ABSRACT

of the dissertation work by Imankulova Binara Bakytzhanovna titled "Design and implementation of a provenance-aware LoRaWAN IoT framework for smart city air-quality data management", submitted for the degree of Doctor of Philosophy (PhD) in the educational program 8D06102 – "Computer and Software Engineering"

General characteristics of the work.

The dissertation presents the design and the implementation of a provenance-aware LoRaWAN Internet of Things (IoT) framework. The targeted application area was chosen as the management of air-quality data in smart cities. Provenance, integrity, and completeness are three main dimensions that the framework addresses trust in monitoring data. Provenance here describes how each stored measurement was received, decoded, validated and then stored. Integrity is the ability to detect subsequent changes to the recorded history as well as completeness reflects whether all expected measurements were received and, if not, how many are missing.

The framework captures provenance on the platform side, after uplink reception. It requires no changes to the LoRaWAN protocol and adds no provenance data to the sensing-node payload. Provenance records follow W3C PROV semantics and they are stored in per-device hash-linked chains with chain-head checkpoints. It means that the design provides tamper evidence under a stated threat model for the stored history. The same records are used to build PROV-style graph views for backward and forward tracing. For each missing measurement, the framework identifies the stage where the loss occurred: the device, upstream communication, platform intake, or processing. The cause of each data loss is identified by means of the frame counters, the reception times, the state of the records and the event log of the network server. When the available data are not sufficient, the cause is classified as undetermined.

Relevance of the work.

Internet of Things (IoT) technologies have become an important component of modern urban infrastructure. Smart city services rely on measurements produced by sensing nodes and transferred through gateways, communication networks, and platform-level processing systems. Current forecasts indicate that the number of connected devices worldwide will reach 32.1 billion by 2030. This expansion increases both the volume and the heterogeneity of data used in environmental monitoring, traffic management, public safety, and other urban applications.

Sensor measurements support monitoring, reporting, operational control, and decision-making. Before these measurements become part of an operational dataset, they pass through sensing, wireless transmission, ingestion, processing, storage, and publication stages. Each stage may add metadata, change the data format, or create a new record. Measurements may also be lost, duplicated, replayed, incorrectly decoded, or modified during platform-side processing.

These problems do not need to result from malicious activity. Packet loss, duplicate gateway receptions, decoding failures, and storage errors can also affect the final dataset. Air-quality monitoring is especially sensitive to such conditions

because missing or incorrectly processed values may influence long-term statistics and environmental indicators. When the processing history is not recorded, it becomes difficult to determine how a stored value was received, transformed, and written to the database.

Data provenance records this history. It describes the source of a measurement, the reception conditions, the processing operations, and the resulting stored record. Provenance information can therefore support traceability, auditability, and integrity verification. These functions are important when environmental data are used for public reporting, scientific analysis, and long-term urban planning.

Existing provenance approaches provide different forms of protection and representation. Blockchain-based systems can preserve tamper-resistant records, but they often introduce additional storage, latency, and consensus overhead. Cryptographic mechanisms can detect record modification, although they may require additional computation, key-management procedures, or protocol changes. These requirements are difficult to accommodate in low-power sensing nodes and LPWAN deployments.

LoRaWAN is widely used for distributed environmental monitoring because it provides long-range communication with low energy consumption. The protocol also provides payload confidentiality and message integrity checks during transmission. However, it does not preserve the processing history of a measurement after uplink reception. Stored and exported records may therefore be changed without a corresponding provenance record. Recent studies show that practical approaches combining LoRaWAN-compatible provenance capture, tamper-evident storage, and traceability remain limited.

Missing data are a further problem that existing approaches rarely address. In long-term LoRaWAN deployments, a substantial share of expected measurements is never stored: devices stop transmitting, frames are lost on the radio path, and messages are lost between the network server and the platform or fail during decoding. Completeness is usually reported as a single percentage, which shows how much data is missing but not where it was lost. Without this information, operators cannot decide whether to repair devices, extend gateway coverage, or correct the platform, and analysts cannot judge whether gaps are random or systematic.

The systematic literature review in Chapter 1 identified four gaps in existing provenance solutions for IoT and smart city systems: provenance capture that requires changes to end devices or covers only part of the processing pipeline; the absence of reception evidence in standard provenance models; integrity claims that are not tied to an explicit threat model; and the absence of stage-level attribution of missing data. These gaps define five requirements for a provenance-aware LoRaWAN platform: stage-complete provenance capture without changes to end devices, a standard provenance model that includes reception evidence, tamper evidence for stored provenance under a stated threat model, bounded overhead, and stage-level attribution of missing data. This dissertation develops a framework that meets these requirements.

The aim and objectives of the research. The aim of this dissertation is to develop a platform-side provenance-aware framework for traceable and tamper-

evident management of air-quality data in LoRaWAN-based smart-city monitoring systems.

To achieve this aim, the dissertation addresses four **objectives**:

1. to analyse existing IoT data provenance approaches and identify their limitations in smart city and LPWAN deployments;
2. to design and formalize a lightweight provenance, integrity, and missing-data attribution model for LoRaWAN-based monitoring;
3. to develop and implement a platform-side prototype that captures, stores, verifies provenance records and reconstructs provenance graph views;
4. to evaluate the prototype in a LoRaWAN monitoring deployment and in simulation in terms of data completeness and attribution of missing data, platform overhead, integrity-verification performance, and provenance traceability.

The object of the research is the lifecycle of IoT measurement data in LoRaWAN-based smart city monitoring systems.

The subject of the research is the set of methods and mechanisms used to capture provenance, protect provenance integrity, and support traceability across the platform-side measurement lifecycle.

Methods of research. The research has combined four types of methods, namely analytical, design, empirical and simulation methods. The analytical stage included a systematic literature review conducted according to PRISMA 2020, which examined IoT data provenance, integrity protection, data completeness and LPWAN communications, and defined the requirements for the proposed framework. Then, the design stage translated such requirements into a platform-side architecture and a formal model of the generation of provenance, the verification of integrity and the attribution of missing data. Moreover, a functional prototype was implemented in order to capture provenance after the reception of each uplink, to store it in hash-linked chains per device, to verify the integrity of each chain and to reconstruct graph views for traceability queries.

Afterwards, the empirical stage evaluated the prototype by means of a 12-month LoRaWAN deployment for air quality monitoring, in terms of data completeness, ingestion performance of the backend, storage overhead, verification time of the provenance chains, tamper detection and traceability. The attribution of missing data was evaluated by means of computer simulations. More specifically, the radio layer was simulated with the ns-3 network simulator and its LoRaWAN module, and failures of devices, gateways and platform with known causes were added, in order to compare the attributed causes with the true ones.

The scientific novelty

A platform-side provenance model for LoRaWAN monitoring is proposed, in which gateway reception evidence is an explicit W3C PROV entity and no changes to devices or the protocol are required, as confirmed by backward and forward tracing of stored measurements to their reception evidence.

A tamper-evident storage scheme adapted to the LoRaWAN uplink lifecycle is introduced, combining per-device hash chains and chain-head checkpoints, which detected all five tested types of record change and verified 403,900 records in 5.87 s.

A method for stage-level attribution of missing measurements is proposed, which, unlike aggregate completeness metrics, identifies whether data were lost at the device, upstream, at platform intake, or during processing, with 3 errors in 270,107 attributed measurements in ns-3 simulation.

Main thesis statements submitted for defense:

1) The architecture of a provenance-aware LoRaWAN framework on the platform side, aimed at air quality monitoring, which captures provenance, verifies its integrity and attributes missing data without any modification of the LoRaWAN protocol or of the payload of the sensing nodes.

2) The provenance model of the lifecycle of LoRaWAN uplinks, which represents the reception, ingestion and storage stages by means of W3C PROV entities, activities and agents, including the evidence of reception at the gateways, and supports both backward and forward traceability of the stored measurements.

3) The method for tamper-evident storage of provenance, based on hash-linked chains per device and checkpoints of the chain heads, which detects the modification, deletion, insertion and reordering of records, as well as the truncation of the chain tail, under the stated trust assumptions.

4) The method for the attribution of missing measurements at the level of each stage, which takes advantage of the evidence recorded by the platform and the network server in order to assign each missing measurement to one of the stages (i.e., device, upstream, platform intake or processing), or to classify it as undetermined.

Scientific results

1. A systematic literature review was conducted according to PRISMA 2020. Among 1,201 records retrieved from Scopus, Web of Science and additional sources, 90 studies were included in the review. According to the results, the existing approaches have rarely combined platform-side capture, suitability for LoRaWAN, tamper-evident storage, traceability and attribution of missing data.

2. A provenance-aware LoRaWAN framework aimed at air quality monitoring was designed and implemented. More specifically, the framework captures the evidence of reception, processing and storage after the reception of each uplink, without any modification of the protocol or of the payload of the sensing nodes.

3. A formal model of provenance and integrity was developed. The model defines stable identifiers for devices, uplinks, gateway receptions, measurements and provenance records, and represents the lifecycle on the platform side by means of three activities (i.e., RX, INGEST and STORE). In particular, the deployment produced 135,000 RX, 135,000 INGEST and 133,900 STORE records.

4. A method for tamper-evident storage, based on hash-linked chains per device and checkpoints of the chain heads, was developed. The verification of 403,900 records took 5.87 s (i.e., 0.0145 ms per record), and all five tested types of change were detected. Among them, the truncation of the chain tail was detected by means of the comparison of checkpoints in 0.94 ms.

5. A method for the on-demand reconstruction of graph views in the PROV style was developed. The method supports both backward tracing, from a stored

measurement to its source uplink, and forward tracing, from an uplink to its stored records, without maintaining a permanent global graph.

6. The framework was evaluated by means of a 12-month deployment with three sensing nodes and two gateways, which produced 135,000 logical uplinks, 133,900 stored measurements and 403,900 provenance records. The effective completeness reached 84.9%, and the database occupied 164.08 MB, of which 114.14 MB corresponded to the provenance records.

7. A paired benchmark of 5,000 logical uplinks over 10 runs was performed. According to the results, the capture of provenance increased the median processing latency from 1.565 ms to 2.189 ms per uplink (i.e., by 0.62 ms).

8. The method for the attribution of missing measurements at the level of each stage was evaluated by means of simulations, with a radio layer based on ns-3 and calibrated to the deployment. Over 20 runs of one year each, only 3 out of 270,107 attributed measurements were assigned to the wrong stage. Moreover, the coverage reached 66.4% (93.7% without the re-joins of devices), and the accuracy decreased to 96.5% when the event log of the network server was not available.

The theoretical and practical significance of the work.

The theoretical significance of the dissertation lies in the proposed provenance model for monitoring systems based on LPWAN technologies. The model represents the lifecycle of IoT measurements on the platform side, from the reception of the uplink to the storage in the database, and combines the capture of provenance, the verification of its integrity and the attribution of missing data. Moreover, the proposed model can be used as a basis for further research regarding verifiable data lineage and data completeness in IoT environments with constrained resources.

The practical significance of the work lies in the provenance-aware LoRaWAN prototype implemented for air quality monitoring in smart cities. The prototype operates on the platform side, without any modification of the LoRaWAN protocol or of the payload of the sensing nodes, and allows the analysis of missing measurements, decoding failures and violations of the integrity of the provenance chains. In particular, the results show that the event log of the network server must be retained independently of the platform collector, since otherwise the failures of the platform cannot be distinguished from the losses on the radio path.

Connection of this work with other research projects. The research presented in this dissertation was started within the framework of the project *"Software and hardware complex for the analysis and monitoring of climatic and environmental changes of the environment*" (grant No. AP05134597-OT-20 of the Ministry of Education and Science of the Republic of Kazakhstan, 2018 - 2020). Furthermore, the continuation of the research will be carried out within the framework of the project for young scientists "Zhas Galym" (grant No. AP32523239, 2026 - 2028).

Approbation of work. The main statements and scientific results of the dissertation have been presented and discussed at the following conferences:

1. seminars of the Department of Computer Engineering at the International Information Technology University, Almaty, Kazakhstan;

2. the 5th International Conference on Smart Information Systems and Technologies (SIST 2025), Astana, Kazakhstan;

3. the X International Scientific Conference “Mathematics, Mechanics and Information Technologies” (ICMMIT 2025), al-Farabi Kazakh National University, Almaty, Kazakhstan;

4. a seminar at Astana IT University, Astana, Kazakhstan (August 28, 2026).

5. the International Conference on Digital Technologies and Applications (ICDTA 2026), Vaughn College of Aeronautics and Technology, New York, USA (October 9-10, 2026).

The results of the dissertation have been implemented at SDU University (Kaskelen, Kazakhstan), as confirmed by the Act of Implementation presented in Appendix E.

Scientific publications. The results of the dissertation have been published in 6 scientific works, namely 3 papers in journals recommended by the Committee for Quality Assurance in Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan (MSHE RK), 2 papers in journals indexed in Scopus and Web of Science (WoS), and 1 paper in the proceedings of an international conference.

1. **B. Imankulova**, W. M. S. Yafooz, K. Alshammari, B. Alshemali, A. E. Yahya and A. Abd Almisreb. Tracing Data Origins in Smart Cities: An IoT Perspective. IEEE Access. 2025. Vol. 13, pp. 130201-130214. DOI: 10.1109/ACCESS.2025.3591277 [WoS Q2, Scopus percentile: 91, Cite Score(2025): 9.3, available: https://ieeexplore.ieee.org/document/11087496]

2. Mukash Zh., Bakranova D., Serikkanov A., Bakranov N., **Imankulova B.** Energy-Aware AI-Enabled Intelligent Sensing Systems for IoT and Edge–Cloud Architectures. Sensors 2026, Volume 26, Issue 21, 6927, 10.3390/s26216927 [WoS Q2, Scopus percentile: 83, Cite Score (2025): 9.4]

3. **B. Imankulova**, A. Sagymbekova. Privacy Challenges in Environmental IoT Systems: A Case Study on Air Quality Monitoring, 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST), Astana, Kazakhstan, 2025, pp. 1-6, doi: 10.1109/SIST61657.2025.11139338 [Scopus]

4. **Imankulova B.B**. Privacy challenges and framework design for IoT devices: A comprehensive approach. Bulletin of AUPET. 2025. Vol. 68, Issue 1. p. 256–267. DOI: https://doi.org/10.51775/2790-0886_2025_68_1_256 [KOKHBO]

5. **Imankulova, B. B**., Kozina, L. A., Duzbayev, N. T., Amanzholova, S. T. Transmission, storage and preprocessing of data received from sensors using LoRa technology. Bulletin of AUPET. 2024. Vol. 64, Issue 1. p.16-24. DOI: https://doi.org/10.51775/2790-0886_2024_64_1_16 [KOKHBO]

6. **Imankulova B.B.**, Amanzholova S.T., Duisebekova K.S. Software and technical aspects of data collection from end nodes: problems and prospects. Bulletin of AUPET. 2020. Vol. 51, Issue 4. p.164-172. DOI: https://doi.org/10.51775/1999-9801_2020_51_4_164 [KOKHBO]

Chapter overview

In Chapter 1 we review the state of the art regarding IoT data provenance for smart city systems. The chapter examines the concepts of provenance, the requirements of smart city data, the representation models, the integrity mechanisms

and data completeness, and presents a systematic literature review conducted according to PRISMA 2020. As a result of the review, four research gaps and five requirements for a provenance-aware LoRaWAN platform are derived.

In Chapter 2 we present the design of the proposed framework. The chapter describes the design principles and the layered architecture, the integration with the LoRaWAN communication stack, as well as the threat model and the precise meaning of the security claim. Afterwards, the formal model of the framework is presented, namely the system components and identifiers, the provenance records and graph, and the models for the generation of provenance, the verification of integrity and the attribution of missing data. At the end of the chapter we describe the storage and verification of provenance.

In Chapter 3 we describe the implementation of the framework, including the platform architecture, the field deployment, the ingestion via MQTT and the capture of provenance, the storage schema, the procedure of integrity verification, the provenance queries and the procedure of attribution of missing data.

In Chapter 4 we evaluate the framework by means of a 12-month deployment, in terms of data completeness, link quality and Adaptive Data Rate (ADR) behaviour, storage and processing overhead, integrity verification and traceability. Furthermore, the attribution of missing data is evaluated against known causes by means of simulations with a LoRaWAN radio layer based on ns-3. At the end of the chapter a comparison with the closest studies and a discussion of the limitations were added.

Finally, the Conclusion summarizes the main findings and contributions of the dissertation, also outlines the limitations of the proposed framework, identifies the directions for future research.

Structure and volume of the thesis

The dissertation consists of 155 pages, including 34 figures, 22 tables, a list of 171 references and 4 appendices.