Collaborative Discussion 1 - The Data Collection Process

Discussion Topic

Critically evaluate the rationale behind the Internet of Things (IOT), in the context of **the article by Huxley et al (2020)**, highlighting the opportunities, limitations, risks and challenges associated with such a large-scale process of data collection.

Instructions

- Go to the discussion forum and create an initial post of your contribution to the discussion.
- Review the Lecturecast and reading for this Unit.
- Review other literature in conjunction with this paper to enhance your post.
- Demonstrate that you understand the topic covered and ensure you use references to academic literature (journals, books, reports, etc.)

Learning Outcomes

- Identify and manage challenges, security issues and risks, limitations, and opportunities in data wrangling.
- Critically analyse data wrangling problems and determine appropriate methodologies, tools, and techniques (involving preparing, cleaning, exploring, creating, optimising and evaluating big data) to solve them.

Initial Post

by Valentina Mercieca - Wednesday, 30 October 2024, 8:56 AM

Number of replies: 3

The Internet of Things (IoT) is transforming data collection by connecting physical devices, generating vast data streams that demand robust wrangling practices. IoT integration within big data systems enables real-time insights that improve decision-making across sectors like healthcare, urban infrastructure, and logistics (Kamel Boulos & Al-Shorbaji, 2014). Azure's lambda architecture facilitates IoT data processing through a combination of batch and real-time streams. This dual-path system allows for immediate analysis in the "hot path" (real-time processing) while accumulating long-term insights through "cold path" (batch) storage and analysis. Such an architecture is valuable for IoT applications that require both rapid responses like anomaly detection, and retrospective analysis for trend identification and resource optimisation (Tejada, N.D.).

Managing IoT data poses unique challenges, particularly in maintaining data quality. IoT data is susceptible to errors caused by device malfunctions, transmission gaps, and environmental factors such as local weather conditions, improper device placement, and range limitations, all of which can impair sensor performance and lead to data inconsistencies (Teh et al., 2020). Rigorous cleaning and validation are essential to prepare IoT data for reliable analysis (Huxley et al., 2020). For instance, data lakes - as recommended in Azure's architecture - provide scalable storage that supports extensive IoT data but also require strong data wrangling practices, such as filtering and normalisation, to address inconsistencies across large datasets. Poor-quality data can compromise analysis accuracy and skew statistical outcomes, a risk underscored by Kozak et al. (2015), who note that even single data anomalies can affect inferential accuracy.

In addition to data quality, IoT systems bring challenges in both security and processing complexity. In Azure's lambda architecture, the hot and cold paths eventually converge at the analytics client application. This approach adds latency since the cold path stores all incoming data in its raw form and performs batch processing on it – taking a very long time. This latency is a trade-off for the high accuracy and comprehensive analysis provided by the batch layer. However, for real-time needs, clients can select data from the hot path, which provides timely but potentially less accurate insights. This dual-path approach requires balancing speed with accuracy, as the hot path captures a small-time window that can later be refined with data from the cold path (Tejada, N.D.). Furthermore, since IoT devices are continuously transmitting data, they are vulnerable to security threats—making data protection practices essential for maintaining data integrity (Roman et al., 2011). Safeguarding IoT data during real-time transmission and in storage is key to ensuring compliant, trustworthy data processing across both immediate and long-term applications.

Addressing these challenges requires targeted data wrangling techniques and tools. Techniques such as data cleaning, normalisation, and imputation help manage IoT data inconsistencies, while tools like Azure Data Lake and Stream Analytics support scalable storage and real-time processing. These tools facilitate efficient handling of high-velocity data, enabling organisations to derive insights without sacrificing security or data quality. As IoT applications expand, using such tools alongside rigorous data wrangling practices will be essential for balancing immediacy with accuracy and security.

References

Huxley, K. (2020) ‘Data Cleaning’, in: P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (eds) *SAGE Research Methods Foundations*. DOI: 10.4135/9781526421036842861

Kamel Boulos, M.N., & Al-Shorbaji, N.M. (2014) On the Internet of Things, smart cities and the WHO Healthy Cities. *International Journal of Health Geographics* 13(10): 1-6. DOI: 10.1186/1476-072X-13-10

Kozak, M., Krzanowski, W., Cichocka, I., & Hartley, J. (2015) The effects of data input errors on subsequent statistical inference. *Journal of Applied Statistics* 42(9): 2030-2037. DOI: 10.1080/02664763.2015.1016410

Roman, R., Najera, P., & Lopez, J. (2011) Securing the Internet of Things. *IEEE Computer* 44(9): 51-58. DOI: 10.1109/MC.2011.291

Teh, H.Y., Kempa-Liehr, A.W., & Wang, K.IK. (2020) Sensor data quality: a systematic review. *Journal of Big Data* 7(11): 1-49. DOI: 10.1186/s40537-020-0285-1

Tejada, Z. (N.D.) Big data architectures. Available from: https://learn.microsoft.com/en-us/azure/architecture/databases/guide/big-data-architectures [Accessed 28 October 2024].

Peer Response

by Opeyemi Adeniran - Monday, 4 November 2024, 10:11 AM

Your exploration of the dual-path processing architecture in Azure for IoT data highlights its benefits and limitations, particularly regarding the trade-off between immediacy and accuracy. This dual-path strategy supports various IoT applications and enables real-time and long-term analysis. However, as you noted, balancing these paths presents challenges, particularly regarding data latency and quality.

To address these challenges, additional measures could improve data integrity and reliability. For example, implementing edge computing could mitigate latency issues by performing initial data processing closer to IoT devices, thereby reducing reliance on cloud-based hot-path processing. Edge processing can provide real-time insights directly from devices, minimizing potential data transmission errors that are common in centralized processing configurations (Shi et al., 2016). Additionally, integrating continuous monitoring and automated anomaly detection at the edge could prevent data accuracy issues by filtering bad data before it reaches the main architecture (Alrawais et al., 2017).

Security is another critical aspect you mentioned, especially since IoT systems are vulnerable to continuous data transfer. Integrating end-to-end encryption from device to cloud and using multi-factor authentication for device access could help address these vulnerabilities. Studies have shown that layered security measures significantly reduce the risk of data breaches, thereby preserving the integrity of real-time and batch-processed data (Fernandes et al., 2017).

The Azure IoT architecture provides a solid foundation, but adding edge processing and robust security practices could address some of the inherent limitations you highlighted. Taken together, these adjustments would further support reliable, scalable data processing, particularly for applications that require both immediate action and long-term analysis.

References

Alrawais, A., Alhothaily, A., Hu, C., & Cheng, X. (2017) 'Fog computing for the Internet of Things: Security and privacy issues'. IEEE Internet Computing, 21(2), pp. 34-42. DOI: 10.1109/MIC.2017.37

Fernandes, E., Rahmati, A., Jung, J., & Prakash, A. (2017) 'Security implications of permission models in smart-home application frameworks'. IEEE Security and Privacy, 15(2), pp. 24-30. DOI: 10.1109/MSP.2017.43

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016) 'Edge computing: Vision and challenges'. IEEE Internet of Things Journal, 3(5), pp. 637-646. DOI: 10.1109/JIOT.2016.2579198

Tutor's Feedback

by Godfried Williams - Thursday, 21 November 2024, 5:34 PM

Hi Valentina,

Useful insight of IOT and proprietary data architecture which has become central to the discussions of handling data in organisations. Why do you think a hot or cold data path is important? Explore additional evidence that backs up or refutes this assertion and proposition.

Response to Tutor's Feedback

by Valentina Mercieca - Sunday, 24 November 2024, 1:41 PM

Thank you for the feedback, Prof. Williams.

A hot data path handles real-time or frequently accessed data, such as sensor readings for monitoring and control systems. Conversely, a cold data path stores historical or infrequently accessed data, such as logs for analytics, compliance, or archival purposes.

In IoT, performance optimisation through hot data paths is essential. Applications like smart cities, autonomous vehicles, and industrial automation depend on low-latency access to sensor data for instant decision-making. For example, real-time traffic systems require immediate processing of road sensor data to dynamically adjust signals. Technologies like edge computing are often employed to process data locally, reducing latency and improving responsiveness (Nguyen & Kortun, 2020). In-memory databases and state-of-the-art SSDs are also commonly used to enable this rapid data processing (VSTL, 2024).

While hot paths focus on immediate processing, cold paths address the need for cost-efficient, long-term storage. IoT systems generate significant amounts of cold data, such as historical sensor logs or rarely accessed video footage. Moving this data to cost-effective cold storage solutions, such as HDDs or cloud-based cold tiers like AWS Glacier, significantly reduces operational costs. For instance, predictive maintenance systems store years of historical sensor data, which is accessed occasionally for trend analysis or compliance audits. Reports from cloud providers show that organisations using tiered storage in IoT environments reduce storage costs by up to 70% while maintaining access when needed (Peram, 2024).

However, IoT deployments face challenges, particularly in managing the complexity of hot and cold paths. Dynamic tiering and seamless data transitions can be technically demanding, but advancements in AI-driven data management and edge-cloud integration are addressing these issues (Firouzi et al., 2022). For example, AI and machine learning enable storage solutions to automatically assign data to the most appropriate tier—whether high-performance SSDs for frequently accessed data (hot path) or cost-effective HDDs for archival purposes (cold path) (Naor, 2024).

By combining these approaches, IoT systems can optimise performance, maintain data integrity, and minimise costs, making hot and cold data paths indispensable for modern IoT applications.

References

Firouzi, F., Farahani, B. & Marinšek, A. (2022) The Convergence and Interplay of edge, fog, and cloud in the AI-Driven Internet of Things (IoT). Information Systems. DOI: https://doi.org/10.1016/j.is.2021.101840

Naor, G. (2024) Why Auto-Tiering is Essential for AI Solutions: Optimizing Data Storage from Training to Long-Term Archiving. Available from: https://insideainews.com/2024/11/11/why-auto-tiering-is-essential-for-ai-solutions-optimizing-data-storage-from-training-to-long-term-archiving/ [Accessed 23 November 2024].

Nguyen, L. & Kortun, A. (2020) Real-time Optimisation for Industrial Internet of Things (IIoT): Overview, Challenges and Opportunities. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems 7(25): 1-7. DOI: 10.4108/eai.16-12-2020.167654

Peram, P. (2024) Optimizing Cloud Computing Performance: A Comprehensive Framework of Strategies and Best Practices. International Journal of Engineering and Technology Research 9(2): 397-419. DOI: https://doi.org/10.5281/zenodo.13851330

VSTL. (2024) How NVMe SSDs Deliver Unmatched Speed, Efficiency, and Scalability. Available from: https://vstl.info/how-nvme-ssds-deliver-unmatched-speed-efficiency-and-scalability/ [Accessed 22 November 2024].