FirstEigen White Paper

Beyond Observability: Data Trustability Platform for Snowflake Data Cloud

Leveraging Machine Learning for Superior Snowflake Data Quality

FirstEigen™ | Gartner Cool Vendor | IDC Innovator

www.FirstEigen.com • contact@firsteigen.com

Snowflake

Snowflake is a cloud data warehouse that aims to deliver fast and modern cloud-based solutions for your business. It enables easy-to-use data storage, fast processing, and analytic solutions. It is also a data marketplace where organizations can share data securely in minutes rather than months.

Snowflake Data Cloud

The Snowflake Data Cloud Support gives you an all-in-one platform for all your data needs – data warehousing, data lakes, data engineering, data science, data application development, and secure sharing and consumption of real-time/ shared data.[1]

Over 400 million SaaS data sets remained isolated in cloud data storage and on-premise data centers around the world. The Data Cloud breaks down these barriers, allowing you to unify, analyze, exchange, and even monetize your data in real-time. It makes it easy for businesses to connect to a single copy of all their data.

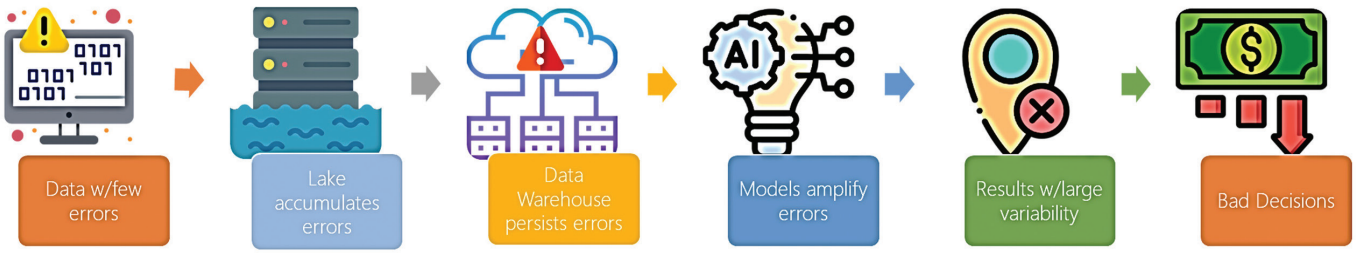
Autonomous Data Trust Score for Snowflake Data Cloud

You can measure how usable your data may be across the enterprise by using an objective 'Data Trust Score.' It allows you to quantify the level of trust in any dataset. The Data Trust Score (DTS) in Snowflake is a metric used to validate and ensure superior Snowflake Data Quality.

What is Trust Score and Why is it Important?

The ability of the consumers of data to determine the data's reliability and quality is known as trust. Data scientists working on advanced use cases need confidence in their data. Minor data errors get amplified into larger errors in analytical model coefficients, resulting in more significant variations in output results. This error amplification is further compounded when the future incoming data also has a small number of errors, as expected. When AI/ML and analytic models constantly learn from data, even < 1% error in data throws their predictions off by 15-20%. This is called the "Bullwhip" effect in the Machine Learning domain. Three pieces of information that help gain trust in a dataset Information are- (1) data quality and coverage, (2) data lineage, and (3) the accountable person, e.g., data steward or data owner. With this information, suitable data sources are more likely to be used, whereas inappropriate data sources that are likely to produce bad results can be avoided.[2]

Exhibit 1



Data Quality and Data Trust are keys to making the most efficient use of data obtained from a public or private marketplace. The more accurate the data, the better the results. A company heavily relies on DTS to ensure correct decision-making and reaching the right customers. It is crucial to measure the accuracy of the data to guarantee that stakeholders and automated systems make decisions based on reliable information.[3]

Imagine if you give $10 to a store, and the cashier is unsure how much the $10 bill is worth. Could it be worth just $5 or maybe $8? Would that not be a fundamental problem for any marketplace? When the product is data, its value depends entirely on your trust. Marketplaces can only function on trust, and trust has to be verified. A DTS is a single number that represents a currency of trust during any data exchange process. DTS is computed using the accepted Data Quality dimensions applied to individual columns in a data set. The individual scores are combined to calculate the total DTS for the entire data set. The combined score is an aggregate of the scores for all columns. It captures the health of the data. The trend and the absolute value of DTS tell us the trustworthiness of the dataset.

Dimensions of Data Quality

Some Data Quality dimensions are discussed here:

1. **Completeness:** It determines the completeness of contextually important fields.
2. **Conformity:** Dataset should contain relevant data and follow specific rules or patterns. This data quality dimension determines conformity to a pattern, length, and format of contextually important fields.
3. **Uniqueness:** This dimension determines the extent of duplicate records. To avoid duplicate records, primary keys are essential to distinguish between different data values and records.
4. **Consistency:** It determines the consistency of intercolumn relationships (e.g., the date of employment must be before the date of retirement).
5. **Drift:** It determines the drift of the key categorical and continuous fields from the historical information.
6. **Anomaly:** It can determine the volume, value, and distribution anomaly of critical columns.

Why Leverage Machine Learning?

Challenges with Traditional approach

- **Knowledge Gap:** Often, data quality analysts, are unfamiliar with the data assets obtained from a third party, either in a public or private context. They need to engage with subject matter experts extensively to build data quality criteria. In a Snowflake Data Cloud, as organizations share datasets, data quality analysts may not have access to subject matter experts from another organization.
- **Time to Use the Dataset:** Even if you are intimately familiar with the dataset, it can take 2 to 5 business days to analyze the data quality. Snowflake Data Cloud reduces the data exchange time drastically. However, adding additional days to perform the data quality manually adds to the timeline and defeats the purpose.

Using a Machine Learning based Approach

Machine Learning is known for solving complex problems and executing results faster than intended without any human error. Using ML in Snowflake Data cloud has many advantages:

- Machine Learning helps to objectively determine data patterns or data fingerprints and translate those patterns to data quality rules.
- Machine Learning can then use the data fingerprints to detect transactions that do not adhere to the rules.
- Implementing an ML approach can help to assess the data health quickly
- ML is usually more comprehensive and accurate than a human-driven data quality analysis

Example – DataSet Name: Analytics Dataset

Let's look at an example to understand the use and purpose of the Data Trust Score. Here, we have a dataset provided by Equifax. This Analytic Free sample dataset contains sample records that represent the entire dataset in terms of data types, organization, and fields. (Please note that this is not actual production data).

The dataset contains granular, loan-level data across credit cycles and asset classes, including vehicle, credit card, mortgage, school loans, and unsecured personal loans, to more accurately predict future performance. The entire dataset is based on an unbiased ten percent statistical sample of the credit-active population in the United States across all geographic boundaries, with data dating back to 2005.

We applied our "Powered by Snowflake" Machine Learning algorithm to validate the trustworthiness of this dataset and generated the following results:

Data Trust Score: 99.6

Test	Data Quality Index	Status	Key Metric-1	Measurement	Key Metric-2	Measurement
Record Count Reasonability			Record Count	2,036		
Length Check (Conformity)	99.0	FAILED	Number of Columns Tested	45	Number of Records Failed	80
Data Completeness	99.0	FAILED	Number of Columns Tested	69	Number of Nulls Identified	462
Data Uniqueness (Primary Keys)	100.0	PASSED	Total Number of Primary Keys	10	Total Number of Duplicates	0
Default Pattern Check (Conformity)	98.0	FAILED	Number of Columns Tested	67	Number of Records Failed	1,618
String Value Drift (Drift and Orphan)			Number of Columns Tested	52	Number of Unique Value Changed	0

Rules applied: ML algorithms discovered a total of 334 rules.

DQ Dimension	Number of Rules
Completeness	69
Conformity	126
Consistency	32
Drift	52
Reasonability	45
Uniqueness	10
Total	**334**

Here is an example of the complexity of rules/patterns discovered by Machine Learning Algorithms:

Rule Id	Rule Type	Column Name	Rule Expression
517144	ColumnRelationship	ECOA	If ECOA='I' then ORIGINATION_PORTFOLIO_TYPE must be 'M','R','I','O','C'
517143	ColumnRelationship	ECOA	If ECOA='J' then ORIGINATION_PORTFOLIO_TYPE must be 'O','C','I','M','R'
517142	ColumnRelationship	ECOA	If ECOA='C' then ORIGINATION_PORTFOLIO_TYPE must be 'I'
517141	ColumnRelationship	ECOA	If ECOA='M' then ORIGINATION_PORTFOLIO_TYPE must be 'I'
517140	ColumnRelationship	ECOA	If ECOA='null' then ORIGINATION_PORTFOLIO_TYPE must be 'R'
517139	ColumnRelationship	ECOA	If ECOA='I' then PORTFOLIO_TYPE must be 'M','R','I','O','C'
517138	ColumnRelationship	ECOA	If ECOA='J' then PORTFOLIO_TYPE must be 'O','C','I','M','R'
517137	ColumnRelationship	ECOA	If ECOA='C' then PORTFOLIO_TYPE must be 'I'
517136	ColumnRelationship	ECOA	If ECOA='M' then PORTFOLIO_TYPE must be 'I'
517135	ColumnRelationship	ECOA	If ECOA='null' then PORTFOLIO_TYPE must be 'R'

Summary

All data, whether from your Data Warehouse or a reputable data provider in a marketplace, must be validated to be trusted. Users have to go beyond mere Observability to Data Trustability. The only practical option to autonomously validate data to obtain an objective Data Trust Score is to leverage Machine Learning. This approach will ensure data from the Snowflake Data Cloud has superior quality data that can be used for its intended purpose.

References

[1] The Snowflake Data Cloud; www.snowflake.com/data-cloud
[2] O. Bieh-Zimmert et al, A capability maturity model for data catalogs (2018), Deloitte
[3] Serverless Autonomous Data Validation in Snowflake, FirstEigen; firsteigen.com/snowflake-data-quality-lp

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