Optimizing Data Quality in Real-Time, Time Series ETL Pipelines Using Monte Carlo Methods and Machine Learning

Robert W. Bakyayita

Uganda Martyrs University Nkozi
Faculty of Science
Department of Computer Science and Information Systems
MSc. Information Systems

July 2025

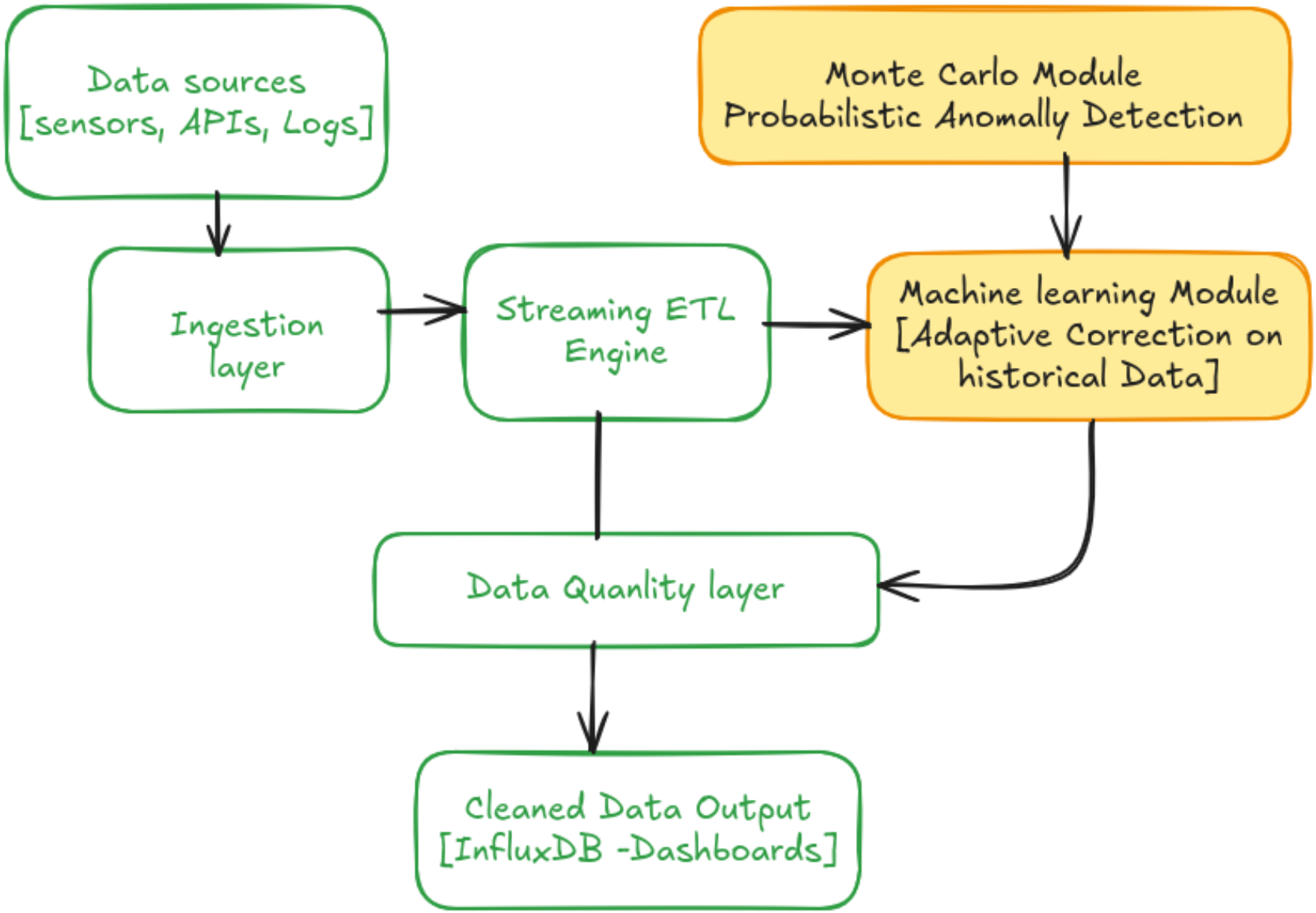
Outline

Background

Real-time ETL pipelines are critical for processing time-sensitive data in many industries. However, ensuring data quality is challenging due to the high velocity, volume, and variability inherent in time series data streams.

- Traditional batch-based data cleaning methods are often inadequate for real-time requirements.
- Monte Carlo methods offer a powerful way to model uncertainty in data.
- Machine learning techniques provide adaptive capabilities to detect and correct anomalies.
- Combining both approaches enables automated, probabilistic, and adaptive data quality control.

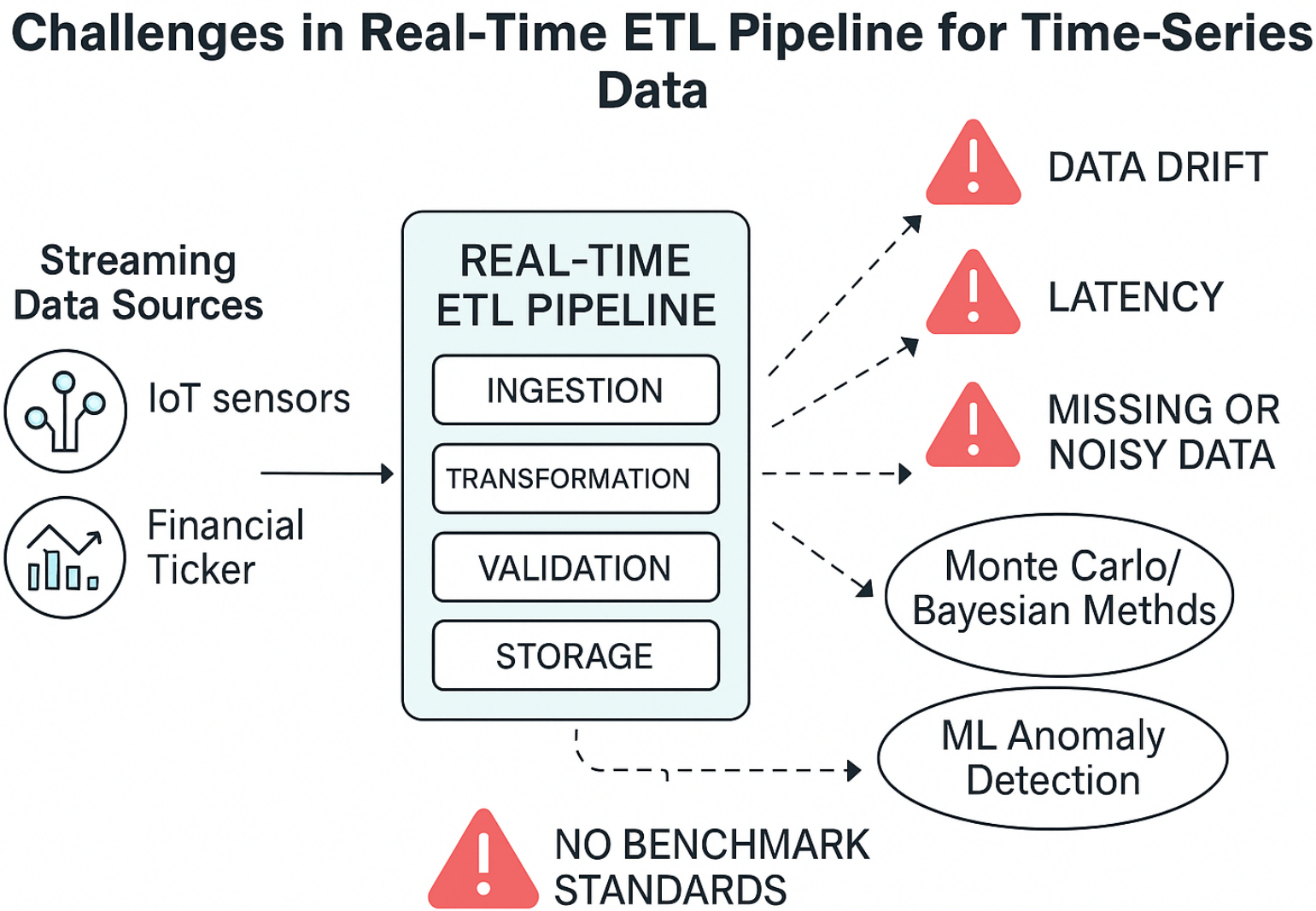
Real-Time ETL Pipeline: Conceptual Overview



Problem Statement

- Current data quality frameworks are not adequately designed to address the dynamic and continuous nature of real-time time-series ETL pipelines.
- Although Markov Chain Monte Carlo (MCMC) methods provide robust mechanisms for uncertainty modeling, their application in streaming data quality contexts remains limited (Brooks et al., 2011).
- Machine learning approaches capable of learning from historical anomaly correction patterns are underutilized in real-time systems.
- The field lacks standardized benchmarks and metrics for evaluating data quality in streaming ETL processes, impeding comparison and improvement of different techniques.

Challenges in Real-Time ETL Pipeline for Time-Series Data



Research Objectives

This research aims to enhance data quality in real-time time series ETL pipelines by pursuing the following objectives:

1. **Develop an MCMC-based module** to simulate realistic data corruptions and correct quality issues dynamically in real-time.
2. **Integrate machine learning models** that adaptively learn from historical anomaly corrections to improve future prediction accuracy.
3. **Design and implement a prototype** ETL pipeline that combines MCMC techniques with machine learning models for real-time data cleaning.
4. **Evaluate the effectiveness** of the proposed system using standard time series anomaly detection metrics and benchmark datasets.

Methodology

This study employs a hybrid methodology that integrates probabilistic modeling and machine learning within a real-time ETL framework.

- **Tools and Technologies:**
  - Data Ingestion: Apache NiFi
  - Storage: InfluxDB
  - Visualization: Grafana
  - Modeling: Python, PyMC3 (MCMC), Scikit-learn, TensorFlow
- **Step-by-Step Approach:**
  1. **Ingestion:** Streaming time-series data is ingested via Apache NiFi into InfluxDB.
  2. **Simulation:** Markov Chain Monte Carlo (MCMC) techniques are used to inject synthetic but realistic anomalies and simulate data corruptions.
  3. **Prediction:** Supervised ML models are trained using historical correction data to detect and predict anomalies.
  4. **Visualization:** Cleaned and corrupted data are visualized using Grafana dashboards via InfluxQL queries for monitoring and evaluation.

Key Findings

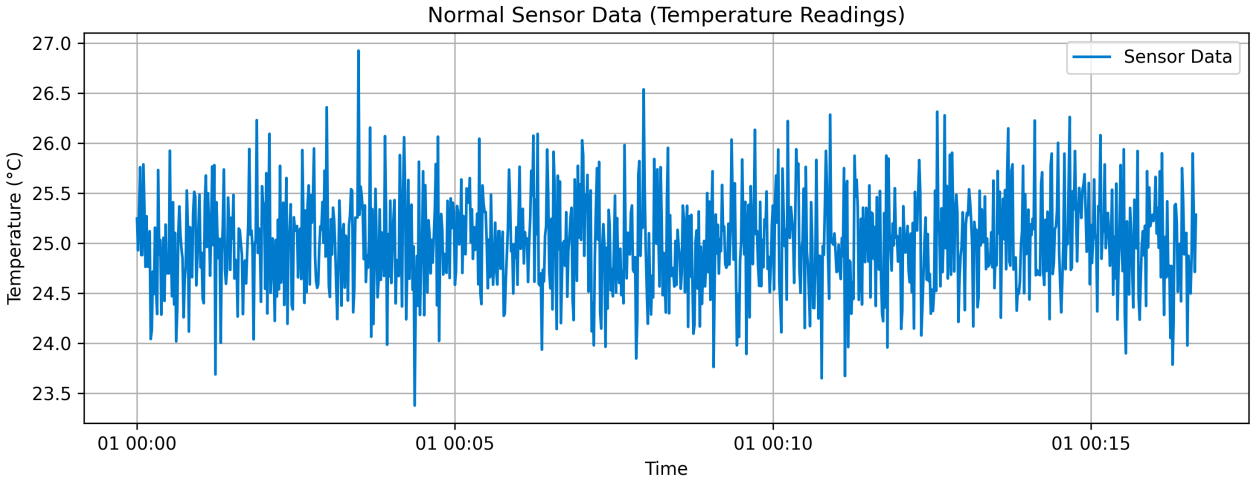
The study yielded several significant insights into improving real-time data quality in time series ETL pipelines:

- **Effectiveness of MCMC:** Markov Chain Monte Carlo (MCMC) techniques proved effective in simulating and identifying corrupted data points in real-time streams, enabling probabilistic error correction.
- **Learning from Historical Patterns:** Machine learning models that incorporated historical correction data showed enhanced accuracy in predicting future anomalies and deviations.
- **Hybrid Approach Superiority:** A combined approach integrating MCMC and machine learning outperformed traditional rule-based methods in terms of adaptability, precision, and robustness in dynamic environments.
- **Limitations in Existing Metrics:** Current data quality assessment metrics for real-time pipelines were found inadequate for capturing the nuances of streaming data, underscoring the need for new, standardized benchmarks.

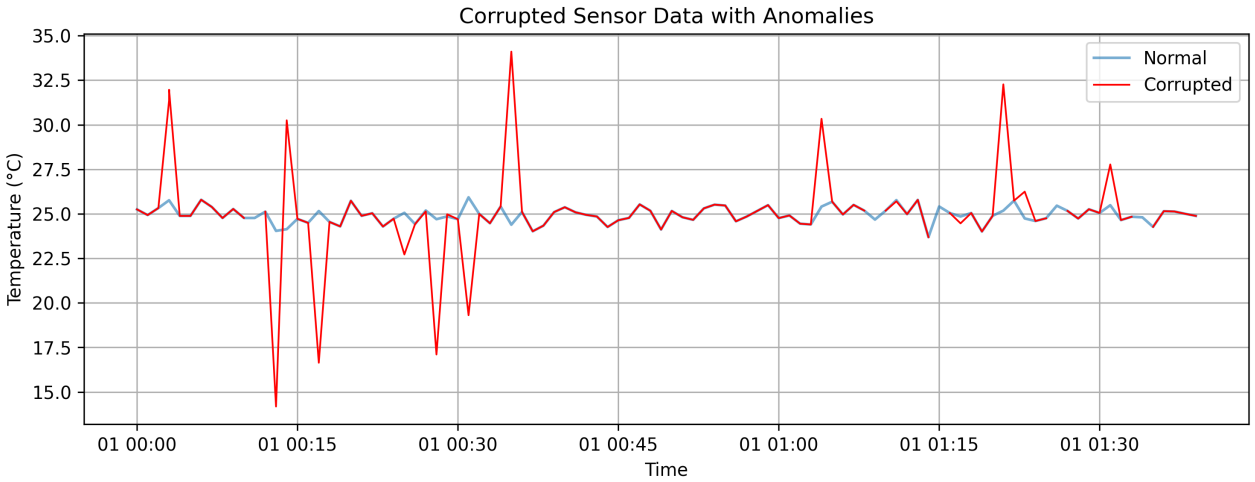
Contributions

- Designed and implemented a flexible, containerized framework integrating Monte Carlo Markov Chain (MCMC) and machine learning for enhancing data quality in real-time ETL pipelines.
- Developed an effective approach for detecting and correcting anomalies in streaming sensor data.
- Validated the system performance using realistic simulated sensor datasets.
- Contributed novel insights to research on data quality improvement in streaming data environments.

Simulation Results (1/2)

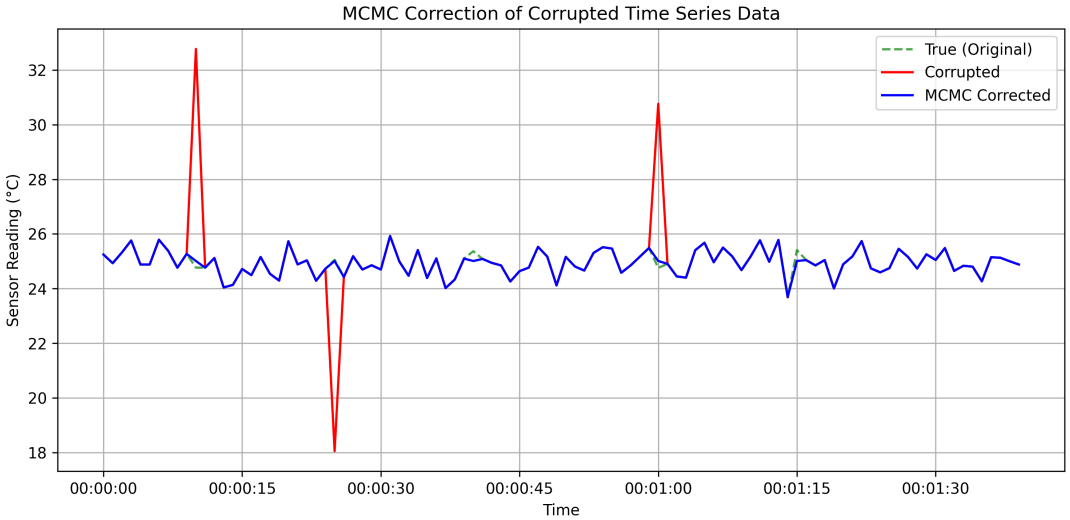


(a) Normal Data

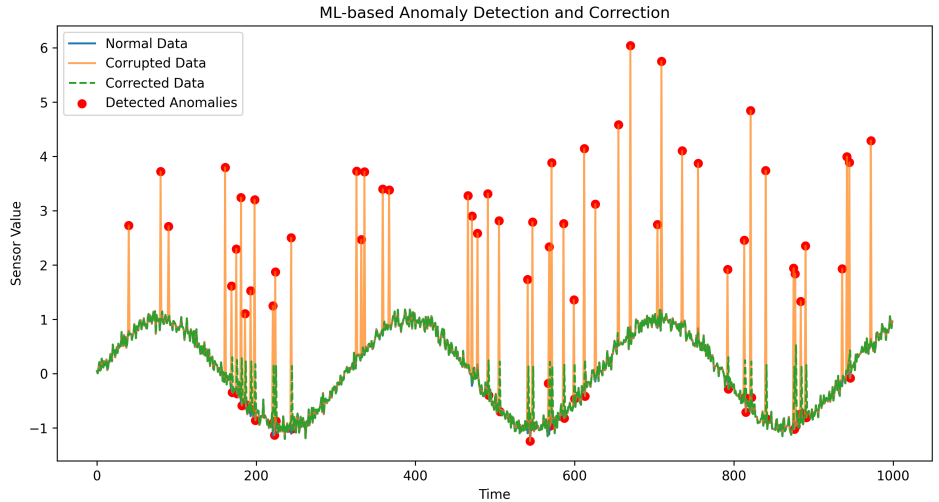


(b) Corrupted Data

Simulation Results (2/2)



(c) MCMC Output



(d) ML Correction

Conclusion and Future Work

Conclusion

- Demonstrated the effective use of MCMC and machine learning to enhance the quality of time series data in real-time ETL pipelines.
- Proposed a flexible framework tailored for streaming data environments.

Future Work

- Integrate federated learning techniques for decentralized and privacy-preserving anomaly detection.
- Extend system evaluation using real-world industrial data streams.
- Design and implement a standard benchmarking suite for real-time ETL pipeline performance assessment.

References I

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Robert, C. P. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer.