Knowledge-Augmented Analytics for Explainable Data Exploration in Smart City Infrastructures

Devis Bianchini, Valeria De Antonellis, Massimiliano Garda, Michele Melchiori
Dept. of Information Engineering - University of Brescia
Via Branze, 38 - 25123 Brescia (Italy)
devis.bianchini,valeria.deantonellis,massimiliano.garda,michele.melchiori@unibs.it

***Abstract*—Smart cities increasingly rely on interconnected digital infrastructures that continuously generate large volumes of heterogeneous data, including sensor streams, operational events, maintenance records, and technical documentation. While Big Data Analytics (BDA) effectively support monitoring and anomaly detection, they often lack explainability and fail to exploit the knowledge embedded in textual information sources. At the same time, Large Language Models (LLMs) enable the extraction and organization of semantic knowledge from unstructured textual sources. This work presents a vision for a knowledge-augmented BDA architecture that integrates quantitative analytics with ontology-guided Knowledge Graphs (KGs) to support explainable data exploration through evidence-grounded responses. The proposed architecture combines Data Lake technologies, composable data services, asset-specific KGs, and LLM-based query answering. Through a representative scenario, we illustrate how semantic knowledge and operational data can be jointly exploited to contextualize anomalies, support root-cause analysis, and enhance decision-making processes.**

***Index Terms*—Big Data Analytics; Large Language Models; Knowledge Graphs; Multi-tier Architecture**

I. Introduction

The digital transformation of cities is driving the deployment of interconnected cyber-physical infrastructures such as transportation systems, energy networks, water-distribution systems, waste-management platforms, and advanced manufacturing facilities. These infrastructures continuously generate large volumes of heterogeneous data that must be transformed into actionable knowledge for operators and decision makers. This vision is aligned with the broader paradigm of cognitive systems, where data analytics, knowledge representation, and artificial intelligence are jointly exploited to support human-centered decision making in complex operational environments [4]. Although existing BDA platforms successfully support monitoring and anomaly detection tasks, two important limitations remain [1], [2]. First, analytical pipelines are largely confined to numerical data and fail to exploit the rich knowledge contained in textual information sources. Second, detected anomalies often lack evidence-based contextual explanations, forcing experts to manually correlate analytical results with historical operational knowledge.

Various smart-city applications, including transportation systems, energy networks, and advanced manufacturing environments, represent challenging features because of the volume, velocity, and heterogeneity of generated data. Modern systems continuously collect measurements from several

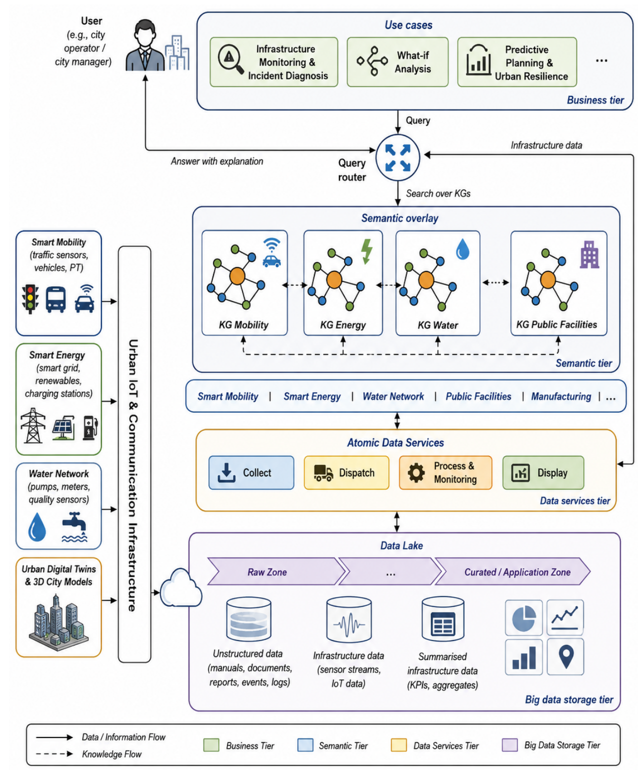

Fig. 1: Overview of the multi-tier BDA architecture.

sensors distributed across smart-city infrastructures. Advanced stream analytics may identify abnormal operating conditions. However, understanding the causes of such anomalies frequently requires consulting maintenance reports, previous interventions, and domain-specific documentation. This separation between numerical evidence and semantic knowledge limits the effectiveness and explainability of data analytics. To address these limitations, we envision a knowledge-augmented architecture that integrates quantitative analytics and semantic reasoning through the combined use of Data Lakes, Knowledge Graphs, and LLMs.

II. Knowledge-augmented architecture

The proposed architecture is organized into four complementary tiers (see Figure 1).

The **Big Data Storage Tier** manages heterogeneous information sources using Data Lakes. Numerical sensor measurements, event logs, maintenance records, technical manuals, and other operational artifacts are collected and stored in a scalable environment capable of supporting both structured and unstructured data.

The **Data Services Tier** exposes reusable services responsible for data collection, processing, monitoring, transformation, and visualization. These services can be composed into analytical pipelines supporting anomaly detection, predictive maintenance, and process-performance assessment.

The **Semantic Tier** enriches data through ontology-guided knowledge extraction. LLMs process unstructured documents to populate asset-specific Knowledge Graphs that provide a structured, verifiable semantic representation of extracted knowledge, according to recent ontology-guided and GraphRAG-inspired approaches [3], [5]. The resulting graphs capture entities such as faults, maintenance activities, materials, tools, and operational procedures, together with the relationships among them. Cross-asset semantic links connect individual graphs into a unified semantic overlay that supports knowledge sharing, in line with recent knowledge-graph solutions for smart maintenance [6].

Finally, the **Business Tier** enables users to interact with the system through natural language queries. A query-routing component coordinates the retrieval of information from both the semantic overlay and the analytical infrastructure. Retrieved evidence is then synthesized by an LLM to generate natural-language explanations explicitly grounded in the retrieved evidence, rather than relying solely on the generative capabilities of the LLM.

III. Knowledge-augmented data exploration

A key innovation of the proposed vision lies in combining semantic retrieval and quantitative analytics during query answering, extending GraphRAG-style retrieval mechanisms toward industrial Big Data environments [3]. When a user submits a natural-language request, the system identifies the relevant assets and retrieves complementary evidence from multiple sources. The semantic overlay provides contextual information extracted from maintenance documents and technical reports, while data services retrieve summarized operational data from sensor streams and historical production records.

Consider an anomaly-contextualization scenario. An operator observes abnormal power consumption in a monitored infrastructure and asks the system to explain the detected behavior. Traditional analytics can identify the anomaly but not its cause, whereas maintenance reports may describe recent interventions without demonstrating their impact. The semantic overlay retrieves notes on recent interventions and reports of temporary power fluctuations, while the analytical layer provides operational summaries confirming unusual data patterns after the intervention. The LLM synthesizes this evidence into a natural-language explanation that faithfully connects numerical observations with documented operational events.

Beyond anomaly explanation, the semantic overlay also enables cross-asset knowledge transfer. Similar faults can be linked through semantic relationships across asset-specific Knowledge Graphs, allowing maintenance engineers to reuse diagnostic knowledge and corrective actions adopted in comparable situations. The proposed approach is domain-independent and applicable to a wide range of smart infrastructures where operational data coexist with extensive maintenance and technical documentation.

IV. Conclusions and research challenges

This extended abstract presented a vision for explainable data exploration in smart-city infrastructures through the integration of Big Data Analytics, Knowledge Graphs, and Large Language Models. In the proposed architecture, the LLM produces human-readable explanations, while their factual grounding is provided by retrieved operational data and Knowledge Graph evidence. The envisioned framework is applicable to various urban domains, including smart mobility, energy networks, water infrastructures, public facilities, and advanced manufacturing. By combining quantitative evidence with semantically grounded knowledge, it supports infrastructure monitoring, anomaly contextualization, knowledge reuse, and explainable decision support.

Open research challenges include the evolution of ontological schemas for knowledge extraction, temporal validity management in Knowledge Graphs, and the scalability, trustworthiness, and faithfulness of LLM-grounded explanations, ensuring consistent support from retrieved operational and semantic evidence. Validation of the proposed architecture in real smart-city settings will assess its effectiveness for infrastructure management and urban decision making.

Acknowledgment. Generative AI tools were used for linguistic editing and manuscript refinement. The authors take full responsibility for the content of this work.

References

[1] Ada Bagozi, Devis Bianchini, Valeria De Antonellis, Massimiliano Garda, and Alessandro Marini. A relevance-based approach for big data exploration. *Future Generation Computer Systems*, 101:51 – 69, 2019.

[2] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. Apache flink: Stream and batch processing in a single engine. *The Bulletin of the Technical Committee on Data Engineering*, 38(4), 2015.

[3] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[4] Fadi El Kalach, Ibrahim Yousif, Thorsten Wuest, Amit Sheth, and Ramy Harik. Cognitive manufacturing: definition and current trends. *Journal of Intelligent Manufacturing*, 36(6):3695–3715, 2025.

[5] Zeno van Cauter and Nikolay Yakovets. Ontology-guided knowledge graph construction from maintenance short texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 75–84, 2024.

[6] Yanying Wang, Ying Cheng, Qinglin Qi, and Fei Tao. Ids-kg: An industrial dataspace-based knowledge graph construction approach for smart maintenance. *Journal of Industrial Information Integration*, 38:100566, 2024.