ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Review Article

Open Access

Threat Detection Using Real-Time Data Engineering Pipelines

Girish Ganachari

USA

ABSTRACT

Due to rising cyber threats and data quantities, existing systems need better threat detection. Real-time data engineering pipelines boost threat detection in this research. This study uses 2016–2020 literature data to examine threat detection methodologies, tools, and models. Results suggest that data security, integration, and standardisation must be addressed before broad use of these pipelines, despite their obvious advantages. Machine learning, blockchain, and IoT are promising.

***Corresponding author**

Girish Ganachari, USA.

Received: January 05, 2023; **Accepted:** January 15, 2023; **Published:** January 25, 2023

Keywords: Real-time Data Pipelines, Threat Detection, Machine learning, Cybersecurity, Data Engineering

Introduction

Cyber threats evolve, requiring new detection and control methods. Traditional threat detection methods fail due to digital information's volume and complexity. Automating data collection, processing, and analysis using real-time data engineering pipelines detects irregularities and threats early. These pipelines outperform batch-based systems in security and reaction speed. This research examines academic secondary data-based real-time data engineering pipelines for threat detection and its pros and cons. The study discusses data security, integration, costs, and standards. Additionally, detection accuracy, operational efficiency, scalability, and flexibility improve. This lengthy analysis highlights cybersecurity's need for real-time data pipelines.

Applications of Real-Time Data Pipelines

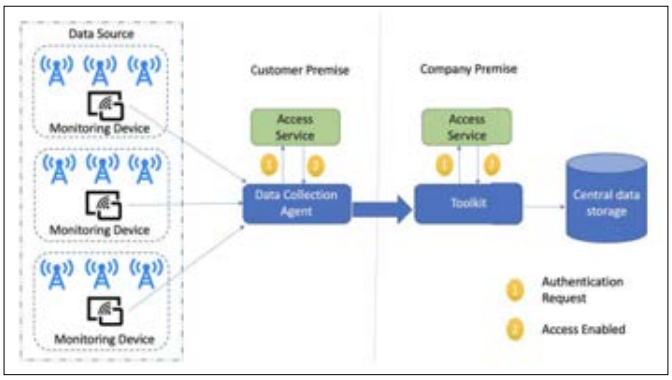

Figure 1: Data Collection Process

Real-time data pipelines increase operational efficiency and aid risk detection in many applications. Pipelines analyse and respond quickly to continuous data streams [1]. Maintaining operational stability and security in changing environments needs this. These sections discuss successful real-time data pipeline applications.

Real-Time Anomaly Detection

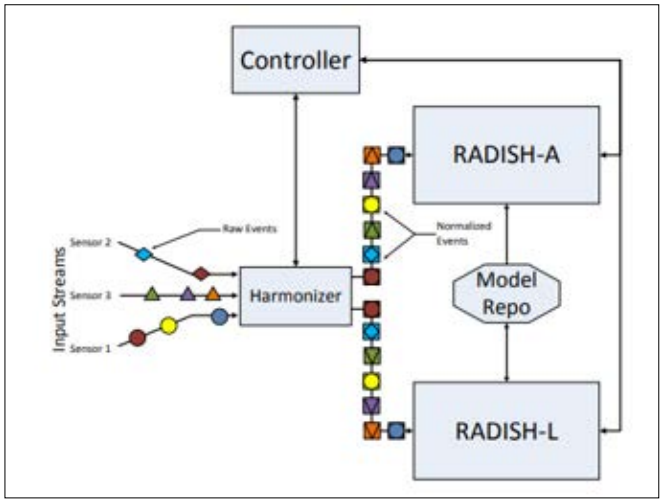

Figure 2: The RADISH System Architecture Normalized Streams

The majority of anomaly detection systems make advantage of real-time data pipelines. RADISH by Böse et al. detects real-time irregularities in different data sources. This technique detects insider threats using machine learning and multiple data sources [2]. The RADISH system illustrates how real-time processing can notify and react to irregularities to increase security. These pipelines use machine learning to learn from past data and adapt to new threats, enhancing system effectiveness.

Data-Driven Machine Learning Algorithms

Another major use is data-driven machine learning predictive analytics. Predictive analytics may minimise pipeline failure risk and increase data pipeline dependability and security, according to Mazumder et al. [3]. By evaluating data streams, machine learning models may predict errors. Preventive maintenance reduces system downtime. Eliminating costly disruptions improves infrastructure safety and efficiency.

Cyberthreat Detection from Social Media

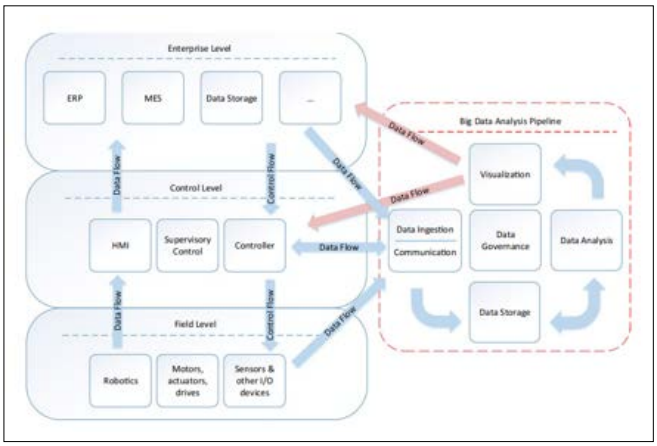

Figure 3: Control and Data Fow Between the Manufacturing Layers and A Big Data Analysis Pipeline

Identifying cyberattacks on social media platforms requires the use of real-time data pipelines. Dionísio et al. used Twitter data to identify cyber risks using deep neural networks [4]. This approach analyses social media streams for suspicious activities or risks using data pipelines in real time. These systems evaluate massive data sets in real time to alert organisations of intrusions. Modern cybersecurity uses social media data to analyse data in real time for threat detection.

Automated Security Risk Identification

Eckhart et al. [5] leverage engineering data to find security risks using Automation ML [8]. Their technology automatically detects engineering data flaws, boosting threat detection. Real-time data pipelines monitor engineering systems, find security issues, and provide timely alerts. Automating security risk identification speeds up and increases threat detection while reducing human activity.

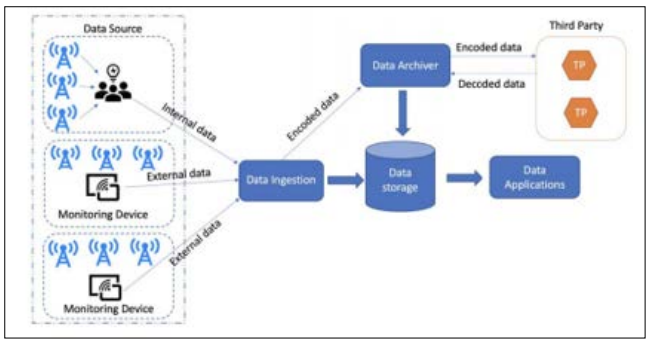

Figure 4: Data Pipeline for Data Governance

Real-time data pipelines in physical infrastructure enable advanced pipeline inspection. Instrumentation and data management are used by Ma et al. [6] to study pipeline corrosion and incursion. These pipeline inspection methods use real-time data analytics to find problems early for rapid repairs. This software illustrates how real-time data pipelines safeguard critical infrastructure.

Efficient Cyber Threat Hunting

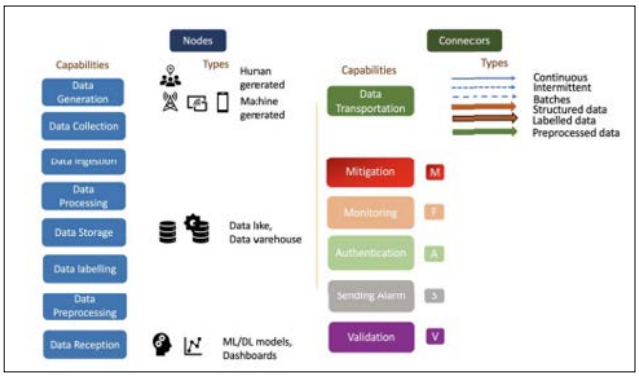

Figure 5: Meta-Model for Building Data Pipeline

Gao et al. advocate real-time data analytics and cyber threat information for threat hunting [7]. Their real-time threat intelligence-data analysis increases cyber threat identification and minimisation. Real-time data stream analysis and threat intelligence may help companies identify and handle risks. This alliance improves cybersecurity and speeds up threat detection.

Benefits of Real-Time Data Pipelines

Sector-wide real-time data pipelines boost operational efficiency and risk detection. These pipelines allow continuous data collection, processing, and analysis, enabling organisations fulfil tight operational efficiency and security criteria.

Enhanced Detection Accuracy

The detection accuracy of real-time data pipelines may be improved by the use of machine learning. These algorithms analyse enormous amounts of data in real time to find complex patterns and abnormalities that may be dangerous [4]. Machine learning algorithms can accurately detect good and bad activities due to past data training. Early threat identification in cybersecurity may prevent major damage. Machine learning algorithms may spot subtle signs of sophisticated cyberattacks or insider threats that traditional methods miss. Models that learn and adapt to new threats increase threat detection and provide organisations proactive cyberdefense [8].

Improved Operational Efficiency

Real-time data pipelines boost efficiency in various industries. These pipelines easily move data across operational components, providing real-time operational insights. Continuous data flow optimises resource utilisation, costs, and procedures [4]. Real-time data streams may uncover inefficiencies, equipment failures, and bottlenecks in manufacturing production processes. Manufacturers may improve productivity and cost by swiftly resolving these issues and reducing downtime. Real-time data pipelines help banks detect fraud and monitor transactions, lowering losses and following regulations [3].

Scalability and Flexibility

Real-time data pipelines are popular and effective due to their scalability and flexibility. These pipelines may be customised for organisational demands and data environments since they handle a variety of data types and amounts. Scalability needs automation to reduce human mistake and preserve pipeline data. During high data loads, automated data pipelines may adjust capacity [9]. This promotes reliability and efficiency. Because of their versatility, these pipelines may interface to unstructured social media feeds and organised databases. Organisations that handle several data sources and integrate them into an analytical framework must

adapt. Real-time data pipelines may assess urban dynamics using environmental monitors, public transportation systems, and traffic sensors in smart cities. This helps distribute resources and make informed judgements [10].

Real-Time Response

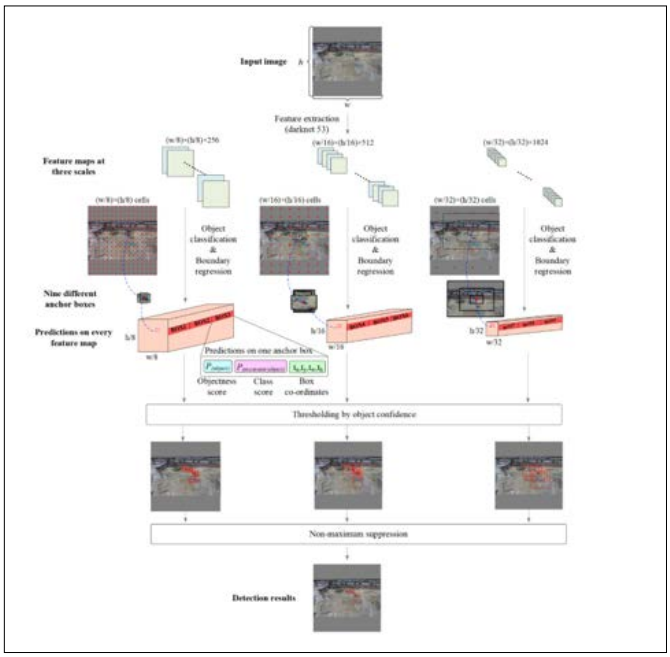

Figure 6: Flowchart of the Excavator Detection Model

Real-time data pipelines are optimal for speedy analysis and response. With this information, businesses can swiftly detect and remove hazards to reduce risk and secure sensitive data. Real-time data analysis reduces vulnerability exploitation by expediting attack response [11]. Monitoring network traffic with real-time data pipelines helps detect illegal access and data theft. Real-time alerts help security teams resolve problems before they worsen. Because timely data may improve patient outcomes, healthcare requires real-time action. Live vital sign monitoring may alert clinicians to critical patient changes, enabling fast intervention and perhaps life-saving therapy [12].

Challenges in Implementing Real-Time Data Pipelines

Real-time data pipeline installation must overcome challenges to ensure system reliability. Issues include infrastructure, prices, data standards, system integration, and security.

Data Security and Privacy

Live data pipelines pose privacy and security issues during processing. Real-time systems handle sensitive data, making them hackable. Security and access control are crucial for sensitive data. Protecting data at rest and in transit requires strong encryption [13]. Limit data access to permitted users to minimise internal hazards. Continuous data pipeline monitoring is necessary. Anomalies in real-time data flow monitoring may indicate a security vulnerability. Complex setups with varying system data may make these security processes challenging to perform. Companies struggle to reconcile data flow speed and security. Need a good balance.

Integration with Existing Systems

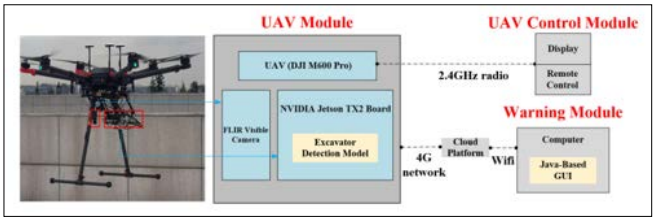

Figure 7: Architecture of the Developed UAV-ED System.

Real-time data streams are challenging to integrate into existing systems because of their complexity. Many firms employ old, non-real-time data processing. Integration of outdated systems with real-time capabilities may be complex and time-consuming [14]. Integration requires fixing data flow and system compatibility issues. However, customised solutions and large infrastructure improvements may be costly and resource-intensive. Without interruption or corruption, data must travel between new and old systems for smooth integration. Comprehensive planning and testing must identify and fix integration issues. Integration is tougher with several data sources and formats, each with its own requirements [9].

Cost and Infrastructure

Starting and monitoring real-time data pipelines may be expensive for smaller companies with limited resources. It is not inexpensive to purchase servers, storage, and networks that have a high-end setup [25]. Real-time data pipelines need specialist software and qualified system administrators and maintainers. New hires or staff training may increase pricing. In addition to setup, a real-time data pipeline needs system upgrades, monitoring, and security. Smaller organisations may struggle to justify these costs if real-time data processing benefits require time. Thus, financial constraints may prevent real-time data pipeline adoption [9].

Data Standardization

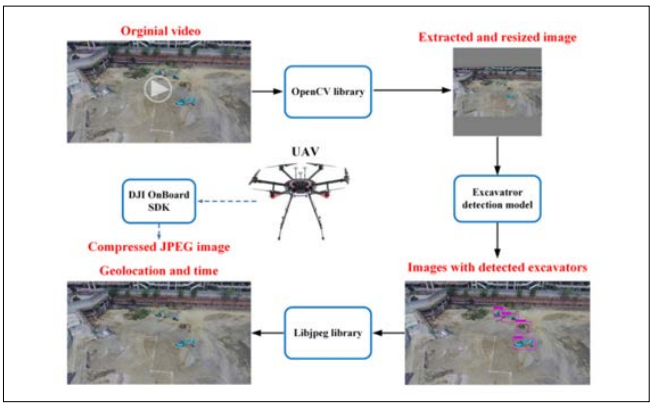

Figure 8: Data-processing Flow in the UAV Module

Standardising data formats and protocols helps real-time data pipelines manage data from several sources [3]. The criteria may be difficult to satisfy. Many sources provide data in different formats and standards to organisations. Data must be standardised before integration and analysis, which might hinder the process. Conversion delays and errors may diminish real-time data processing benefits. Data standardisation needs cooperation from parties with various agendas and standards. Large companies or those with uneven data standards may have problems. For many organisations, creating and implementing data formats and standards requires collaboration [3].

Future Directions

Their future is bright as tech improves real-time threat detection data pipelines. The convergence of wearable technologies, IoT, blockchain for data security, AI and ML integration, and system interoperability will drive future progress. By continually collecting real-time data, AI and ML may enhance predictive analytics, identify hidden patterns, and detect threats [13]. Healthcare and other businesses need immediate replies and constant monitoring from wearable and IoT devices [15]. Blockchain provides secure, immutable data exchanges and traceability. These qualities are essential for regulatory compliance and cooperative threat detection [16]. Protocols and defined data formats enable systems communicate, enabling data transfer for analysis and streamlined operations [17]. These advances will make real-time data pipelines crucial for cybersecurity and operational management [18-25].

Conclusion

Real-time data engineering pipelines automate data gathering, processing, and analysis, enhancing security and management. Threat detection nowadays requires these pipelines. Automation reduces risk by detecting irregularities quickly. Benefits surpass data format standardisation, security, and system integration. Scalable and adaptable pipelines enhance detection accuracy and operating efficiency. Future research should develop advanced machine learning models and scale to handle more data. Data security and operational efficiency will need real-time data pipelines as blockchain, AI, IoT, and interoperability technologies advance.

References

1. Habeeb RAA, Nasaruddin F, Gani A, Hashem IAT, Ahmed E, et al. (2019) Real-time big data processing for anomaly detection: A survey. International Journal of Information Management 45: 289-307.
2. Böse B, Avasarala B, Tirthapura S, Chung YY, Steiner D (2017) Detecting insider threats using radish: A system for real-time anomaly detection in heterogeneous data streams. IEEE Systems Journal 11: 471-482.
3. Mazumder RK, Salman AM, Li Y (2021) Failure risk analysis of pipelines using data-driven machine learning algorithms. Structural safety 89: 102047.
4. Dionísio N, Alves F, Ferreira PM, Bessani A (2019) Cyberthreat detection from twitter using deep neural networks. In 2019 international joint conference on neural networks (IJCNN) IEEE 1-8.
5. Eckhart M, Ekelhart A, Weippl E (2020) Automated security risk identification using AutomationML-based engineering data. IEEE Transactions on Dependable and Secure Computing, 19: 1655-1672.
6. Ma Q, Tian G, Zeng Y, Li R, Song H, et al. (2021) Pipeline in-line inspection method, instrumentation and data management. Sensors 21: 3862.
7. Gao P, Shao F, Liu X, Xiao X, Qin Z, et al. (2021) April. Enabling efficient cyber threat hunting with cyber threat intelligence. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) IEEE 193-204.
8. Tejedor J, Macias-Guarasa J, Martins HF, Pastor-Graells J, Corredera P, et al. (2017) Machine learning methods for pipeline surveillance systems based on distributed acoustic sensing: A review. Applied Sciences 7: 841.
9. Peng Z, Jian J, Wen H, Gribok A, Wang M, et al. (2020) Distributed fiber sensor and machine learning data analytics for pipeline protection against extrinsic intrusions and intrinsic corrosions. Optics Express 28: 27277-27292.
10. Raj A, Bosch J, Olsson HH, Wang TJ (2020) August. Modelling data pipelines. In 2020 46th Euromicro conference on software engineering and advanced applications (SEAA) IEEE 13-20.
11. Rangnau T, Buijtenen RV, Fransen F, Turkmen F (2020) Continuous security testing: A case study on integrating dynamic security testing tools in ci/cd pipelines. In 2020 IEEE 24th International Enterprise Distributed Object Computing Conference (EDOC) IEEE 145-154.
12. Ismail A, Truong HL, Kastner W (2019) Manufacturing process data analysis pipelines: a requirements analysis and survey. Journal of Big Data 6: 1-26.
13. Gao P, Xiao X, Li D, Li Z, Jee K, et al. (2018) {SAQL}: A stream-based query system for {Real-Time} abnormal system behavior detection. In 27th USENIX Security Symposium (USENIX Security 18) pp 639-656.
14. Morfino V, Rampone S (2020) Towards near-real-time intrusion detection for IoT devices using supervised learning and apache spark. Electronics 9: 444.
15. Meng L, Peng Z, Zhou J, Zhang J, Lu Z, et al. (2020) Real-time detection of ground objects based on unmanned aerial vehicle remote sensing with deep learning: Application in excavator detection for pipeline safety. Remote Sensing 12: 182.
16. Barker JLP, Macleod CJ (2019) Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities. Environmental modelling & software 115: 213-227.
17. Hafsa M, Jemili F (2018) Comparative study between big data analysis techniques in intrusion detection. Big Data and Cognitive Computing 3: 1.
18. Khan F, Yarveisy R, Abbassi R (2021) Risk-based pipeline integrity management: A road map for the resilient pipelines. Journal of Pipeline Science and Engineering 1: 74-87.
19. Senthivel S, Dhungana S, Yoo H, Ahmed I, Roussev V (2018) March. Denial of engineering operations attacks in industrial control systems. In Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy 319-329.
20. Ardagna CA, Bellandi V, Ceravolo P, Damiani E, Bezzi M, et al. (2017) A model-driven methodology for big data analytics-as-a-service. In 2017 IEEE international congress on big data (BigData Congress) IEEE 105-112.
21. Fedushko S, Ustyianovych T, Gregus M (2020) Real-time high-load infrastructure transaction status output prediction using operational intelligence and big data technologies. Electronics 9: 668.
22. Zhou M, Yang Y, Xu Y, Hu Y, Cai Y, et al. (2021) A pipeline leak detection and localization approach based on ensemble TL1DCNN. IEEE Access 9: 47565-47578.
23. Therrien JD, Nicolaï N, Vanrolleghem PA (2020) A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. Water Science and Technology 82: 2613-2634.

24. Al-Hawawreh M, Sitnikova E, Den Hartog F (2019) An efficient intrusion detection model for edge system in brownfield industrial internet of things. In Proceedings of the 3rd international conference on big data and internet of things pp 83-87.

25. Khan R, Maynard P, McLaughlin K, Laverty D, Sezer S (2016) August. Threat analysis of blackenergy malware for synchrophasor based real-time control and monitoring in smart grid. In 4th International Symposium for ICS & SCADA Cyber Security Research BCS pp 53-63.