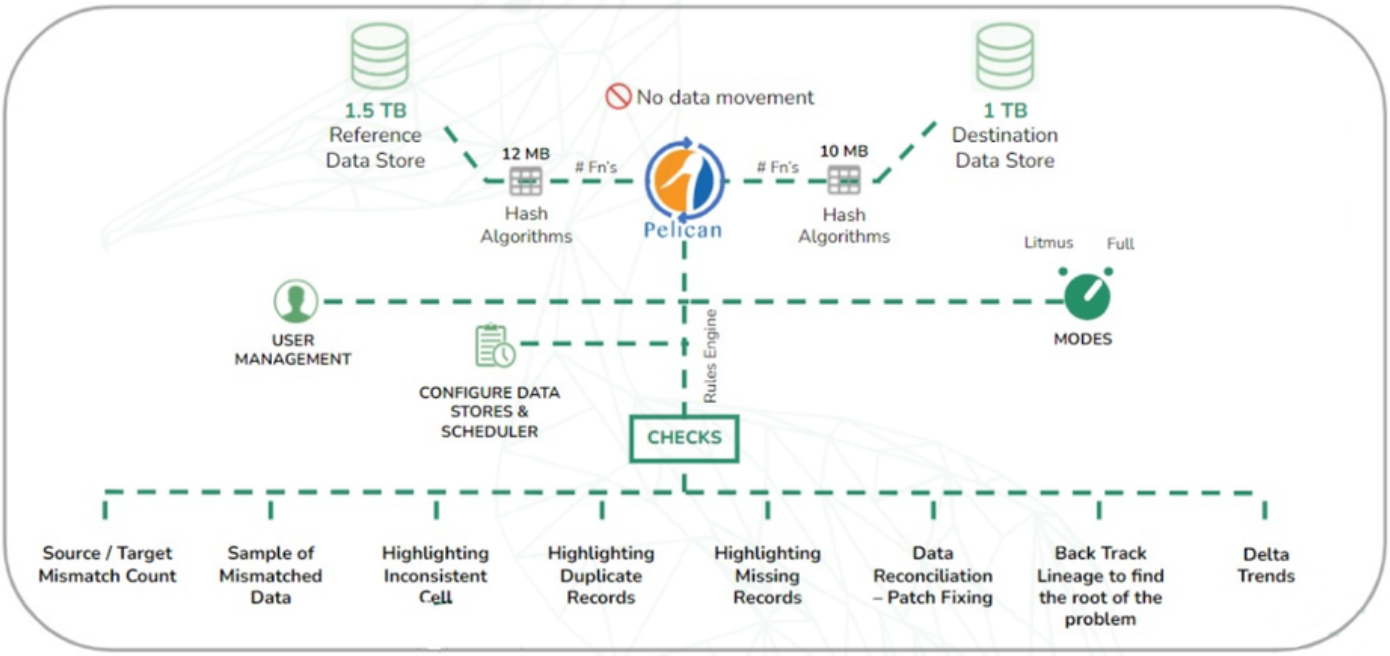


*Image of Pelican Architecture

- **Datametica's Pelican: Key Features that Make it a Market Leader in Automated Data Validation**

There are some automated data validation and reconciliation tools available in the market, but some key features of Pelican differentiate it in the market, make it more unique, and project it as a data validation leader. Here is a quick comparison:

Features	Pelican	Industry Peer	Industry Peer	Industry Peer	Industry Peer
Support	All Legacy Data systems and Cloud	TD, NZ, Oracle, SQL Server, SAS Vertica, Hadoop	Relational, Flat Files, XML, NoSQL, Cloud, and Big Data Sources	All Legacy Data systems and Cloud	RDMBs
Zero Coding Requirement	Yes	Yes	Unknown	Yes	Unknown
No Data Movement	Yes	No	No	No	No
Cell-level Comparison	Yes	No	No	No	No
Checks Data Type	Yes	Yes	Yes	Unknown	Yes
Match Tables and Views	Yes	Yes	No	No	Yes

Pelican: The Validator

Datametica's Pelican is an automated data validation and reconciliation technology that performs data validation testing automatically by comparing and reconciling petabyte-scale data at the cell level across the diverse source and destination systems with zero data movement.

It helps to minimize the risks associated with data migration by offering cell-level validation (including table, column, and row-level comparison), delivering 100% accuracy, executing both new and old systems concurrently, and minimizing the unit testing associated with a modernization process. During the validation process, Pelican also ensures data security and protects personally identifiable information (PII).

Pelican has 8 major modules that solve all the challenges faced during data warehouse modernization and data validation processes.

- **Module 1: Metadata Validation**

Pelican validates the source and destination table structures with respect to the mapping documents. On both the target and source sides, it validates the following:- It validates Table name, Column name, Column ordering, Data Types on both source and target side.

  - **Table name:** Pelican compares Table name on both the source and target sides.
  - **Column names:** Pelican compares column names on both the source and target sides.
  - **Column ordering:** Pelican compares column ordering on both the source and target sides.
  - **Data types:** Pelican compares data types on both the source and target sides.

- **Module 2: Data Validation**

Pelican performs data validation with no data movement and zero coding. It uses a rule engine and compares cell level validation (Litmus and Full).

  - **LITMUS Mode:**

  The Litmus mode supports analyzing if the tables at source and target are matching or mismatching. It does not support cell level differences or samples.

  - **Full Validation Mode:**

  Along with data validation as in Litmus mode, Full mode displays the cell-level differences through the sample rows fetched from both the source and target tables.
  It performs incremental data validation and supports 26 data pairs.
  Once the validation is done, the application generates statistics with the following information:

    - Count of total rows at source
    - Count of total rows at destination
    - Count of mismatch rows at destination
    - Count of extra rows at destination
    - Count of missing rows at destination
    - Total mismatch row count
    - Validation Status
    - Samples of mismatch data

- **Module 3: Mapping Creation**

Pelican creates source and target table mapping (single table mapping), creates bulk mapping (Bulk mapping functionality enables the user to map all the tables at the same time), and clone mapping (clone the mapping and allow editing in the cloned ones). It also builds a default expression mapping.

- **Module 4: Scheduling**

In Pelican, the user can create a scheduler for a saved mapping, so that, after a specific time period, the scheduler executes the process and validates the source table with the destination tables. Once the user completes the table mapping process, the respective schedulers can be configured. Pelican allows recurrence scheduling. The Pelican APIs also help in orchestration via external tools, e.g., Control M, Tidal, etc. Pelican also allows group scheduling.

- **Module 5: Metadata Migration**

This feature is used to export data from one Pelican instance to another instance of the same version. With this functionality, the user will be able to export and import Pelican metadata, viz., table mappings and their scheduler, for selected pairs of data stores. Pelican also creates import/export mappings for metadata migration.

- **Module 6: Custom SQL**

Pelican gives businesses database views and also allows them to filter data with WHERE clauses. With this feature, Pelican enables the user to get a filtered view of the database according to business requirements. For example, you can search x-form of 100 records, from specific location records.

- **Module 7: Configurations and Infrastructure**

The following configurations and infrastructure-related tasks are performed by Pelican:

- User Management: The User Management section enables secured authorization to the users. Each user is assigned a unique identity to authenticate the application.
- Role: A role can be described as a set of permissions that are given to a user to define its access and permission rights in Pelican. A role can be assigned to multiple users, but a user can only have one role.
- Workspace: The workspace section allows you to add multiple mappings, and that workspace can be assigned to multiple users. A User can have multiple workspaces.

- Auto-metadata backup: The frequency of automatic backups is configurable. The frequency can be changed using a cron pattern in the application.properties file.
- Parameter: Parameter-based execution tailors to specific validation requirements.

- **Module 8: Reports**

Pelican showcases scheduler execution history. It also shows reports in the following manner:

- Sample: It will show list of mismatch data
- Dashboard: Dashboard screen displays result, trends, graph of the scheduled mappings based on datastore, database, tags, and date.
- Lineage: The Lineage feature enables the user to view the last execution of source tables and display them graphically. It becomes easy for the tester or developer to backtrack and identify the table where the issue started. This results in faster root-cause analysis.

Case Study

Pelican helped a top retailer in the U.S. save 90% on data validation costs and a remarkable amount of time.

- **Context**

One of the top retailers in the United States wanted to perform data validation during the process of data warehouse modernization. The client was migrating its legacy data warehouse to the Google Cloud Platform and wanted to gain confidence in order to decommission its existing data warehouse. The client was planning to perform the validation process manually by employing 25 BQ engineers. Thanks to Pelican, that helped the client save a significant amount of time and cost.

- **The Objective**

Top US retailers desired data validation during the process of data warehouse migration to the GCP. Initially, they requested 25 BQ engineers for it, which required a huge investment of more than $3.4 million. This method of data validation was a very time consuming and laborious job. Then the client was introduced to Datametica's automated data validation technology, Pelican.

- **The Challenges**

Manual data validation would have been the work of 25 BQ engineers. They would only perform sample testing (not on the complete database). Additionally, they would be unable to perform multiple instances of data validation. This was also a highly time-consuming task with a significant risk of human error, which would hamper the overall validation effort.

- **Solutions**

Datametica utilized Pelican for data validation and reconciliation during the migration to the GCP. Pelican is a one-of-a-kind tool that validates and reconciles petabyte-scale data at the cell level and across heterogeneous systems to automate, accelerate, and ease data validation.

Key Features of Pelican:

- Automation: Automated data validation & reconciliation
- Zero Factor: No data movement & zero coding requirement
- Accurate: Intelligent comparison at the cell level
- Parallel run: Validation during migration
- Insights: Detailed reports identifying discrepancies

Read more case studies about the noteworthy projects done by Datametica.

Conclusion

In comparison to the prior manual technique, the customer saved 90% of the cost of the data validation process by collaborating with Pelican. Pelican also helped save millions of dollars by replacing 25 BigQuery engineers. It also contributed to significant time savings.
The decision to deploy Pelican also had the added benefit of validating the entire dataset at the cell level each and every time it was required, instead of the usual approach of sampling and one time-validation. All of this guaranteed 100% accuracy in the data validation process and built confidence for decommissioning the legacy system.

Book your Demo Now!!!

About Datametica

Datametica is a global leader in data warehouse modernization and cloud migration. Datametica empowers organizations to streamline their modernization journey by automating data, workload, ETL, and analytics migration to the cloud. Datametica automates and accelerates data migration to the cloud with the help of their automated product suite: Eagle - Data Warehouse Assessment & Migration Planning Product; Raven - Automated Workload Conversion Product; and Pelican - Automated Data Validation and Reconciliation Technology. This enables us to eliminate anxiety from the migration process, making it faster, with greater accuracy, with less risk, and at a lower cost.

Datametica specializes in transforming legacy Teradata, Oracle, Hadoop, Netezza, Vertica, and Greenplum databases, as well as ETLs such as Informatica, Datastage, AbInitio, and others, to cloud-based data warehousing, with additional capabilities in data engineering, advanced analytics solutions, data management, data lake implementation, cloud optimization, and data lake management.