FirstEigen White Paper

Autonomous Data Trust Score for Data Catalogs

FirstEigen™ | Gartner Cool Vendor | IDC Innovator

www.FirstEigen.com • contact@firsteigen.com

Autonomously Assign Data Trust Score to Each Data Asset Present in the Data Catalog

With the accelerating adoption of Data Catalogs as the core component for enterprise data governance, the need to provide information about the health and usability of the data assets has become critical. With the availability of standardized data trust scores within the data catalog, users can easily determine the usability and relevancy of the dataset in their particular use case.

Why Should the Data Governance Team Care About Data Trust Score?

When working with data, a key question that comes to mind is, "Is the data of the right quality for my purpose?" Data scientists working on advanced use cases need confidence in the data they use. Information about data quality, data lineage, and the accountable person, e.g., data steward or data owner, are three sorts of information that help gain trust in a dataset. With this information, suitable data sources are more likely to be used, whereas inappropriate data sources that are likely to produce erroneous results can be avoided.

Data quality and trust are keys to making the most efficient use of data. The more accurate the data, the better the results. It is crucial to measure the accuracy of the data to guarantee that stakeholders and automated systems make decisions based on reliable information.

A Data Trust Score is a single number that represents a currency of trust during any data usage or exchange process. Data Trust Score can serve as a solution to the information asymmetry problem that exists in modern catalogs. The consumer and the producer of data do not have the level of information related to the usefulness of the data.

Data Trust score is generally computed on certain data quality dimensions applied to individual columns in the data set. The scores are then combined to calculate the total quality score for the entire data set. The combined score is an average of the scores for all columns.

Data quality is a mandatory element for enabling large-scale and better use of data in any organization. By adding data quality information within the catalog, users can find insights about the data without ever needing to switch contexts. Data Catalog users often find the need for data quality information and other metadata of the data assets[1].

Although not a data quality tool, a data catalog can help enterprises in providing data quality information[2]. Most catalog provides vendor-neutral integrations with dedicated data quality platforms, thus allowing data catalog search functions to surface higher-quality data first. Data assets with high scores are promoted which are indicative of good data quality. Other metrics such as usage and company certifications can also guide catalog users to the highest quality data assets.

How to Calculate Data Trust Score?

Data Trust score can be computed based on the weighted average data quality score on data quality rules. Progressive organizations categorize the data quality rules along the data quality dimensions. A trust score is then calculated for each dimension and then rolled up using weighted average methods for the dataset.

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

Challenges With Incorporating Data Trust Score into the Data Catalog

Chief Data Officers (CDOs) drive the data strategy for the entire organization. They are accountable for maintaining data quality. The volume of data assets is increasing rapidly, and data teams are struggling to keep up because existing data validation approaches are resource-intensive, time-consuming, costly, and not scalable for 1000s of data assets. Business stakeholders often find errors in their reports and dashboards repeatedly, which leads to mistrust of data.

In general, the data team experiences the following operational challenges while integrating data quality into the data catalog:

- Analyzing data and consulting the subject matter experts to determine what data quality rules need to be implemented is time-consuming. 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
- Implementation of the rules is specific to each data set/table. So, the effort is linearly proportional to the number of tables/buckets/folders in the Data Catalog.
- Existing open-source tools/approaches 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 is cumbersome.

Why Leverage Machine Learning?

ML models can learn from massive volumes of data and uncover hidden data patterns. Moreover, they train overtime on the new data as it changes. Using ML in calculating data trust scores has several 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.

More importantly, when a trust score is assigned by the consumer of data, the score becomes susceptible to the same temptations to hide or even falsify information as the producers of the underlying data assets. Machine learning-based Data Trust score eliminates human biases and provides an objective score that both producer and consumer can agree on[3,4].

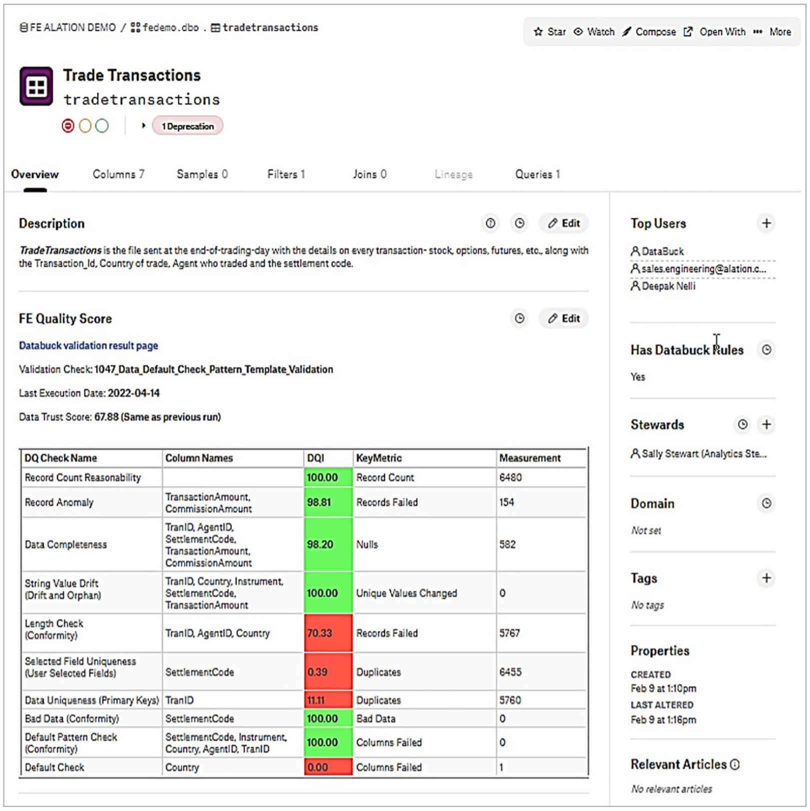
Autonomous Data Trust Score for Data Catalog – an Example

In the example to the right, Machine Learning based was used to calculate the data trust score for each of the data assets registered in a data catalog. The result of the analysis is then presented using standardized data quality dimensions as shown below and embedded within the data catalog using standard APIs.

The Data Trust Score displayed in the summary of the Data Quality analysis shows how the quality score changed between the last two analyses. Every violation discovered can be double-clicked for further information:

- Expand the dimension to see which columns are affected at the data asset level. Click a column name to see the dimension details for that column.
- At the column level, click the dimension name for further details.

This provides the catalog users with enough information regarding the fitness/usability of the dataset in the context of their use case.



Trade Transactions
tradetransactions

Description

TradeTransactions is the file sent at the end-of-trading-day with the details on every transaction- stock, options, futures, etc., along with the Transaction_Id, Country of trade, Agent who traded and the settlement code.

FE Quality Score

Databuck validation result page

Validation Check: 1047_Data_Default_Check_Pattern_Template_Validation

Last Execution Date: 2022-04-14

Data Trust Score: 67.88 (Same as previous run)

DQ Check Name	Column Names	DQI	KeyMetric	Measurement
Record Count Reasonability		100.00	Record Count	6480
Record Anomaly	TransactionAmount, CommissionAmount	98.81	Records Failed	154
Data Completeness	TranID, AgentID, SettlementCode, TransactionAmount, CommissionAmount	98.20	Nulls	582
String Value Drift (Drift and Orphan)	TranID, Country, Instrument, SettlementCode, TransactionAmount	100.00	Unique Values Changed	0
Length Check (Conformity)	TranID, AgentID, Country	70.33	Records Failed	5767
Selected Field Uniqueness (User Selected Fields)	SettlementCode	0.39	Duplicates	6455
Data Uniqueness (Primary Keys)	TranID	11.11	Duplicates	5760
Bad Data (Conformity)	SettlementCode	100.00	Bad Data	0
Default Pattern Check (Conformity)	SettlementCode, Instrument, Country, AgentID, TranID	100.00	Columns Failed	0
Default Check	Country	0.00	Columns Failed	1

Top Users: DataBuck
sales.engineering@alation.c...
Deepak Nelli

Has Databuck Rules: Yes

Stewards: Sally Stewart (Analytics Ste...

Domain: *Not set*

Tags: *No tags*

Properties
CREATED Feb 9 at 1:10pm
LAST ALTERED Feb 9 at 1:16pm

Relevant Articles: *No relevant articles*

References

[1] K. Kashalikar, Integrating Data Quality with Data Catalog (2022), LinkedIn

[2] A CDO's Guide to the Data Catalog, Alation

[3] Certification as Solution to the Asymmetric Information Problem? Georg von Wangenheim, University of Kassel, 2019

[4] From Big Data to Big Profits: Success with Data and Analytics, Russell Walker, 2015

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