FirstEigen White Paper

How to Architect Data Quality on Snowflake – Serverless, Autonomous, In-Situ Data Validation

FirstEigen™ | Gartner Cool Vendor | IDC Innovator

www.FirstEigen.com • contact@firsteigen.com

Executive Summary

Without adequate and comprehensive validation, a data warehouse becomes a data swamp.

With the accelerating adoption of Snowflake as the cloud data warehouse of choice, the need for autonomously validating data has become critical.

While existing data quality solutions provide the ability to validate Snowflake data, these solutions rely on a rule-based approach that is not scalable for 100s of data assets and often prone to rules coverage issues. More importantly, these solutions provide an easy way to access the audit trail of results.

Solution: Organizations must consider a scalable solution that can autonomously monitor 100s of tables to detect data errors as soon as the data lands.

Current Approach & Challenges

The current focus in Snowflake Data Warehouse projects is on data ingestion, the process of moving data from multiple data sources (often of different formats) into a single destination. After data ingestion, business stakeholders use and analyze data, which is where data errors/issues begin to surface. As a result, business confidence in the data hosted in Snowflake reduces. Our research estimates that an average of 20-30% of any analytics and reporting projects in Snowflake is spent identifying and fixing data issues. In extreme cases, the project can get abandoned entirely.

Current data validation tools are designed to establish data quality rules for one table at a time. As a result, there are significant cost issues in implementing these solutions for 100s of tables. Table-wise focus often leads to an incomplete set of rules or often not implementing any rules for certain tables resulting in unmitigated risks.

The data engineering team experiences the following operational challenges while integrating current data validation solutions:

- It takes time to analyze data and consult the subject matter experts to determine what rules need to be implemented.
- Implementation of the rules has to be specific for each table. So, the effort is linearly proportional to the number of tables in Snowflake.
- Data needs to be moved from the Snowflake to the data quality tool resulting in latency and significant security risks.
- Existing tools come with limited audit trail capability. Generating an audit trail of the rule execution results for compliance requirements often takes time and effort from the data engineering team.
- Maintaining the implemented rules as the data evolves.

Solution Framework

Organizations must consider data validation solutions that, at minimum, meet the following criteria:

1. **Machine Learning Enabled**
   Solutions must leverage AI/ML to:
   - Identify and codify the data fingerprint for detecting data errors related to **Freshness, Completeness, Consistency, Conformity, Uniqueness, and Drift.**
   - Efforts required for establishing validation checks should not depend on the number of tables. Ideally, Data Engineer/Stewart should be able to develop validation checks for 100s tables with a single click.

2. **In-Situ**
Solutions must validate data at the source without moving the data to another location to avoid latency and security risks. Ideally, the solution should be powered by Snowflake for performing all the data quality analysis.

3. **Autonomous**
Solution must be able to:
    - Establish validation checks autonomously when a new table is created.
    - Update existing validation checks autonomously when the underlying data within a table change.
    - Perform validation on the incremental data as soon as the data arrives and alert relevant resources when the number of errors becomes unacceptable.

4. **Scalability**
The solution must offer the same level of scalability as the underlying Snowflake platform used for storage and computation.

5. **Serverless**
Solutions must provide a serverless scalable data validation engine. Ideally, the solution must be using SNOWFLAKE's underlying capability.

6. **Part of the Data Validation Pipeline**
The solution must easily integrate as part of the data pipeline jobs.

7. **Integration and Open API**
Solutions must open API integration for easy integration with the enterprise scheduling, workflow, and security systems.

8. **Audit Trail/Visibility of Results**
Solutions must provide easy to navigate audit trail of the validation test results.

9. **Business Stakeholder Control**
Solutions must provide business stakeholders complete control of the auto-discovered implemented rules. Business stakeholders should be able to add/modify/deactivate rules without involving data engineers.

Conclusion

Data is the most valuable asset for modern organizations. Current approaches for validating data, particularly SNOWFLAKE, are full of operational challenges leading to trust deficiency, time-consuming, and costly methods for fixing data errors. There is an urgent need to adopt a standardized autonomous approach for validating the SNOWFLAKE data to prevent Data Warehouses from becoming a data swamp.

DataBuck - Serverless, Autonomous, In-Situ Data Validation

FirstEigen's DataBuck is recognized by Gartner and IDC as the most innovative data validation software for the Lake and the Cloud. By leveraging AI/ML it's >10x more effective in catching unexpected data errors in comparison to traditional DQ tools. DataBuck's out-of-the-box 9-Point Data Quality checks needs no coding, minimal human intervention, and is operationalized in just a few clicks. It increases the scalability of discovering and applying essential data quality checks to 1,000's of tables by auto-discovering relationships and patterns, auto updating the rules, and continuously monitoring new incoming data. It reduces man-years of labor to days.

DataBuck

DataBuck – Benefits

People Productivity Boost >80%

Reduction in Unexpected Errors: 70%

Cost Reduction >50%

Time Reduction to Onboard Data Set ˜90%

Increase in Processing Speed >10x

Cloud Native