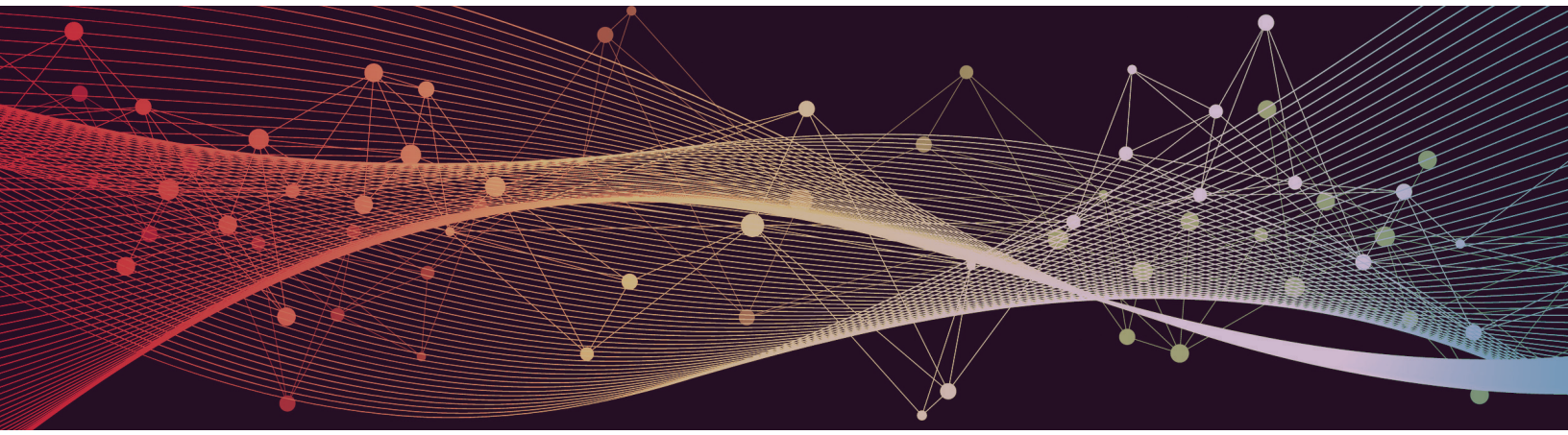
FirstEigen White Paper



How to Establish Continuous Data Validation in Snowflake in 60 Seconds?

FirstEigen™ | Gartner Cool Vendor | IDC Innovator

www.FirstEigen.com • contact@firsteigen.com

Use DataBuck for Autonomous Data Validation of Snowflake tables.

With the accelerating adoption of Snowflake as the cloud data warehouse of choice, the need for validating data has become critical. According to a 2021 study by Boston Consulting Group (www.linkedin.com/pulse/data-quality-key-element-driving-competitive-advantage-lucas-quarta), data quality is lagging in most companies.

Despite significant investments in data quality solutions, most organizations are not able to ensure quality in their data assets because of the following challenges:

- **High Cost of Implementation:** Existing Data Quality solutions rely on a rule-based approach. As a result, implementation effort is linearly proportional to the number of tables in Snowflake. Maintaining thousands of implemented rules as the data evolves adds to the total cost of ownership.
- **Architectural Limitations:** Many existing tools are not architected to validate billions of records that some of the Snowflake tables may contain. In addition, Data needs to be moved from Snowflake to the Data Quality solution, resulting in latency and significant security risks.
- **Knowledge Gap:** Data quality analysts are often unfamiliar with the data assets. To create data quality rules, they need to consult subject matter experts extensively. In the Snowflake Data Cloud, as organizations share datasets – Data Quality analysts may not have access to the subject matter experts from another organization.

FirstEigen developed DataBuck to address these issues within the Snowflake environment.

What is DataBuck?

DataBuck is an autonomous "Powered by Snowflake" data validation solution for Snowflake. It establishes a data fingerprint and an objective data trust score for each data asset (Schema, Tables, Columns) present in Snowflake using its ML capabilities. Trust in data will no longer be a popularity contest. There is no need for individuals to give their subjective opinion on the health of a table/file. All stakeholders can universally understand the objective Data Trust Score.

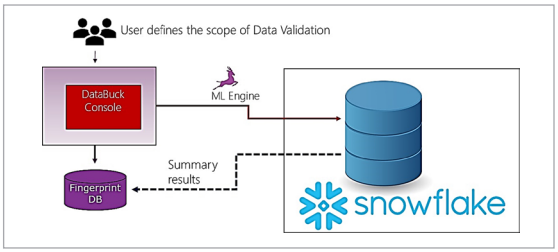
More specifically, it leverages machine learning to measure the data trust score through the lens of standardized data quality dimensions, as shown below:

1. **Freshness** — determine if the data has arrived before the next step of the process
2. **Completeness** — determine the completeness of contextually essential fields. Contextually important fields should be identified using various mathematical and machine learning techniques.
3. **Conformity** — determine conformity to a pattern, length, and format of contextually important fields.
4. **Uniqueness** — determine the uniqueness of the individual records.
5. **Drift** — determine the data drift of the key categorical and continuous fields from the historical information
6. **Anomaly** — determine volume and value anomaly of critical columns.

DataBuck can auto-trigger Data Trust Score as soon as new data lands in a Snowflake table or can be scheduled to run at a specific time or as part of the data pipeline.

How Does DataBuck Work?

In DataBuck, the user provides Snowflake connection information and the database details to trigger the continuous data validation process. Once the data validation process is activated, DataBuck sends its ML engine to Snowflake to analyze the data and identify data quality issues. Summary results are then presented to the user through the web console. The user does not need to write rules or move data out of Snowflake at no point in this process.



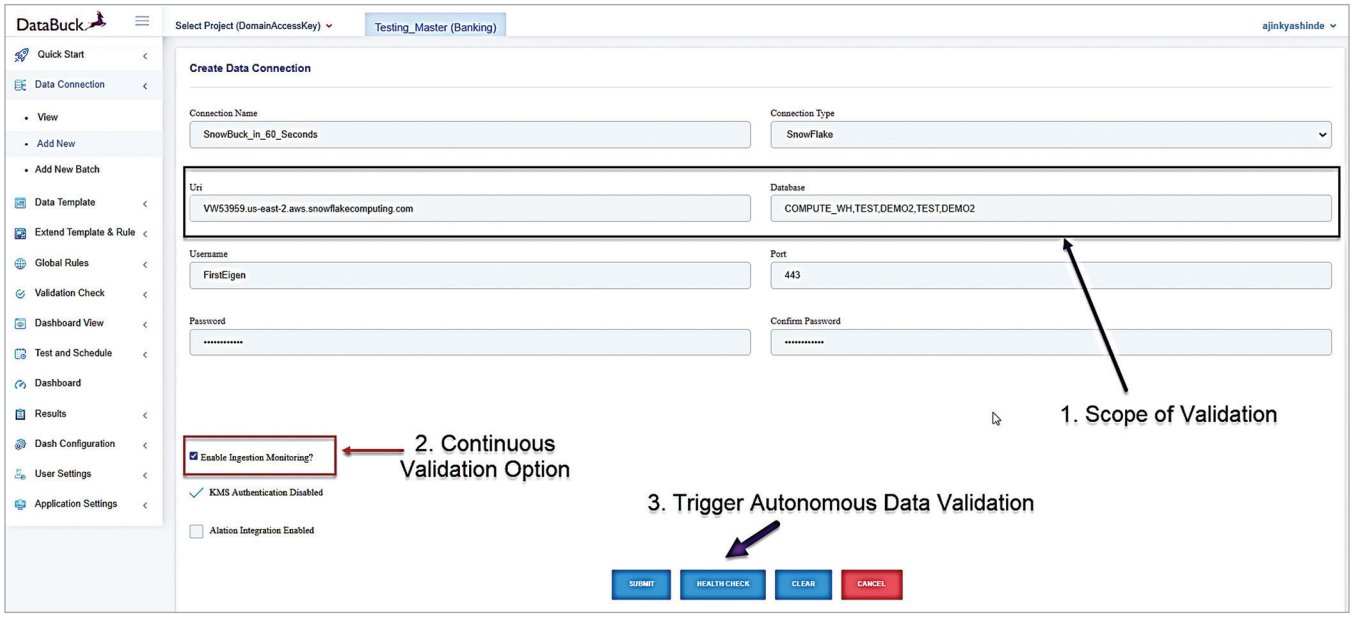
Setting up DataBuck in 60 Seconds

As shown below, the user follows the following process to set up DataBuck in 60 Seconds:

1. Provides database and schema name for which data validation needs to be done

2. Indicates whether continuous data validation needs to be performed or not

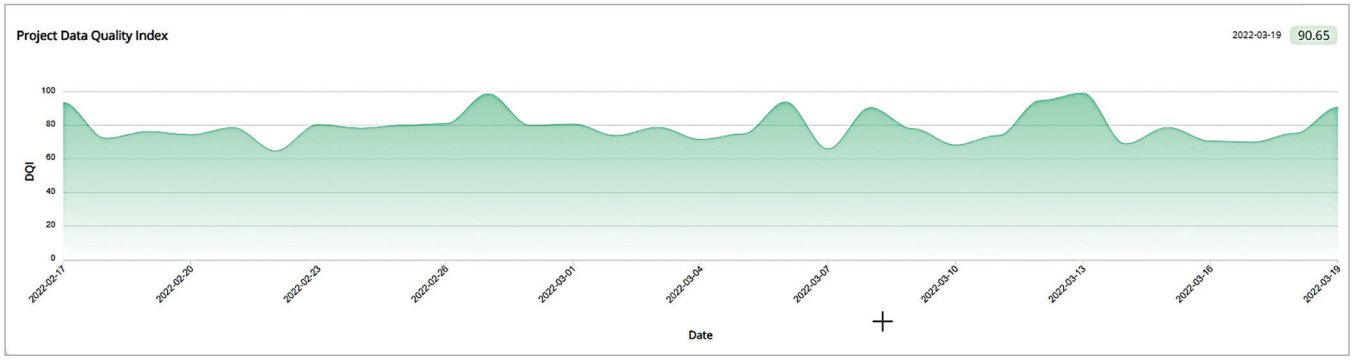
3. Trigger the data validation process by clicking the health check button.



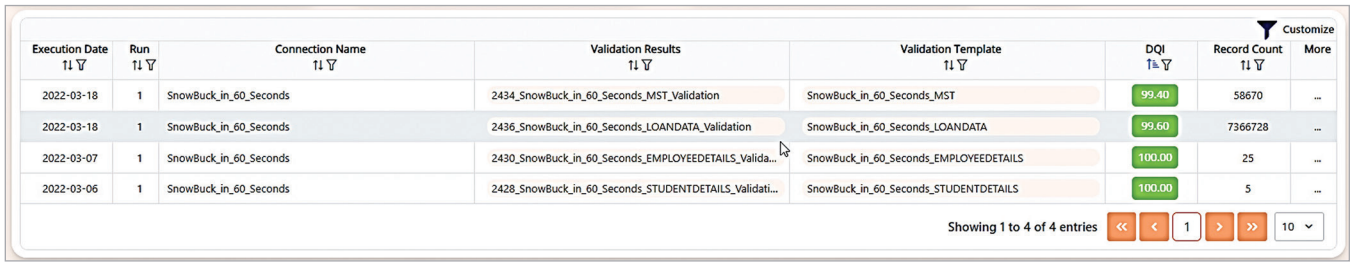
DataBuck Results

DataBuck can validate a snowflake database regardless of the number of tables and size of each individual table. DataBuck provides the following results:

1. Data Quality of a Schema Overtime:



2. Summary Data Quality Results of Each Table

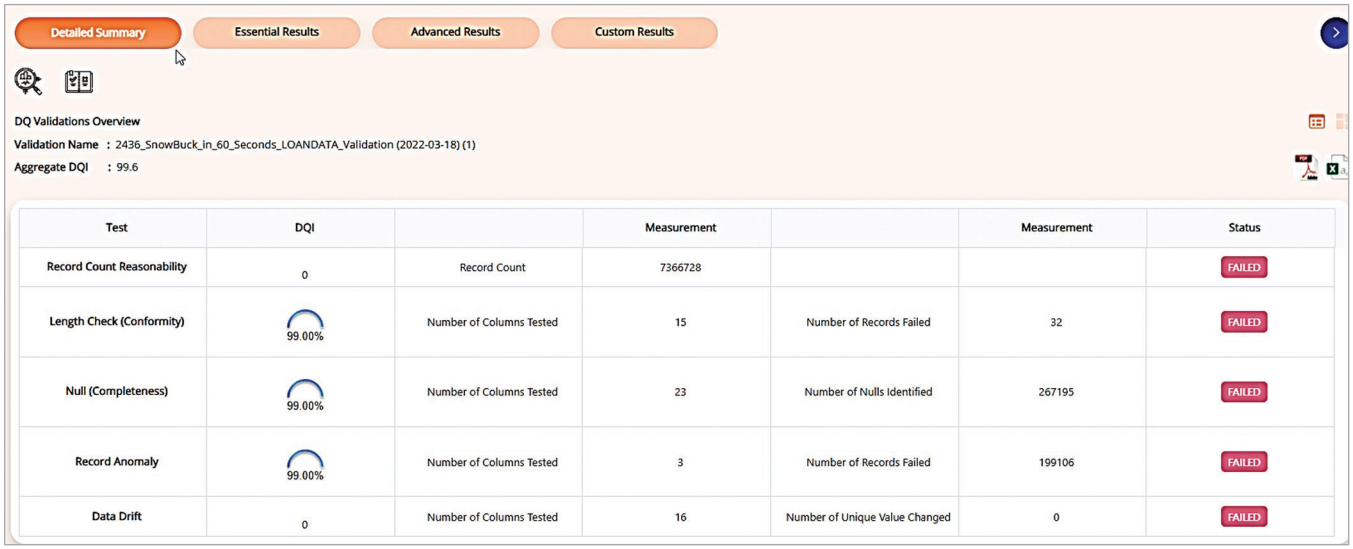


Customize

Execution Date	Run	Connection Name	Validation Results	Validation Template	DQI	Record Count	More
2022-03-18	1	SnowBuck_in_60_Seconds	2434_SnowBuck_in_60_Seconds_MST_Validation	SnowBuck_in_60_Seconds_MST	99.40	58670	...
2022-03-18	1	SnowBuck_in_60_Seconds	2436_SnowBuck_in_60_Seconds_LOANDATA_Validation	SnowBuck_in_60_Seconds_LOANDATA	99.60	7366728	...
2022-03-07	1	SnowBuck_in_60_Seconds	2430_SnowBuck_in_60_Seconds_EMPLOYEEDETAILS_Valida...	SnowBuck_in_60_Seconds_EMPLOYEEDETAILS	100.00	25	...
2022-03-06	1	SnowBuck_in_60_Seconds	2428_SnowBuck_in_60_Seconds_STUDENTDETAILS_Validati...	SnowBuck_in_60_Seconds_STUDENTDETAILS	100.00	5	...

Showing 1 to 4 of 4 entries

3. Detailed Data Quality Results of Each Table



Detailed Summary | Essential Results | Advanced Results | Custom Results

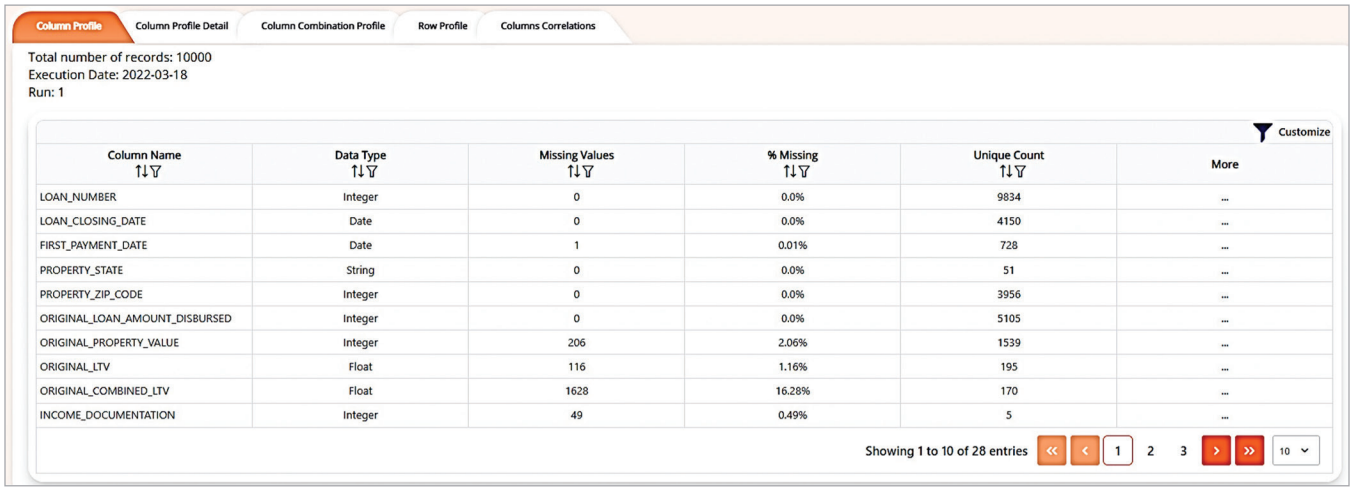
DQ Validations Overview

Validation Name : 2436_SnowBuck_in_60_Seconds_LOANDATA_Validation (2022-03-18) (1)

Aggregate DQI : 99.6

Test	DQI		Measurement		Measurement	Status
Record Count Reasonability	0	Record Count	7366728			FAILED
Length Check (Conformity)	99.00%	Number of Columns Tested	15	Number of Records Failed	32	FAILED
Null (Completeness)	99.00%	Number of Columns Tested	23	Number of Nulls Identified	267195	FAILED
Record Anomaly	99.00%	Number of Columns Tested	3	Number of Records Failed	199106	FAILED
Data Drift	0	Number of Columns Tested	16	Number of Unique Value Changed	0	FAILED

4. Detailed Data Profile of Each Table



Column Profile | Column Profile Detail | Column Combination Profile | Row Profile | Columns Correlations

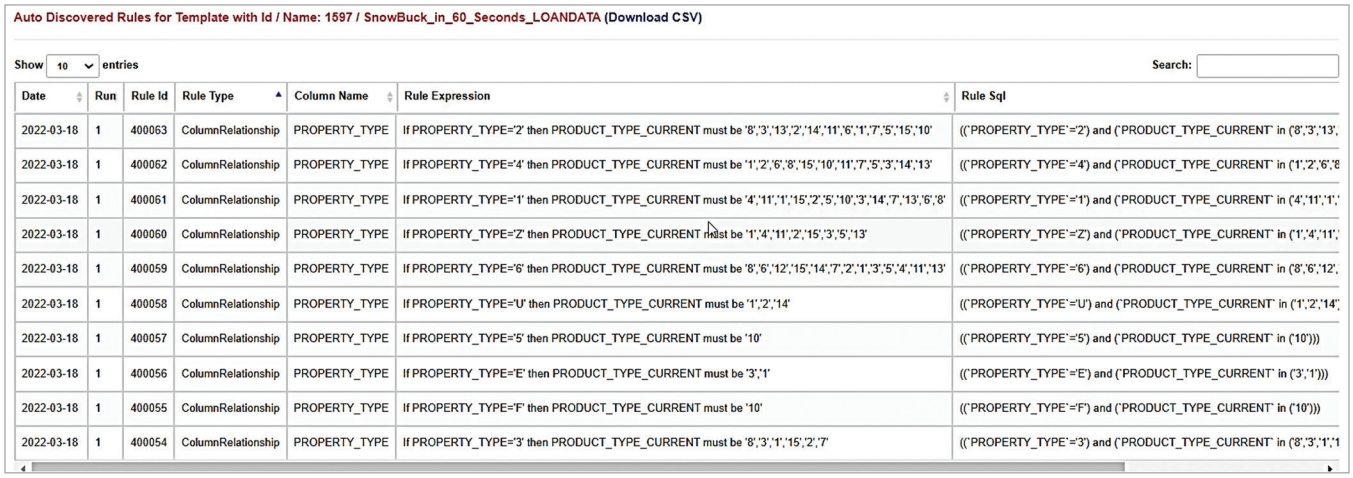
Total number of records: 10000
Execution Date: 2022-03-18
Run: 1

Customize

Column Name	Data Type	Missing Values	% Missing	Unique Count	More
LOAN_NUMBER	Integer	0	0.0%	9834	...
LOAN_CLOSING_DATE	Date	0	0.0%	4150	...
FIRST_PAYMENT_DATE	Date	1	0.01%	728	...
PROPERTY_STATE	String	0	0.0%	51	...
PROPERTY_ZIP_CODE	Integer	0	0.0%	3956	...
ORIGINAL_LOAN_AMOUNT_DISBURSED	Integer	0	0.0%	5105	...
ORIGINAL_PROPERTY_VALUE	Integer	206	2.06%	1539	...
ORIGINAL_LTV	Float	116	1.16%	195	...
ORIGINAL_COMBINED_LTV	Float	1628	16.28%	170	...
INCOME_DOCUMENTATION	Integer	49	0.49%	5	...

Showing 1 to 10 of 28 entries « ‹ 1 2 3 › » 10

5. Discovered Data Quality Rules for Each Table



Auto Discovered Rules for Template with Id / Name: 1597 / SnowBuck_in_60_Seconds_LOANDATA (Download CSV)

Show 10 entries Search:

Date	Run	Rule Id	Rule Type	Column Name	Rule Expression	Rule Sql
2022-03-18	1	400063	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='2' then PRODUCT_TYPE_CURRENT must be '8','3','13','2','14','11','6','1','7','5','15','10'	(('PROPERTY_TYPE'='2') and (`PRODUCT_TYPE_CURRENT` in ('8','3','13',
2022-03-18	1	400062	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='4' then PRODUCT_TYPE_CURRENT must be '1','2','6','8','15','10','11','7','5','3','14','13'	(('PROPERTY_TYPE'='4') and (`PRODUCT_TYPE_CURRENT` in ('1','2','6','8
2022-03-18	1	400061	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='1' then PRODUCT_TYPE_CURRENT must be '4','11','1','15','2','5','10','3','14','7','13','6','8'	(('PROPERTY_TYPE'='1') and (`PRODUCT_TYPE_CURRENT` in ('4','11','1',
2022-03-18	1	400060	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='Z' then PRODUCT_TYPE_CURRENT must be '1','4','11','2','15','3','5','13'	(('PROPERTY_TYPE'='Z') and (`PRODUCT_TYPE_CURRENT` in ('1','4','11',
2022-03-18	1	400059	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='6' then PRODUCT_TYPE_CURRENT must be '8','6','12','15','14','7','2','1','3','5','4','11','13'	(('PROPERTY_TYPE'='6') and (`PRODUCT_TYPE_CURRENT` in ('8','6','12',
2022-03-18	1	400058	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='U' then PRODUCT_TYPE_CURRENT must be '1','2','14'	(('PROPERTY_TYPE'='U') and (`PRODUCT_TYPE_CURRENT` in ('1','2','14',
2022-03-18	1	400057	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='5' then PRODUCT_TYPE_CURRENT must be '10'	(('PROPERTY_TYPE'='5') and (`PRODUCT_TYPE_CURRENT` in ('10')))
2022-03-18	1	400056	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='E' then PRODUCT_TYPE_CURRENT must be '3','1'	(('PROPERTY_TYPE'='E') and (`PRODUCT_TYPE_CURRENT` in ('3','1')))
2022-03-18	1	400055	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='F' then PRODUCT_TYPE_CURRENT must be '10'	(('PROPERTY_TYPE'='F') and (`PRODUCT_TYPE_CURRENT` in ('10')))
2022-03-18	1	400054	ColumnRelationship	PROPERTY_TYPE	If PROPERTY_TYPE='3' then PRODUCT_TYPE_CURRENT must be '8','3','1','15','2','7'	(('PROPERTY_TYPE'='3') and (`PRODUCT_TYPE_CURRENT` in ('8','3','1','1

Summary

DataBuck provides a secure and scalable approach to validate snowflake data in an ongoing manner. All it takes is a single click, and you can validate hundreds of your Snowflake tables.

DataBuck by FirstEigen: 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

FirstEigen™ | Gartner Cool Vendor | IDC Innovator