Entropy-Driven Data Quality Scoring for Large-Scale Analytics Pipelines

Meher Deepika Uppaluri
AI Financial Consultant
meherdeepikauppaluri@gmail.com

Kuldeep Chowdary Raavi
AI Cyber Security Consultant
kuldeepchowdaryraavi@gmail.com

Aravind Kumar Karpoorapu
AI/ML Research Specialist
aravindkumarkarpoorapu@gmail.com

Abstract

Data quality is critical to the scalability and effectiveness of large-scale analytics pipelines deployed in enterprise environments. In this situation, inaccuracies, incompleteness, or inconsistencies in the data cause a propagation error through the analytical models. This can lead to inaccurate insights, such as operational inefficiencies

and increased company risk. Traditional data quality evaluation approaches may rely on static or manual tests, which are challenging to scale and insufficiently flexible to the evolving data stream. As the complexity and volume of analytic pipelines increase, there is a need to automate the data-driven mechanism to ensure a continual assessment of data quality.

This paper describes an entropy-driven data quality approach used in large-scale analytic pipelines. This methodology leverages information and theoretical entropy to assist in quantifying the uncertainties, variability, and irregularities that exist within streams. This will aid and support the goal of ensuring quality in the framework enterprise on an ongoing basis. As a result, computing the entropy-based score across the feature ensures that the time window and data source are approaching the detection of abnormalities, degradations, and shifts in the distribution of signals with diminishing data quality. The scores are aggregated into a consistent quality metric that can

be integrated into the workflow metrics that already exist.

Keywords*: Entropy, Large Scale Analytic Pipelines, Data quality, Enterprise Analytics, Conventional data*

Introduction

The Enterprise analytics pipeline is demonstrating a growth in support missions, which are crucial in decision-making across domains such as finance, cybersecurity, and even health care. The effectiveness of pipelines is determined by the quality of incoming data. It introduces inaccuracy at an early level, which can cascade via downstream analytics and ML models. Despite the dependency, data quality management is persistent in large-scale challenges characterised by heterogeneous data sources, including high velocity and data distributions.

Conventional data quality techniques often rely on predetermined validation of rules, manual inspections, and even periodic audits. In contrast to usefulness, their techniques tend to treat data quality as a way of demonstrating a static attribute and the problem of adjusting to changes in data behaviour. In reality, rule-based systems often fail to catch subtle kinds of degradation, such as a steady increase in noise or unexpected variability across a feature.

To address limitations, this paper introduces an enterprise-driven strategy that enables the assessment of data quality. Entropy provides a well-principled measure of uncertainty and information content, making it suitable for spotting anomalies and even loss of structure in data streams [1]. As a result, by incorporating entropy-based scoring directly into analytic pipelines, businesses may continue to monitor data quality and identify possible issues before they affect downstream analytics.

The key objective of this paper is to demonstrate how an entropy-based metric may be utilized to provide scalable, interpretable, and automated data quality scores in big analytic pipelines. Through simulations of enterprise-relevant situations and research, the article demonstrates that entropy-driven monitoring aids in

monitoring and improving data reliability, as well as supporting proactive data governance.

Simulation Report

The study involves evaluating the usefulness of a suggesting entropy-driven data quality assessment framework by using enterprise inspiration data set representatives from large-scale analytic pipelines. The data set is organized in a way that is replicating common data quality conditions found in real-world settings. These could be including clean data streams, noisy measurements, missing values, and progressive distributional changes. Furthermore, controlled perturbation contributes to the systematic decline of data quality over time, allowing for the observation and analysis of the influence on entropy-based scoring [7].
Additionally, computing the entropy score across a sliding time window and the feature dimension will help to capture both temporal and structural change in the data [2]. The ratings are aggregated into normalized data quality indicators, which allow for comparison across different pipeline stages. The stimulation architecture is built for continuous data input scenarios, ensuring that the findings match realistic operational realities.
The simulation results will be displayed in simple and easily interpretable visualizations, including a time series plot of the entropy evolution. This will also aid in the presentation of feature-level quality comparisons and reliability curves, which link data quality scores to analytic performance. The visual analytics will demonstrate how entropy-driven scoring detects data degradation, helps target remediation, and improves the reliability of downstream analytics in large-scale corporate pipelines.

Real-Time Scenarios

Financial transaction analytics; in the case of large financial institutions, the analytic pipelines provide a constant flow of transaction data for data reporting, fraud detection, and compliance checks . Data quality issues include missing transaction files, delayed updates, and format errors, which lead to inaccurate financial summaries and regulatory risk. The entropy-driven data quality scores enable continuous monitoring of the transaction stream, detecting unexpected increases in uncertainty across parameters such as amount, timestamp, and even account identification.

The IoT Sensoring Data in Smart Industry: The industrial environment relies on high-frequency sensor data to monitor equipment operation and optimize production processes. Sensor drift, noise, and intermittent failures can result in a steady degradation of data quality without immediately triggering rule-based alarms [5]. In this example, entropy-based scoring is used to sensor data streams; the analytic pipeline detects an increase in unpredictability or a loss of structure in sensor readings. This ensures that the malfunctioning sensor or communication issue is detected early, allowing for preventive maintenance while also boosting the reliability of the analytic procedure.

The customer analytics and personalized platform. The customer analytics pipeline aggregates data from a variety of sources, including web interactions, mobile applications, and CRM systems [6]. Inconsistencies in data or missing variables

can have a negative influence on personalization models and client segmentation. The entropy-driven quality scoring is a feature that highlights features with a high level of uncertainty, such as an incomplete user profile or inconsistency in behavioural signals. As a result, the data engineering team can prioritise data cleaning and enrichment to ensure accurate and trustworthy downstream analytics and recommendations [10].

Graphs

Data Condition	Entropy Scoring	Quality Level
Clean data	Low	High
Noisy Data	Moderate	Medium
Missing values	High	Low

Table 1: Illustration of the data quality under different conditions

Table 1 shows how entropy scores tend to rise as data quality declines. Higher entropy represents a greater probability of entropy and irregularity, indicating lower data quality.

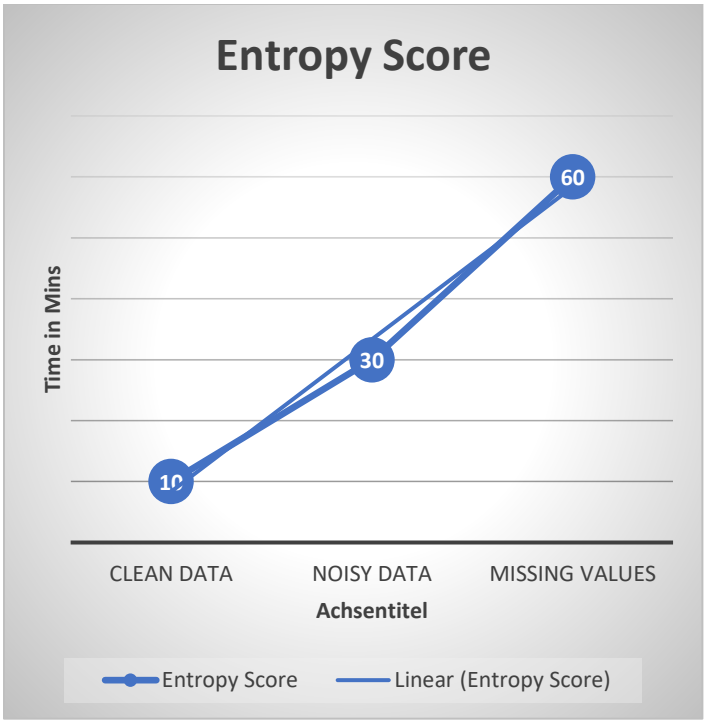


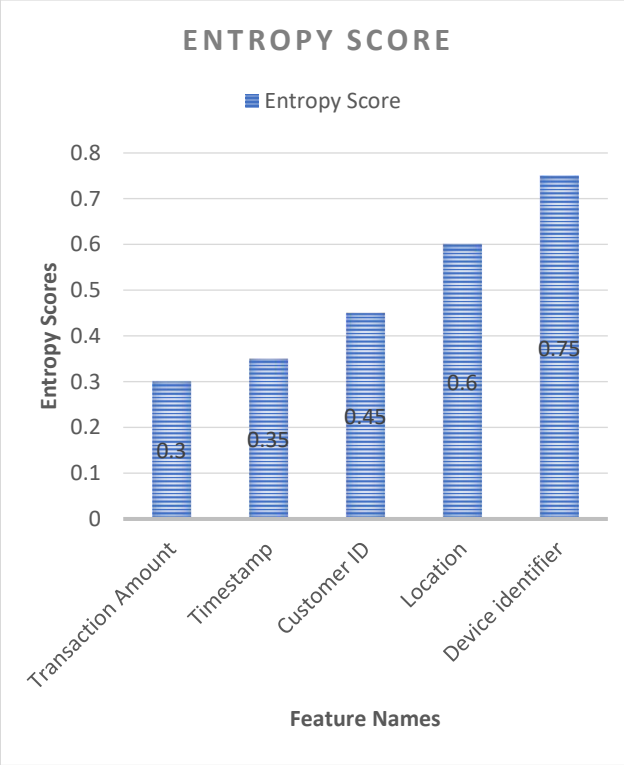
Figure 1 illustrates the entropy score with time.

The developing entropy patterns provide a signal in progressive degradation of the data quality, allowing easy detection of pipeline faults.

Table 2: The Feature Level Entropy-Based Data Quality Score

Feature name	Entropy Score	Quality interpretations
Transactional Amount	0.30	High quality
Time stamp	0.35	High quality
Customer ID	0.45	Moderate quality
Location code	0.60	Low quality
Device	0.75	Low quality

identifier		



.

Table 2 displays the entropy-based score computed for each unique feature inside the analytic pipelines. The feature is containing transaction amounts and timestamps with low entropy, that indicates reliable and high-quality data. On the other side, features with a higher entropy, such as location codes and device identifiers, indicate increased unpredictability or missing values. As a result, the feature-level score supports Fig. 2 by quantitatively highlighting data quality weaknesses, allowing for target correction and prioritization inside large-scale pipelines.

Data	Analytic	Interpretations

Quality Score	reliability (%)	
Low	60	Unreliable
Medium	80	Moderately Reliable
High	90	Highly Reliable

Table 2 illustrates the data quality score and analytic reliability.

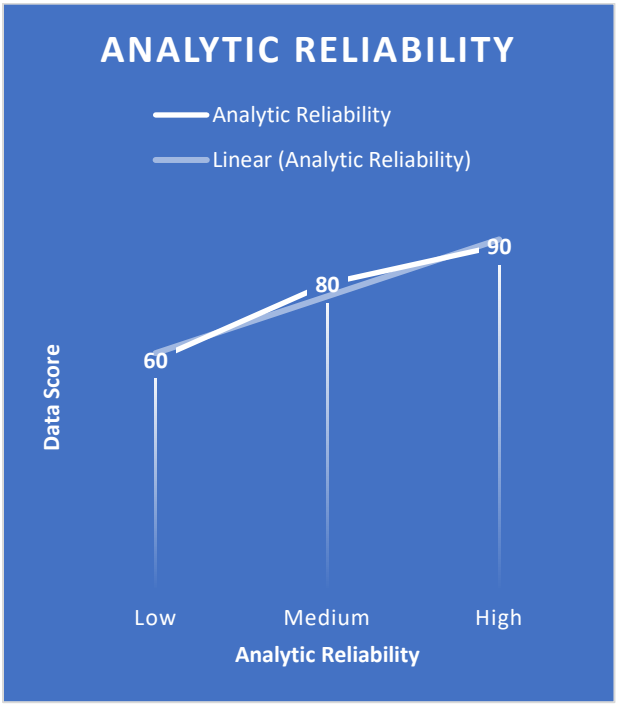


Figure 3: Data Quality Score against Analytic Reliability

Table 3 is showing a quantitative relationship between the entropy-driven data quality score and the analytical dependability. Low-quality data is associated with poor reliability, which raises the chance of erroneous or unstable analytic outcomes [3]. As data quality improves, analytic reliability increases considerably, demonstrating the value of constant monitoring. This table provides Figure 3 by numerically validating the favourable association between data quality and

analytic performance in large-scale corporate pipelines.

Challenges and solutions

The major challenge in entropy-driven data quality assessment is to distinguish between valid data variability and actual quality decline [9]. In this circumstance, some applications may be naturally exhibiting a significant degree of variability, which may be misinterpreted as poor quality. This can be solved by calibrating entropy thresholds against a historical baseline and domain-specific context.

Another challenge stems from the computational overheads of large-scale pipelines. Entropy calculationsis across high-dimensional data can be resource-intensive [1]. This problem can be addressed by implementing a framework that uses window-based calculations and a selectable feature that monitors and reduces processing costs without sacrificing detection capabilities.

Entropy interpretation can also be problematic for non-stakeholders; to make it easier to understand, a normalised quality score technique should be used, as well as a simple depiction. The visualization can be used to translate the entropy values into understandable quality categories [4].

Finally, integrating an entropy-based score into the existing process may add to the operational complexity. The modular deployment and compatibility with standard data processing frameworks can be enabled to allow seamless adoption and expansion of the rollout.

Conclusion

The paper describes an entropy-based quality assessment framework for large-scale analytics pipelines. As a result, leveraging the information in theoretical entropy will suggest a solution that provides automation and a scalable mechanism for continuous assessment of data quality. The simulation findings and corporate scenarios are demonstrating how entropy-based methods may successfully detect data degradation, identify problem features, and promote proactive governance. Furthermore, the visualizations improved the interpretation value and operating usability. As a result, the findings give a high value entropy that is driven and monitored as a dependable foundation component in enterprise analytics, allowing organizations to boost trust in data-driven decisions while also assuring managerial complexity at scale.

References

1. Yu, Y. W., Daniels, N. M., Danko, D. C., & Berger, B. (2015). Entropy-scaling search of massive biological data. *Cell systems*, *1*(2), 130-140. https://www.cell.com/cell-systems/fulltext/S2405-4712(15)00058-7
2. Shakya, S. R., Zhang, C., & Zhou, Z. (2018). Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data. *Int. J. Mach. Learn. Comput*, *8*(6), 577-582. https://www.researchgate.net/profile/Sarbagya-Shakya/publication/328725395_Comparative_Study_of_Machine_Learning_and_Deep_Learning_Architecture_for_Human_Activity_Recognition_Using_Accelerometer_Data/links/5c89557092851c1df93ff1ef/Comparative-Study-of-Machine-Learning-and-Deep-Learning-Architecture-for-Human-Activity-Recognition-Using-Accelerometer-Data.pdf
3. Abdelsamie, A., Janiga, G., & Thevenin, D. (2017). Spectral entropy as a flow state indicator. *International Journal of Heat and Fluid Flow*, *68*, 102-113. https://www.sciencedirect.com/science/article/pii/S0142727X1630933X
4. Mathews, S. M. (2019, July). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In *Intelligent computing-proceedings of the computing conference* (pp. 1269-1292). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-22868-2_90
5. Nakao, A., Du, P., Kiriha, Y., Granelli, F., Gebremariam, A. A., Taleb, T., & Bagaa, M. (2017). End-to-end network slicing for 5G mobile networks. *Journal of Information Processing*, *25*, 153-163. https://www.jstage.jst.go.jp/article/ipsjjip/25/0/25_153/_article/-char/ja/
6. Albietz, B., & Assimakopoulos, D. G. (2014). Understanding organization-CRM System misfits and their evolution: A path to improving usage. In *Academy of Management Proceedings* (Vol. 2014, No. 1, p. 11174). Briarcliff Manor, NY 10510: Academy of Management. https://journals.aom.org/doi/abs/10.5465/ambpp.2014.11174abstract
7. Farag, Y., Yannakoudakis, H., & Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898*. https://arxiv.org/abs/1804.06898
8. Atri, P. (2018). Optimizing Financial Services Through Advanced Data Engineering: A Framework for Enhanced Efficiency and Customer Satisfaction. *International Journal of Science and Research (IJSR)*, *7*(12), 1593-1596. https://www.academia.edu/download/120680826/sr24422184930.pdf

9. Minhas, A. S., Singh, S., Malhotra, J., & Kumar, N. (2018). Machine deterioration identification for multiple nature of faults based on autoregressive-approximate entropy approach. *Life Cycle Reliability and Safety Engineering*, *7*(3), 185-192. https://link.springer.com/article/10.1007/s41872-018-0056-6
10. Patil, D. J., & Mason, H. (2015). *Data Driven*. " O'Reilly Media, Inc.". https://books.google.com/books?hl=en&lr=&id=yOFwBgAAQBAJ&oi=fnd&pg=PP1&dq=related:I2eZqrtgVZAJ:scholar.google.com/&ots=x35Rn3tWEg&sig=Wzelo6e39KqYJmh7obqoXn439GU