UDC 004.62
DOI https://doi.org/10.36994/2788-5518-2023-02-06-07

ENHANCING DISTRIBUTED STREAM PROCESSING THROUGH ARTIFICIAL INTELLIGENCE: A REVIEW AND PROPOSAL FOR LSTM INTEGRATION

Ilin M.O., *PhD student, Department of Computer Systems Software, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine*
https://orcid.org/0009-0001-0803-3726, max.ilin@yahoo.com

Oleshchenko L.M., *Candidate of Technical Sciences, Associate Professor at the Computer Systems Software Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine*
https://orcid.org/0000-0001-9908-7422, oleshchenkoliubov@gmail.com

Abstract. *The proliferation of Internet of Things (IoT) devices and sensors is generating massive volumes of real-time, heterogeneous data streams. Organizations need to be able to integrate and analyze these high-velocity, multi-formatted data to extract value. However, traditional batch analytics paradigms struggle with the volume and latency demands. This paper reviews a distributed stream data processing architecture designed to address these challenges. The proposed information system provides a pipeline for ingesting high-volume, heterogeneous data streams; automatically homogenizing the data into a consistent schema; and enabling Complex Event Processing (CEP) to detect situations of interest through real-time pattern matching. The feasibility of the architecture was demonstrated by implementation in a smart water management system to process and analyze real-time meter data. However, further enhancements using artificial intelligence (AI) technologies could improve the efficiency, accuracy, and adaptability of the system. Specifically, this paper hypothesizes integrating Long Short-Term Memory (LSTM) neural networks to add predictive analytics capabilities. The LSTM models would forecast expected values from historical data, identify deviations, and augment the stream events with predictions. This would enable the CEP engine to operate at a higher level detecting pattern anomalies rather than individual outliers. The AI integration is projected to increase the accuracy of anomaly detection based on more holistic data context. However, sizable research is still required to prototype, implement, and validate the fusion of stream processing with recurrent neural networks. This paper provides an initial perspective on the future convergence of big data pipelines and artificial intelligence to support rapidly evolving, mission-critical streaming applications in IoT domains. It lays the foundation for an intelligent distributed architecture that can handle heterogeneous data at scale.*

Key words: *software methods, heterogeneous streaming data, Python, data processing information systems, distributed computing, AI technologies, LSTM, CEP.*

ОСОБЛИВОСТІ ОБРОБКИ ПОТОКОВИХ ДАНИХ РОЗПОДІЛЕНОЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ: ОГЛЯД І ПОКРАЩЕННЯ ІСНУЮЧИХ ПІДХОДІВ ЗА ДОПОМОГОЮ ІНТЕГРАЦІЇ LSTM

Максим Ільїн, *аспірант кафедри програмного забезпечення комп’ютерних систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна*
https://orcid.org/0009-0001-0803-3726, max.ilin@yahoo.com

Любов Олещенко, *канд. техн. наук, доцент кафедри програмного забезпечення комп’ютерних систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна*
https://orcid.org/0000-0001-9908-7422, oleshchenkoliubov@gmail.com

Анотація. *Поширення пристроїв і датчиків інтернету речей сприяє генерації величезних об’ємів гетерогенних потоків даних у реальному часі. Організації повинні мати можливість інтегрувати й аналізувати потокові дані різного формату для отримання бізнес-цінності. Однак традиційні парадигми пакетної аналітики мають проблеми з об’ємом даних і затримкою в часі. У статті розглядається архітектура розподіленої потокової обробки даних, спрямована на вирішення цих проблем. Запропонована інформаційна система забезпечує конвеєр для приймання великого обсягу неоднорідних потоків даних, автоматичного упорядкування даних у послідовну схему, а також увімкнення комплексної обробки подій для виявлення специфічних ситуацій за допо-*

могою зіставлення шаблонів у реальному часі. Здійсненність архітектури було продемонстровано впровадженням в інтелектуальну систему управління водними ресурсами для обробки й аналізу даних лічильників у реальному часі. Подальші вдосконалення за допомогою технологій штучного інтелекту (ШІ) можуть підвищити ефективність, точність і адаптивність інформаційної системи. Зокрема, у статті висувається гіпотеза про інтеграцію нейронних мереж довготривалої короткочасної пам'яті (LSTM) для додавання можливостей прогнозної аналітики. Моделі LSTM дають змогу прогнозувати очікувані значення на основі історичних даних, визначати відхилення та доповнювати потоки подій прогнозами. Завдяки цьому механізм комплексної обробки подій працюватиме на вищому рівні, виявляючи типові аномалії, а не окремі викиди. Очікується, що інтеграція ШІ підвищить точність виявлення аномалій на основі більш цілісного контексту даних. На сьогодні відсутні дослідження для прототипування, реалізації та перевірки інтеграції рекурентних нейронних мереж у потокову обробку даних. У роботі представлено початковий погляд на майбутню конвергенцію конвеєрів великих даних і штучного інтелекту для підтримки критично важливих програм потокового передавання даних у доменах інтернету речей.

Ключові слова: *програмні методи, гетерогенні потокові дані, Python, інформаційні системи оброблення даних, розподілені обчислення, технології штучного інтелекту, LSTM, CEP.*

Introduction

Processing and analyzing real-time data streams is an increasingly important capability in domains ranging from IoT to finance. However, integrating and making sense of heterogeneous data formats poses challenges. In 2020 Corral-Plaza et al. proposed a novel stream processing architecture to address these issues. The paper [1] proposes a novel architecture for processing streaming data from diverse Internet of Things (IoT) devices. The architecture enables real-time data integration, filtering, and analytics for various applications, including predictive maintenance, quality control, and environmental monitoring. The authors evaluate the architecture using several case studies and demonstrate its effectiveness in handling complex data streams with varying levels of accuracy and latency.

Today, there are no studies for prototyping, implementation and verification of the integration of recurrent neural networks in streaming data processing, so this problem is relevant and requires further research.

This paper reviews distributed information system in detail, including its layers, components, and sample implementations. Building on this foundation, potential augmentations using artificial intelligence are explored to enhance real-time data analytics. Finally, conclusions are presented on the feasibility and projected improvements from integrating AI into stream processing systems.

Research

Today, distributed information systems have been developed for stream data processing of various applications. The research paper by Hu et al. (2016) proposes a information system for sensor data processing from multiple sources in real-time. The system is designed to handle heterogeneous sensor data and can be used in various applications such as environmental monitoring, health monitoring, and traffic management. The authors present a novel stream processing algorithm that can handle data from different sources with different data formats and characteristics [2]. The paper Montori F. et al. (2016) discusses the challenges and opportunities of integrating multiple data sources in the context of the Internet of Things (IoT). The authors propose a framework for integrating heterogeneous data sources, including sensor data, social media, and context-aware systems. They discuss the importance of data quality and security in this integration and highlight the potential benefits of collaborative data sharing for IoT applications [3]. The authors of the article by Nadal S. et al. in 2017 proposed a framework for building semantic-aware Big Data systems. The authors present a software reference architecture that integrates semantic technologies with Big Data processing to enable more efficient and effective data analysis. The architecture consists of several components, including a semantic data lake, a semantic metadata management system, and a set of semantic-aware data processing modules. The paper provides a detailed description of each component and their interactions, as well as use cases that demonstrate the effectiveness of the architecture [4]. The research paper by D'Silva et al. (2017) proposes a system for real-time processing of IoT events and historic data using Apache Kafka and Apache Spark with the Dashing framework. The system enables efficient processing and analysis of large datasets from IoT devices, enabling real-time insights and decision-making [5]. The article by Estévez-Ayres I. et al. (2017) proposes a novel infrastructure called Lostrego, designed to streamline the collection and analysis of diverse educational data in real-time. Lostrego adopts a distributed architecture, allowing for efficient processing and scalability, and supports diverse data sources, including learning management systems, student information systems, and online collaboration platforms. The research demonstrates the effectiveness of Lostrego through a series of case studies, showcasing its ability to provide real-time insights into educational processes and outcomes [6]. The research by Traore et al. (2016) explores the use of artificial neural networks (ANNs) for short-term forecasting of evapotranspiration (ET) using public weather forecast restricted messages. The authors use a dataset from the southeastern United States and demonstrate the potential of ANNs in improving ET forecasts. The study contributes to the development of accurate and reliable ET forecasting methods, which are essential for water resource management and crop planning [7]. The by Greff et al. (2017) investigates the search space of Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN) widely used in Natural Language Processing (NLP) tasks. The authors propose a novel approach to explore the vast search space of LSTM architectures, enabling the identification of optimal hyperparameters and evaluating the performance of LSTM models [8]. The paper by Karim F., Majumdar S., Darabi H., and Chen S. (2018) proposes a novel deep learning architecture that combines the strengths of Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) for time series classification. The authors introduce the LSTM-FCN model, which utilizes fully convolutional layers to improve the feature extraction process and reduce computational complexity. The paper provides a detailed evaluation of the LSTM-FCN model on several benchmark datasets, demonstrating its superior performance compared to traditional time series classification methods [9].

Corral-Plaza et al. proposed an architecture for processing and analyzing heterogeneous data streams in real-time from Internet of Things (IoT) devices.

The information system is organized into three layers: data producers, data processing and data consumers.

On the data producers layer IoT devices and sensors generate heterogeneous data streams. The data can be structured (e.g. JSON, XML) or unstructured (e.g. CSV files).

Data processing layer contains two main components – a stream processing platform (e.g. Kafka Streams) and a complex event processing (CEP) engine (e.g. Esper). The stream processing platform consumes the heterogeneous data streams, homogenizes the data into a common format, and generates simple events. The CEP engine analyzes the simple events to detect situations of interest.

On the data consumers layer endpoints like databases, smartphones, or web services consume the processed data and detected situations from the system.

Architecture

The data processing layer contains the main components for homogenizing and analyzing the heterogeneous data streams:

Kafka Cluster

A Kafka cluster stores the incoming heterogeneous data streams in Kafka topics.

Data producers publish messages to input topics, while consumers can subscribe to output topics (Fig. 1). Kafka provides high throughput and scalability [10].

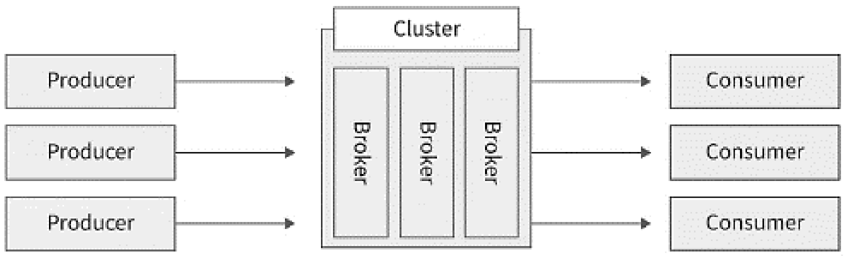

Fig. 1. Kafka cluster structure

In the water management implementation, there are:

- an input topic for sensor data from smart water meters;
- an output topic for detected anomalies and alerts.

Kafka Streams Application

The Kafka Streams application consumes messages from the input topic, homogenizes the data, and generates simple events for analysis (Fig. 2).

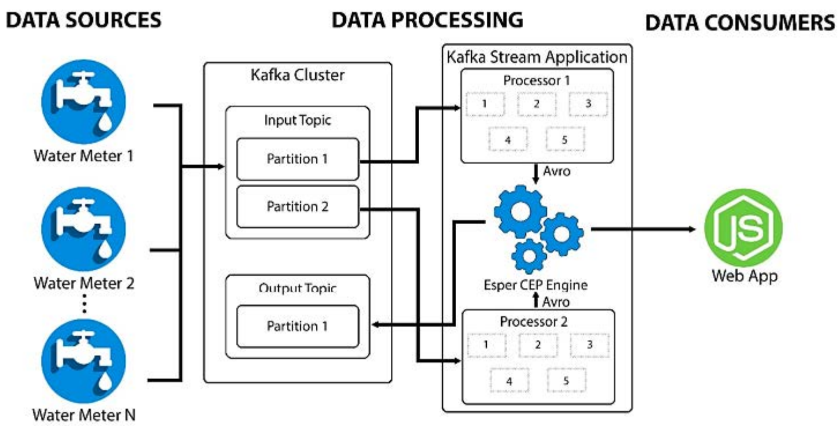

Fig. 2. Software architecture adapted to the water management scenario [1]

Key steps to use the information system are next.
1. Consume data from the input topic partitions.
2. Detect data format (JSON, XML, CSV).
3. Homogenize data into a common JSON format.
4. Generate Avro schema for the data.
5. Create Avro simple events for analysis.

Esper CEP Engine

The Esper CEP engine receives the Avro simple events from Kafka Streams. It allows defining EPL patterns that match event conditions and detect situations of interest. When a pattern matches, a new complex event is generated and published to the output topic. For example, anomalies like water meter reading errors can trigger alerts. New patterns can also be deployed at runtime without stopping the system.

Implementation Example

Corral-Plaza et al. implemented the architecture for a smart water supply network. Heterogeneous data from smart water meters was processed and analyzed. For example, heterogeneous sensor data from smart water meters is consumed and homogenized into a JSON format with a schema (Table 1). This generates simple "WaterMeasurement" events for analysis.

Table 1

Sample sensor data in JSON format

```
{
  "serialNumber": "A87SAA77188",
  "dateTime": "2018-07-06T05:00:00.000Z",
  "volumeM3": 5.4366397,
  ...
}
```

The Kafka Streams application homogenizes the sensor data into Avro "WaterMeasurement" events (Table 2).

Table 2

Avro "WaterMeasurement" events in Java

```
class HomogenizeSensorData {
  @StreamListener("input-topic")
  @SendTo("cep-topic")
  public WaterMeasurement process(String message) {
    // Detect format
    if (isJSON(message)) {
      // Deserialize JSON
    } else if (isCSV(message)) {
      // Parse CSV
    }
    // Homogenize and create Avro event
    WaterMeasurement event = ...;
        return event;
  }
}
```

The Esper CEP engine detects anomalies using EPL patterns (Table 3).

Table 3

Example of an EPL pattern, SQL

```
INSERT INTO WaterMeterAnomaly
SELECT *
FROM WaterMeasurement
WHERE volumeM3 < 0
```

Overall, the architecture provided real-time processing of heterogeneous data streams in a water management IoT system.

The main advantages of the architecture are:
- real-time processing of high volumes of heterogeneous data streams;
- automatic homogenization of data into a common format;
- real-time analysis and detection of situations of interest;
- high scalability by adding more stream processing nodes.

Proposed AI Integration Method

Long Short-Term Memory (LSTM) neural networks are a type of recurrent neural network (RNN) that have proven to be effective for processing and forecasting streaming data, particularly in time series analysis. LSTMs are designed to handle sequences of data. They are capable of processing data points one at a time in a sequential manner, making them well-suited for streaming data where the order of the data points matters. LSTMs can capture long-term dependencies in the data, which is important for understanding patterns and trends in time series data. Unlike traditional feedforward neural networks, LSTMs can remember information from much earlier in the sequence. LSTMs can be adapted to work with streaming data in an online learning setup. As new data points arrive, the LSTM can be updated and retrained in real-time, allowing it to adapt to changing patterns and trends. LSTMs can handle sequences of varying lengths. In a streaming data scenario, the length of the data sequence may change over time, and LSTMs are flexible enough to handle this. LSTMs can be used in a stateful mode, where the internal state is carried over from one batch of data to the next. This is particularly useful for streaming data, as it allows the network to maintain context across multiple time steps. LSTMs can automatically learn relevant features from the data, reducing the need for manual feature engineering. This is beneficial when dealing with streaming data, where the data distribution may change over time. LSTMs can handle multivariate time series data, which is common in many streaming data applications. They can model dependencies between multiple variables simultaneously. LSTMs can be used for anomaly detection in streaming data. By modeling the normal behavior of a system, they can detect deviations from this normal behavior in real-time. LSTMs are often used for time series forecasting. They can be trained to predict future data points in a sequence, which is valuable for making predictions in streaming data applications. LSTMs can be integrated into systems for real-time decision making. For example, in finance, they can be used to make trading decisions based on streaming market data.

The architecture of the recurrent neural network is shown in Fig. 3.

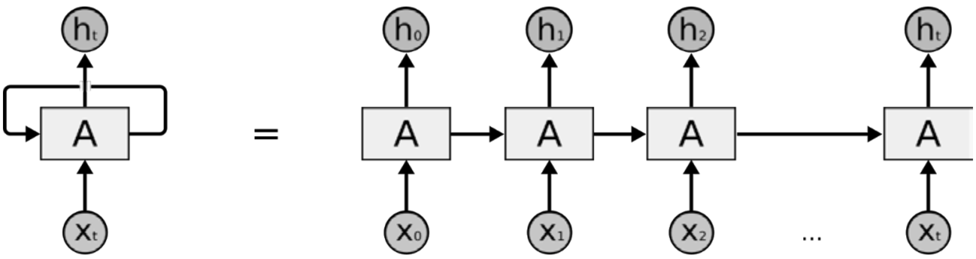

Fig. 3. Recurrent neural network in sweep

This network is particularly adept at forecasting time series data due to its ability to factor in past experience. The key distinction between LSTM and traditional RNN lies in the inclusion of specialized nodes dedicated to retaining information over extended time intervals. This is achieved by these nodes abstaining from using the activation function, ensuring that data remains unchanged as it's stored for future reference.

LSTM takes the form of a chain of recurrent neural network modules. Fig. 4 shows an LSTM structure that has four hidden layers and uses a sigmoid function and a hyperbolic tangent as the activation function.

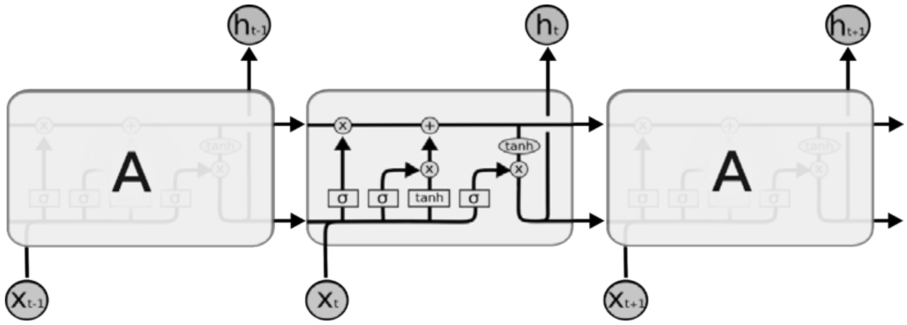

Fig. 4. LSTM structure

The neural network is trained using backpropagation in time, for which iterative gradient descent is used, which changes each weight coefficient in proportion to its derivative with respect to the error. Traditional LSTM with forgetting gates $c_0 = 0$, $h_0 = 0$:

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (1)$$

$$i_t = \sigma_g\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{2}$$

$$o_t = \sigma_g\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{3}$$

$$c_t = f_t c_{t-1} + i_t \sigma_c\left(W_c x_i + U_c h_{t-1} + b_c\right) \tag{4}$$

$$h_t = o_t \sigma_h\left(c_t\right) \tag{5}$$

Variables: x_t is input vector; h_t – output vector; c_t – vector of states; W, U, and b are parameter matrices and a vector; f_t, i_t and o_t – gate vectors: f_t – the forgetting gate, the weight of remembering old information, i_t – the input gate, the weight of receiving new information, o_t – the output gate, the candidate for output (Fig. 5). Activation functions: σ_g – based on the sigmoid, σ_c – based on the hyperbolic tangent, σ_h – based on the hyperbolic tangent, in this work it is assumed that $\sigma_h(x) = x$.

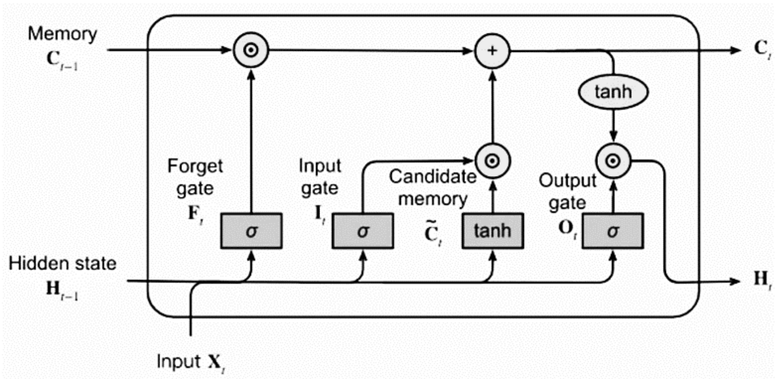

Fig. 5. A simple LSTM block with three gates: input, output and forget

To enable more efficient and intelligent processing, a predictive analytics module based on LSTM neural networks could be added between the Kafka Streams application and the Esper CEP engine [11].

The LSTM model would take the homogenized data streams from Kafka Streams as input and make multi-step predictions about expected values. For example, in the water management use case it could forecast anticipated consumption and detect when actual readings diverge significantly from the prediction.

The LSTM module would have the following components:

- Data preparation – the Avro event streams are formatted as sequences for time series prediction.
- LSTM model – Long Short-Term Memory network is trained on historical data to make multi-step forward predictions.
- Anomaly detection – errors between predicted values and actual events are calculated. Significant deviations are flagged as anomalies.
- Integration – the predicted values and anomalies are appended to the Avro events emitted to the CEP engine.

Table 4

Pseudocode for proposed AI integration

```
DataPreparation(AvroStream):
   format avro data as time series
 LSTMModel(HistoricalData):
   train LSTM network
   return model
 Predict(Model, AvroStream):
   make multi-step predictions
   calculate prediction errors
   flag anomalies
 AugmentAvroStream(AvroStream, Predictions, Anomalies):
   add predictions and anomalies to Avro events
   return AugmentedStream
```

Adding predictive abilities would allow the CEP engine to operate on a higher level, detecting pattern anomalies rather than individual outliers. This is expected to improve the accuracy of anomaly detection.

The LSTM module would run continuously alongside Kafka Streams and incrementally update its model on new data. This would allow the predictions to adapt to evolving conditions. It's important to note that while LSTMs have many advantages for processing and forecasting streaming data, they are not without challenges. They can be computationally expensive, and training them with large datasets in real-time may require specialized hardware or distributed computing. Additionally, choosing the right architecture and hyperparameters for specific streaming data application is crucial to achieve the best performance (Fig. 6).

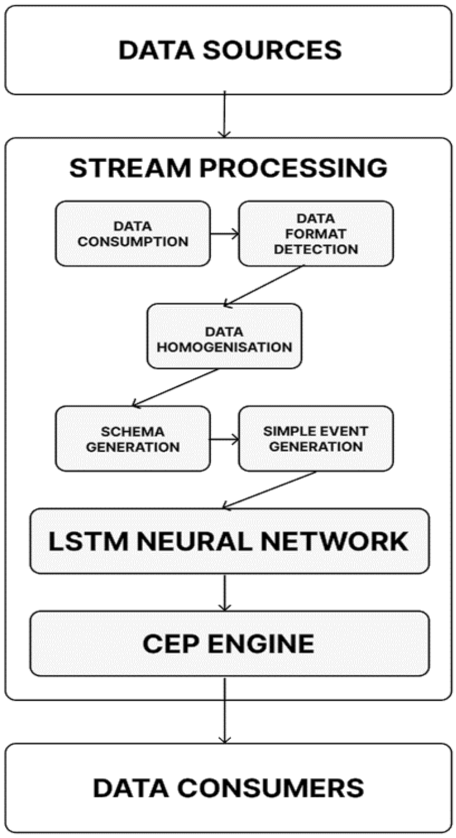

Fig. 6. Proposed architecture enhanced with AI usage

Overall, integrating recurrent neural networks could significantly improve the intelligence of the distributed stream processing system. While requiring further research, this AI-based approach has potential for making heterogeneous data analytics more efficient.

Conclusions

In this paper, we reviewed a distributed system stream processing system architecture for handling heterogeneous data that provides an effective pipeline for ingesting high-volume, real-time streams of heterogeneous data formats, automatically homogenizing the data into a consistent schema and enabling complex event processing for real-time pattern detection and alerts.

The feasibility of the approach was demonstrated through an implementation in a smart water management IoT system. However, further enhancements using artificial intelligence could improve the efficiency, accuracy, and adaptability of the system.

Specifically, we proposed integrating LSTM neural networks to add predictive analytics capabilities. By forecasting expected values and preemptively identifying anomalies, this AI integration could potentially improve the anomaly detection accuracy. While promising, this AI approach requires further research into production deployments, model training requirements, and integration architectures. As next steps, a proof-of-concept prototype should be implemented to quantify the concrete benefits for real-world streaming applications.

In the long-term, blending stream processing with artificial intelligence will be essential to handling accelerating volumes of heterogeneous and complex data. This paper presented initial perspectives on accomplishing that fusion through a distributed, real-time analytics pipeline augmented with machine learning. There remains ample open challenges and opportunities for future innovations in this domain.

References

1. Corral-Plaza, D., Medina-Bulo, I., Ortiz, G., & Boubeta-Puig, J. (2020). A stream processing architecture for heterogeneous data sources in the Internet of Things. *Computer Standards & Interfaces*. 70 (2020) 103426. DOI: 10.1016/j.csi.2020.103426.

2. Hu, L., Sun, R., Wang, F., Fei, X., Zhao, K. (2016). A stream processing system for multisource heterogeneous sensor data, *Sens.*, https://doi.org/10.1155/2016/4287834.

3. Montori, F., Bedogni, L., Bononi, L. (2016). On the integration of heterogeneous data sources for the collaborative Internet of Things, *IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow (RTSI)*, 2016, pp. 1–6, https://doi.org/10.1109/RTSI.2016.7740616.

4. Nadal, S., Herrero, V., Romero, O., Abelló, A., Franch, X., Vansummeren, S., Valerio, D. (2017). A software reference architecture for semantic-aware Big Data systems, *Inf. Softw. Technol.* 90. 75–92, https://doi.org/10.1016/j.infsof.2017.06.001.

5. D'silva, G.M., Khan, A., Gaurav, S.B. (2017). Real-time processing of IoT events with historic data using apache Kafka and Apache Spark with dashing framework, *2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pp. 1804–1809, https://doi.org/10.1109/RTEICT.2017.8256910.

6. Estévez-Ayres, I., Arias Fisteus, J., Delgado-Kloos, C. (2017). Lostrego: a distributed streambased infrastructure for the real-time gathering and analysis of heterogeneous educational data, *Netw. Comput. Appl.* 100, pp. 56–68, https://doi.org/10.1016/j.jnca.2017.10.014.

7. Traore, S., Luo, Y.F., Fipps, G. (2016). Deployment of artificial neural network for short-term forecasting of evapotranspiration using public weather forecast restricted messages. *Agricultural Water Management* 163 (1), 363–379. https://doi.org/10.1016/j.agwat.2015.10.009.

8. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R. and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, doi: 10.1109/TNNLS.2016.2582924.

9. Karim, F., Majumdar, S., Darabi, H. and Chen, S. (2018). LSTM Fully Convolutional Networks for Time Series Classification, *IEEE Access*, vol. 6, pp. 1662–1669, doi: 10.1109/ACCESS.2017.2779939.

10. Stojanović Natalija, and Dragan Stojanović. (2020). Big mobility data analytics for traffic monitoring and control. *Facta Universitatis, Series: Automatic Control and Robotics* 19, no. 2: 087. http://dx.doi.org/10.22190/fuacr2002087s.

11. Cheng H., Xie Z., Wu L. et al. (2019). Data prediction model in wireless sensor networks based on bidirectional LSTM. *Wireless Com Network* 2019, 203. https://doi.org/10.1186/s13638-019-1511-4.