Article:

Perera, I.R., Speight, V., Young, S. et al. (2026) Uncovering gaps and opportunities in water distribution system operations through graph-database-driven data utilization and process mapping. AQUA — Water Infrastructure, Ecosystems and Society. ISSN: 2709-8028

https://doi.org/10.2166/aqua.2026.115

AQUA
Water Infrastructure, Ecosystems and Society

AQUA — Water Infrastructure, Ecosystems and Society Vol 00 No 0, 1 doi: 10.2166/aqua.2026.115

Uncovering gaps and opportunities in water distribution system operations through graph-database-driven data utilization and process mapping

Ishara R. Perera [a,*], Vanessa Speight [a], Scott Young [b] and Joby Boxall [a]
[a] School of Mechanical, Aerospace and Civil Engineering,University of Sheffield, Sheffield S1 3JD, UK
[b] Scottish Water, Edinburgh EH10 6YQ, UK
*Corresponding author. E-mail: irperera1@sheffield.ac.uk

IRP, 0009-0008-4602-5867; VS, 0000-0001-7780-7863; SY, 0009-0006-3685-2469; JB, 0000-0002-4681-6895

ABSTRACT

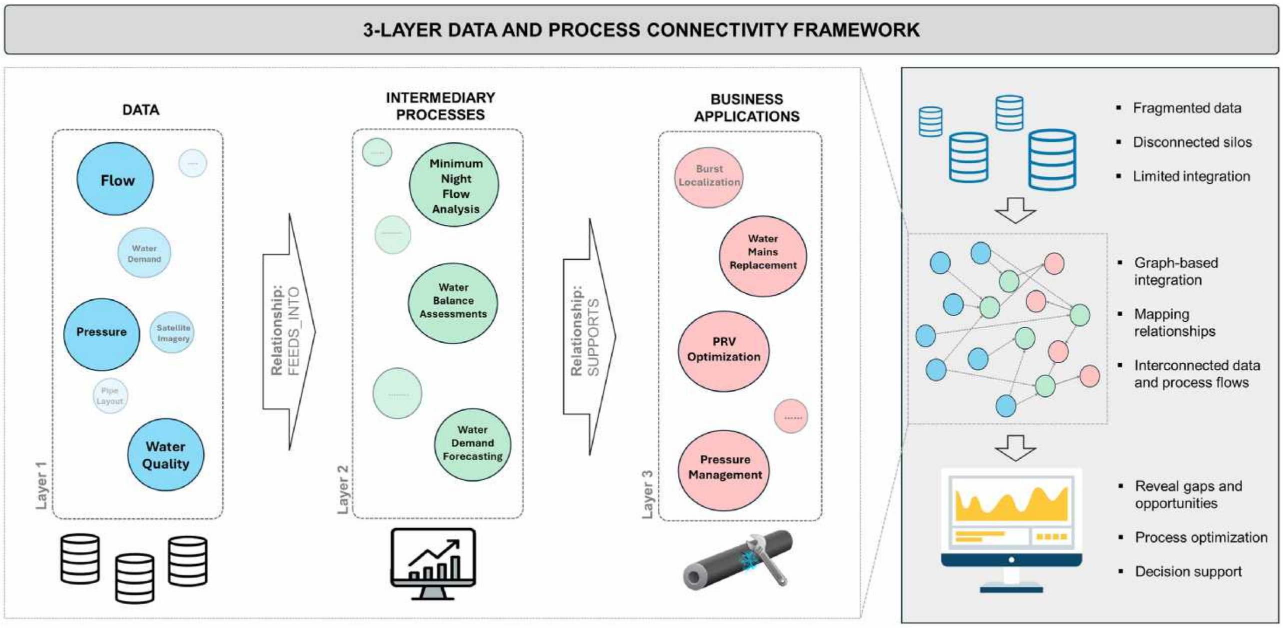
The water industry faces increasing challenges in ensuring efficient and sustainable delivery of safe drinking water through ageing infrastructure while meeting operational, regulatory, and customer demands. Large volumes of data are generated but are stored and utilised in disconnected silos. Such fragmentation limits the potential for comprehensive analysis, informed decision-making, and operational optimisation. To better understand the impact of this fragmentation, we developed a three-layer data and process connectivity framework, a novel graph-database-driven approach designed to identify, analyse, and address gaps in data utilisation and data process efficiency. The framework systematically maps critical data flows and reveals interdependencies between data, intermediary processes, and business applications. We validated the framework through a case study with Scottish Water, constructing a graph comprising 127 nodes and 911 relationships across water distribution data and processes. Using node centrality measures, relationship-based graph querying, and impact simulations, we demonstrate how this approach exposes inefficiencies, uncovers overlooked data relationships, and generates actionable insights for process integration. The findings highlight the potential of graph-database-driven approaches to overcome data fragmentation, bridge silos, and enable more data-driven and strategically optimised operations, with applicability across the wider water sector and other industries.

Key words: data and process mapping, decision support frameworks, graph databases, water distribution systems, water utility management, water utility operations

HIGHLIGHTS

- A novel three-layer, graph-database-driven framework that maps data flows and process utilisation in water utility management.
- Reveal data silos and criticality to support more resilient, data-driven decisions.
- Graph analytics to uncover underutilized datasets and process opportunities.
- A generalizable methodology with applicability to other linear infrastructure sectors with complex, fragmented data ecosystems.

GRAPHICAL ABSTRACT



1. INTRODUCTION

The water industry plays a critical role in ensuring the quality, availability, and equitable distribution of water, a vital resource for human survival, economic development, and environmental sustainability. However, this sector faces increasing challenges driven by ageing infrastructure, climate change, urbanization, and resource scarcity (Wallis-Lage 2020; Weintraut *et al.* 2022). One of the major factors in overcoming these challenges is a fundamental shift in how data is managed and utilized across operations, so that the performance of the vast, complex, ageing infrastructure systems can be understood and decisions made appropriately.

Despite generating vast amounts of data across multiple departments, ranging from sensor-based technical measurements to customer relations, water utilities often operate within fragmented and siloed data environments that hinder comprehensive analysis and strategic decision-making (Ponce Romero *et al.* 2017). Isolated data repositories prevent cross-functional integration and limit the ability to derive actionable insights (Weintraut *et al.* 2022). These silos and systems are often overly focused on purely achieving regulatory requirements, which must not be compromised, but the opportunity is then missed to derive deeper insights. This fragmentation weakens the industry's ability to leverage analytical tools such as machine learning, predictive modelling, and data fusion for operational optimization (Kyritsakas *et al.* 2023).

The aim of this paper is to explore and develop data mapping strategies that work across data silos and enable understanding and quantification of the levels of utilisation of different data, in current and potential future forms. We propose that the key to understanding the complex challenges in data utilisation is the application of advances in graph database technologies, which present a transformative opportunity for integrating and analysing complex interconnected datasets. Unlike traditional relational databases that rely on rigid schemas, graph databases represent data as nodes and relationships, allowing for a more flexible and intuitive mapping of real-world interactions (Bollen *et al.* 2021).

Building on this concept, we propose a graph-database-driven framework that is particularly suited for modelling the intricate interdependencies between raw data sources, analytical processes, and business applications within water utilities. By adopting such an approach, utilities gain the knowledge to break down data silos, uncover hidden relationships, and further enhance operational efficiency.

While the framework is designed to be general and applicable across multiple operational domains of a water utility, this paper focuses specifically on drinking water distribution systems (DWDS). This focus reflects the inherent complexity and criticality of DWDS operations, which are often characterised by ageing infrastructure, large-scale spatial distribution, limited accessibility, and the need for continuous, safe, and reliable service delivery. As such, DWDS presents a particularly compelling context in which to demonstrate the value of the proposed approach.

2. BACKGROUND

Water utilities operate within a highly complex environment where diverse datasets are generated from numerous sources. Research in multiple sectors has revealed and documented the challenges posed by the fragmentation of operational data (Grueau *et al.* 2019; Weintraut *et al.* 2022). Water utilities typically collect high-frequency sensor data from pressure or water level meters, flow sensors, and quality monitors; operational data from maintenance logs and customer complaints; and use external data like environmental data from meteorological agencies. However, these data sources are often stored in disparate systems, each with its own data format, structure, and update frequency. As a result, a holistic understanding of the operational state is difficult to achieve, and potential synergies between datasets are overlooked (Platfoot 2020; Kyritsakas *et al.* 2023).

The problem of data fragmentation is compounded by the fact that water utility operations have traditionally been managed using legacy systems designed for specific individual functions, often dominated by regulatory requirements, rather than for integrative analysis. These systems are built around relational databases and centralized data warehouses, which, although effective at handling structured data and enforcing transactional integrity, tend to enforce rigid schemas (Robinson *et al.* 2015; Medhi & Baruah 2017). This rigidity limits the ability to easily merge heterogeneous data types, leading to isolated data silos that fail to capture the underlying interdependencies (Weintraut *et al.* 2022). Sensor readings, for example, often come with varying frequencies, units, and levels of granularity that can be cumbersome to model in a relational format. Furthermore, the siloed nature of data in traditional systems limits cross-domain insights. For instance, while pressure data may be stored separately from customer complaint records, an integrated view could reveal correlations between sudden pressure drops and increased complaints, providing actionable intelligence for preventive maintenance. These missed opportunities to gain additional insight from data that is already collected are particularly important for water distribution systems.

The challenges posed by siloed data systems are well documented across various sectors, and water utilities are no exception. In siloed systems, different departments or functions within an organization maintain their own data repositories, often without standardized protocols for data sharing. This lack of integration is not merely a technological issue but also a managerial one, as organizational structures and legacy practices often reinforce data isolation. In water utilities, siloed data can result in suboptimal asset management, inefficient resource allocation, and delayed response to operational issues (Grueau *et al.* 2019; Weintraut *et al.* 2022). Consequently, utility managers often rely on incomplete, inaccurate, or outdated information (Carriço *et al.* 2020), and the fragmentation of processes and data creates inefficiencies that impede the implementation of effective management strategies (Halfawy 2008). For example, maintenance teams might rely solely on historical records without incorporating real-time sensor data, while planning departments may work with aggregated metrics that miss critical fluctuations in system performance. Research has demonstrated that breaking down these silos is essential (Dell 2005) for developing predictive maintenance models, optimizing resource allocation, and ultimately enhancing service delivery. Despite the growing recognition of these challenges, many utilities remain entrenched in legacy systems that are resistant to change, thereby limiting the full realization of data-driven management strategies.

2.1. Graph databases as a paradigm shift in data integration

Graph databases represent a shift from traditional relational models by storing data as nodes and relationships, enabling flexible representation of real-world interactions and interdependencies (Robinson *et al.* 2015; Qiao *et al.* 2024). Recent studies have also explored graph-based data models for managing sensor-enabled infrastructure and utility networks, further demonstrating the growing interest in property-graph approaches for representing complex networked systems (Bollen *et al.* 2024; Javaherian Pour *et al.* 2025). They support rapid querying of complex networks, such as tracing data influences on outcomes, which aids proactive maintenance and risk mitigation (Cheng *et al.* 2009). Additionally, graph databases excel in real-time analytics, dynamic integration of evolving datasets, and scalability, avoiding performance bottlenecks in traditional systems (Vaquero *et al.* 2014; Kumar & Huang 2020).

The adoption of graph databases in numerous sectors provides compelling evidence of their potential. For instance, in the financial industry, graph databases have been used to detect fraud by mapping complex transaction networks (Henderson 2020; Al-Mansouri 2021). In the healthcare sector, graph-based models have facilitated the integration of patient records, genomic data, and clinical trials, thereby improving diagnostic accuracy and treatment planning (Park *et al.* 2014; Turhan 2023). In supply chain management, graph models have been used to visualize and optimize complex logistics networks, leading to significant cost reductions and improved efficiency (Lawand *et al.* 2024; Rauch *et al.* 2024). In telecommunications, graph-based approaches have enabled the detection of network bottlenecks and the optimization of service delivery by mapping the interconnections between different network nodes (Lehotay-Kéry & Kiss 2020; Javier Zorzano Mier & Iglesias 2023). These applications highlight the transformative potential of graph databases in providing actionable insights from complex data ecosystems. These cross-sector applications underscore the versatility and robustness of graph databases in handling complex, interconnected data, a capability that is equally relevant for water utilities facing the challenges of data fragmentation and siloed systems.

Recent studies demonstrate early forays into incorporating graph databases within the water sector, marking some of the earliest applications of this technology in the domain. Bollen *et al.* (2021) represented and queried hydrological data across the Flemish river system, highlighting the conceptual alignment between graph structures and river networks. Qiao *et al.* (2024) proposed a comprehensive graph-based framework for modelling complex hydraulic systems, focusing on spatial, structural, and operational interconnections to support advanced water resource management applications. While these studies demonstrate improved integration and querying in hydrological or hydraulic contexts, they are narrowly scoped to physical representations and do not address broader utility operations.

2.2. Gaps and future directions for graph-database-driven approaches in water utility industry

Within the drinking water distribution sector specifically, there are even fewer studies applying graph-database-driven techniques. Rose (2015) provides one of the earliest examinations of graph databases in this domain, highlighting their use for hydraulic queries such as leak detection, valve isolation, and flow path analysis. This work emphasizes the modelling and performance advantages of graph structures over relational databases for network connectivity tasks. However, like Bollen *et al.* (2021) and Qiao *et al.* (2024), Rose's study remains rooted in using graph database techniques for infrastructure representation, with limited exploration of how graph-database-driven models can support data flows and analysis supporting higher-level utility operations, cross-functional process mapping, or data-driven decision-making.

Based on the existing research in the water industry, it is evident that current graph-database-driven applications often target specific technical challenges in isolation (e.g. leakage) rather than supporting holistic operational intelligence. One of the key limitations to deeper analysis is the absence of standardized methodologies for integrating heterogeneous data sources into a coherent graph data model. This absence restricts the ability of utilities to leverage their full data ecosystem for cross-cutting insights, proactive decision-making, and strategic planning.

In this study, we recognize this as an opportunity and seek to address these gaps by proposing a novel approach that maps data and process flows pertaining to DWDS operations, using a graph database platform. By doing so, we aim to demonstrate how it introduces a more integrated and systemic perspective on data utilization in water utilities, moving beyond isolated technical applications toward a foundation for strategic, data-informed utility management.

3. METHODOLOGY

In this section, the methodology used to develop and implement the three-layer framework is presented. The focus is not on mapping the physical infrastructure of water distribution, but rather the information infrastructure, which is the data flows that support intermediary processes and, in turn, enable core business operations in drinking water distribution. The methodology comprises a graph-database-driven approach for mapping the data flows, processes and business operations as three distinct layers, defining their relationships, and applying analytical techniques to extract actionable insights.

The three-layer framework was applied in a case study to encapsulate the operations of Scottish Water, a water utility company serving 2.6 million households and 160,000 businesses across Scotland. Drinking water is supplied from 229 water treatment works, which process raw water from reservoirs, lochs, and rivers to meet quality standards regulated by the Drinking Water Quality Regulator for Scotland. The company operates an extensive and complex water distribution network, including 30,586 miles of water mains, which collectively deliver 1.49 billion litres of water per day, (Scottish Water 2024). They collect and manage a typically vast array of data, intermediate and business processes to manage these DWDS.

3.1. Conceptualization of the three-layer framework

The methodology proposed in this paper is designed to recognise, and specifically to exploit the potential of mapping and understanding data flows and utilisation from data types, via processing, to application. We propose a three-layer connectivity structure to capture management functions of a water utility company that are data driven, particularly for analysing data flow and process utilisation (Figure 1). This structure is designed to be generalizable and adaptable across different functional areas of water utilities, including both clean water and wastewater operations. For the purposes of this study, the framework is demonstrated and validated within the context of operational environments specific to DWDS. The three layers broadly capture the raw data that is collected ('Data' in Figure 1), the intermediary processes that manipulate this data including calculations, models, and other analytics ('Intermediary Processes' in Figure 1), and the business outcomes and applications that directly support decision making and operations ('Business Applications' in Figure 1).

The development process of this framework was informed by a series of discussions with data managers and business personnel across Scottish Water, ensuring that the resulting graph representation reflects real-world operational structures, information flows, and decision-making pathways.

The discussions were carried out sequentially by framework layer, beginning with Data (Layer 1), then moving to Intermediary Processes (Layer 2), and finally Business Applications (Layer 3). This ordering ensured that discussions about processes and applications were grounded in an understanding of available data. For each layer, an initial broad set of questions was asked to map known elements, followed by probing questions to capture overlooked or less-visible components. Participants were selected from various departments including asset management, network operations, hydraulic modelling, water quality, IT/data management, and planning to ensure perspectives across the functions of the utility relevant to DWDS.

These engagements provided critical insights into the practical use of data, existing process interdependencies, and the challenges faced in integrating information across domains. The three-layer framework was then developed to map individual dataset origins, intermediary transformations, and business applications.

Although the framework is constructed by initially mapping existing datasets to intermediary processes and business applications, it is not intended to represent a purely bottom-up workflow. Once represented within the graph database, the relationships can be interrogated in both directions, allowing users to trace how data contributes to operational decisions

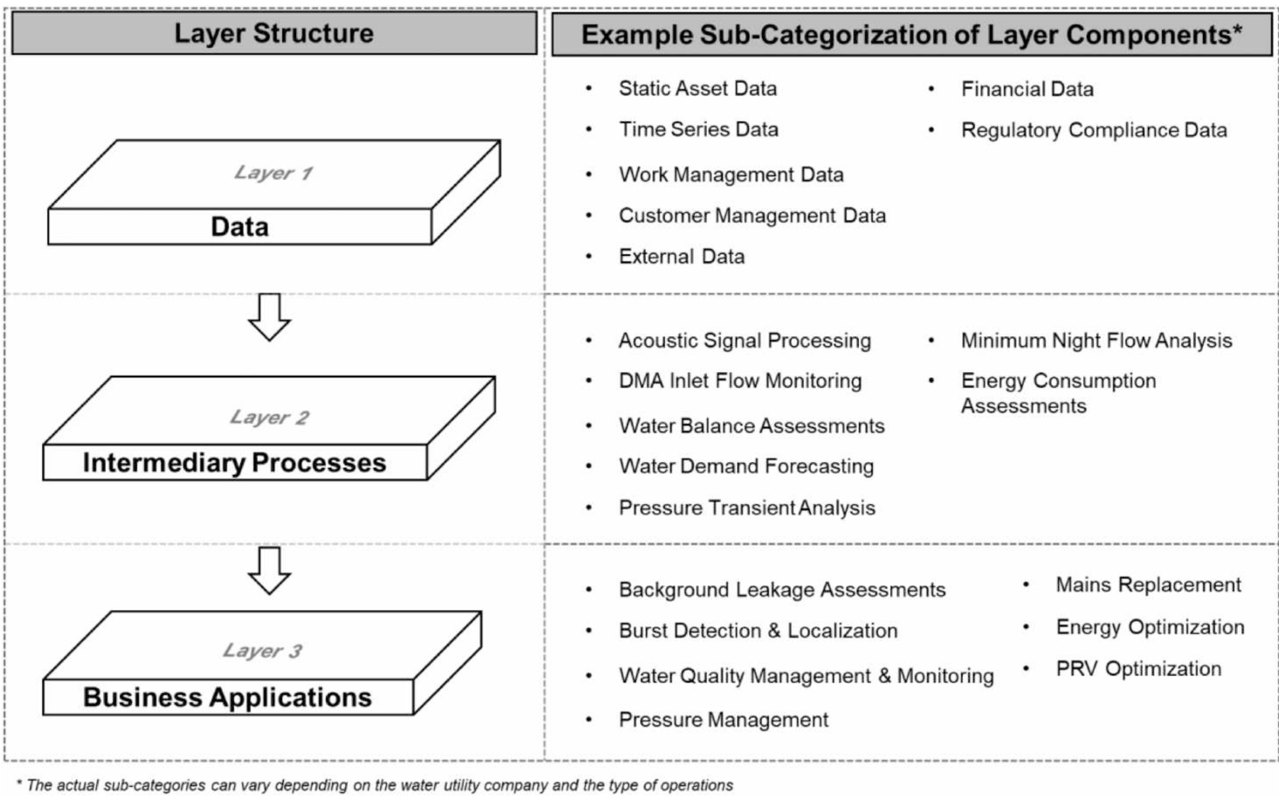


Figure 1 | The three-layer conceptualization of the framework.

or, conversely, to identify the data and processes required to support specific business applications. In this way, the framework supports both exploratory analysis of current data utilisation and purpose-driven evaluation of information needs. This perspective is consistent with purpose-driven information management approaches such as the Level of Information Need concept defined in ISO 7817-1:2024, which emphasises specifying the quantity, quality, and granularity of information required for particular purposes.

3.1.1. Identification of datasets for creating layer 1 of the framework

The first stage focused on identifying the types of data currently collected by water companies for DWDS operational and planning purposes. This includes static or asset data, such as pipe network layouts, pipe materials, asset age, and the locations of measurement and sensing instruments, as well as time-series data such as flows and pressures. It also includes operational data from work management systems and operational logs, which provide insights into maintenance activities, repairs, and incident responses. In addition, environmental data from external sources (e.g., rainfall and temperature records) and customer-related data (e.g., feedback and complaint logs) were considered.

Table 1 lists the types of questions asked when creating the 'Data' layer, along with their intended purposes. These questions provided a structured guide for the discussions, which often extended to consider how datasets are used within analytical processes and operational applications across the different framework layers, reflecting the interconnected nature of operational workflows within the utility. As shown in the table, care was taken to account not only for datasets that are extensively used for various processes by water companies but also for those used relatively infrequently. To achieve this, we specifically inquired about datasets that participants knew existed but were rarely or never accessed, often due to barriers such as format, availability, or awareness.

3.1.2. Intermediary processes for creating layer 2 of the framework

Once a comprehensive dataset list was constructed, the next step was to identify the intermediary processes into which these data feed. Layer 2 (Intermediary Processes) encompasses activities that transform raw data into structured outputs, such as calculations, statistical summaries, visualisations, and model predictions. These processes were defined as those requiring computational and analytical tasks, possibly converting data into more usable forms of information but stopping short of informing any specific decision making or reporting.

If we were to compare the three layers of this framework to the Data–Information–Knowledge–Wisdom (DIKW) pyramid (Rowley 2007), commonly used in the data science domain, this layer corresponds to the transition from raw data to information, where meaning is first derived through structured processing.

The types of questions asked during the construction of Layer 2 of the framework, and their intended purposes are shown in Table 2. These were designed not only to capture the common intermediary processes, but also to elicit expert insights into where bottlenecks and inefficiencies are most frequently encountered during analytical workflows.

Intermediary processes for DWDS include tasks such as performing water balance assessments, conducting minimum night flow analysis, carrying out hydraulic modelling and simulations, conducting GIS-based spatial analysis and analysing customer complaints. In some cases, advanced data fusion techniques, anomaly detection algorithms, or temporal disaggregation methods may also be applied, particularly when data is sparse or noisy.

Table 1 | Guiding questions for constructing the 'Data' layer of the framework

Example Questions	Purpose
What datasets do you generate or use in your daily operations?	Inventory of available datasets
Where is each dataset stored, and in what format (e.g., GIS, spreadsheets)?	Understand accessibility and formats
How frequently is this dataset updated or refreshed?	Capture temporal characteristics
Who is responsible for maintaining and ensuring the quality of this dataset?	Identify ownership and governance
Do you combine this dataset with others? If so, how?	Map dataset linkages
Are there datasets you know exist but are difficult for you to access?	Expose silos and barriers
Are there datasets you wish you had but are not currently collected?	Identify gaps and opportunities

Table 2 | Guiding questions for constructing the 'Intermediary Processes' layer of the framework

Example Questions	Purpose
Which datasets feed into the analyses or models you perform?	Establish input–process linkages
What types of analyses or calculations do you routinely conduct?	Identify process types
How often are these analyses carried out, and under what circumstances?	Capture process frequency and triggers
Are these processes automated, semi-automated, or manual?	Assess efficiency and scalability
Which teams or processes rely on the outputs of your work?	Map process dependencies
What challenges or bottlenecks do you face in carrying out these analyses?	Identify inefficiencies
Are there analyses you think should be done but are not currently performed?	Reveal missed opportunities

3.1.3. Business applications captured in layer 3 of the framework

One or more intermediary processes inform downstream tasks, which constitute the business applications carried out by water companies. An important consideration in applying the three-layer framework is determining whether an activity should be represented as an intermediary process (Layer 2) or as a business application (Layer 3). While the two are closely connected, the distinction was drawn by recognising that Layer 3 (Business Applications) represents the stage at which outputs from Layer 2 are interpreted and acted upon to inform or enable operational or strategic decisions or reporting.

In terms of the DIKW hierarchy, this layer aligns with 'Knowledge' (and, where interpretation involves foresight or judgement, 'Wisdom'), as information is applied to guide business actions and operational choices. In the context of DWDS, examples include deciding when and where to dispatch repair crews for burst localisation and repairs, scheduling mains replacement, adjusting pressure zones based on modelled results, water conservation programs and implementing emergency response actions.

It was found that, just as multiple datasets can be combined to perform intermediary processes, multiple intermediary processes often collectively enable end-user applications, and that intermediary tasks can feed directly to applications as well as into other intermediary analysis and then into other applications. Table 3 lists the types of questions that were asked to understand business applications, not only to capture the common applications within a water utility but also to understand the consequences and risks that may arise if the outputs from relevant intermediary processes are not properly considered.

3.2. Graph-database-driven mapping framework

To map the relationships between various types of data collected by water utility companies, intermediary processes, and final business applications requires an approach that can efficiently handle complex, interconnected data structures. To accomplish this, we selected Neo4j (Robinson *et al.* 2015), a Java-based graph database that stores data as nodes, relationships, and associated properties in a graph structure. Although platforms such as Neo4j are often used for constructing knowledge graphs, the objective in this study is not to develop a semantic knowledge graph with formal ontologies, but rather to use a property graph database to represent and query operational relationships between different components

Table 3 | Guiding questions for constructing the 'Business Applications' layer of the framework

Example Questions	Purpose
What decisions or business outcomes are supported by your analyses?	Link intermediary processes to business applications
Can you describe a recent example where analysis outputs influenced an operational decision?	Capture concrete applications
Which business functions depend most critically on these outputs?	Map dependencies and priorities
What would be the consequence if this analysis was delayed or unavailable?	Assess criticality and risk
Are there upcoming or emerging challenges (e.g., climate resilience, regulatory changes) that current processes do not yet support?	Identify future needs

of the utility data ecosystem. Neo4j was chosen for its property graph model and expressive query language, which support the efficient representation and analysis of the interconnected structures central to this study.

3.2.1. Node level attributes

In the graph database representation of the data flows and processes, each element within the three layers of the structure (i.e., Data, Intermediary Processes, and Business applications) were stored as nodes. At the node level, additional properties (or attributes) related to the nodes were specified. For Data nodes (layer 1), these include data categories, unique identifiers, information about data sources, and data formats. Attributes related to Intermediary Processes in layer 2 include properties such as process type/description, processing algorithms, and software used. For business applications (layer 3), properties such as application function, impact metrics, and stakeholders were included. Figure 2 shows a detailed breakdown of the node-level attributes proposed for this graph-based framework.

3.2.2. Edge level attributes

The flow from Data to Intermediary Processes and then to Business Applications was mapped using links (or relationships). Various data sources aid intermediary processes, and multiple data types often feed into a specific intermediary process. A unique labelling identifier, *FEEDS_INTO*, was assigned to all relationships extending from Data to Intermediary processes. Under this label, several attributes were specified, such as *current_usage*, *potential_usage*, *industry_benchmark* (Figure 2). Similarly, intermediary processes that support business applications were mapped with a unique relationship identifier *SUPPORTS*. In the Neo4j implementation, relationships are also assigned unique identifiers within the database; however, in this study the analysis focuses on the type of connection between nodes (e.g., FEEDS_INTO or SUPPORTS) rather than on individual relationship identifiers. Further attributes such as *current_usage*, *potential_usage*, and *maturity* were included under this.

Two key attributes that describe both *FEEDS_INTO* and *SUPPORTS* relationships are *current_usage* and *potential_usage*. By categorizing the frequency or scale of usage (e.g., High, Medium, Low, or None) and opportunities for expansion (e.g., High, Moderate, Low or Untapped), gaps and potential opportunities can be uncovered through queries and various analytical methods. By incorporating these layers of metadata and attributes, the graph database will not only represent the connectivity between different components but also provide richer contextual insights. This enables more nuanced analyses, such as identifying critical data sources, detecting redundant processes, or uncovering potential efficiency improvements.

Populating these node and edge-level attributes presented challenges due to the inherent complexity of utility operations and the diverse nature of departmental workflows. Accurately specifying attributes requires either a strong, system-wide understanding of how data, processes, and business outcomes interconnect, or access to detailed input from domain experts across different parts of the organization. Findings from discussions carried out with personnel from relevant departments combined with practical judgement was used when populating attributes pertaining to the node and edges, particularly the ones which involves qualitative responses. For example, when populating attributes such as *current_usage* (the extent to which a dataset or process is presently used) and *potential_usage* (the perceived scope for additional application), the responses ranged from explicit ('this process is essential to our daily operations') to more tentative or speculative ('this dataset could be very useful if we had the capacity to integrate it') opinions. Such responses were translated into categorical values such as Low, Medium, High. For instance, *current_usage* values were assigned based on how frequently and broadly the dataset or process was reported to be in use. If multiple interviewees described it as central to daily workflows, it was classified as High. If it was used occasionally or only by a single team, it was classified as Medium. If it was rarely used or mentioned only as a legacy dataset/process, it was classified as Low. Similarly, for *potential_usage*, categorical values were assigned by reflecting both interviewee suggestions and the researcher's synthesis of responses. Where multiple participants independently indicated that a dataset or process could support additional applications (e.g., 'this could be valuable for leakage analysis'), it was classified as High. Where suggestions were more tentative or mentioned by only one participant, it was classified as Moderate. Low was used where participants expressed little or no perceived scope for further application.

3.3. Analytical techniques

Several analytical techniques were used for gaining insights about where gaps and opportunities lie in terms of data and process utilization.

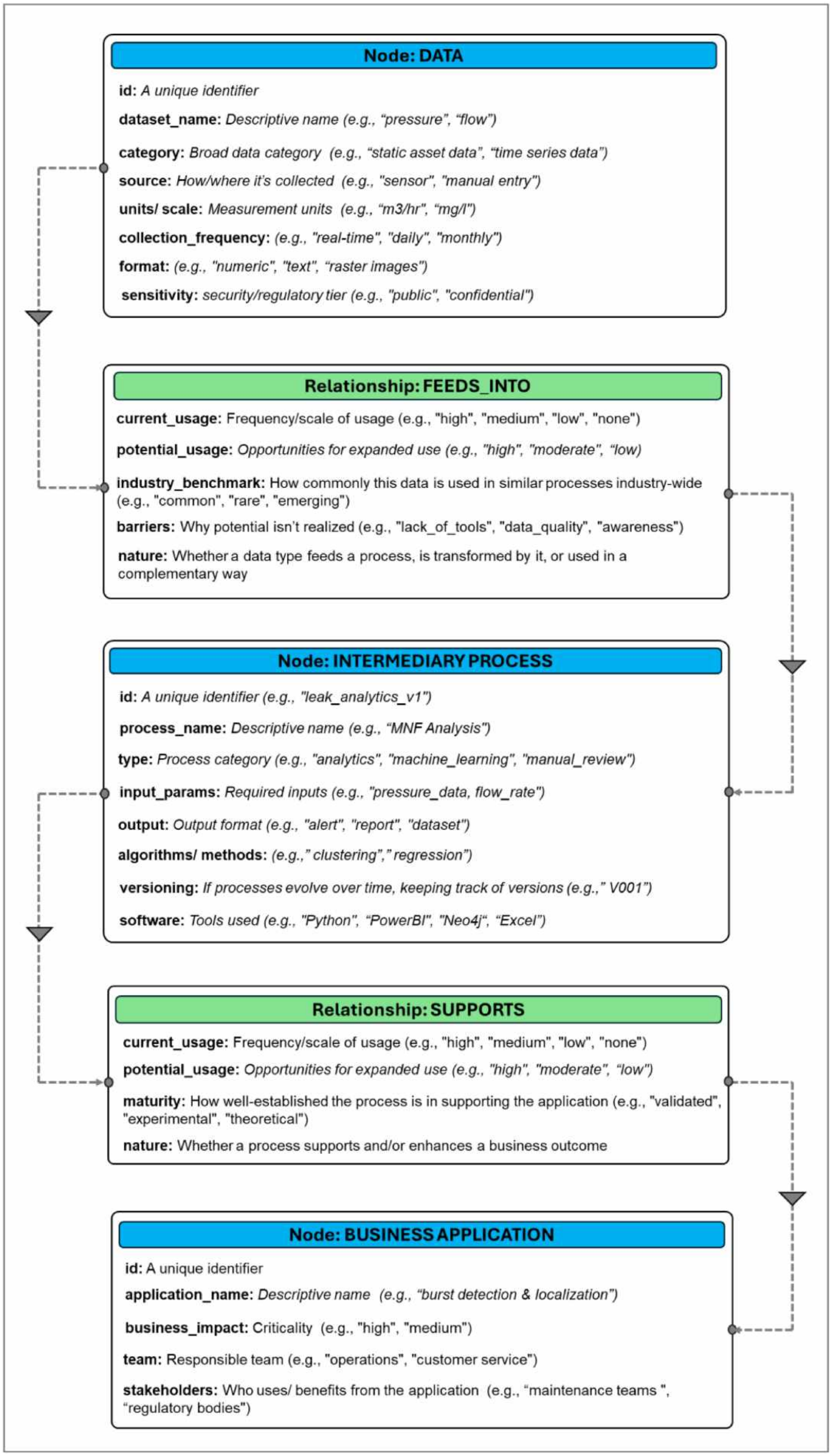


Figure 2 | Three-layer mapping framework as implemented in Neo4J.

3.3.1. Node centrality measures

Centrality algorithms help identify which nodes are most important or influential in a network. In a graph database, a node, such as a specific data type or an intermediary processing step, with high centrality could indicate its critical role in many downstream applications.

Centrality measures vary in approach. For example, degree centrality identifies the importance of nodes based on the number of connections they have, while PageRank determines which nodes receive a lot of 'influence' by being connected to other important nodes (Zhang *et al.* 2022; Chicho & Mohammed 2023). Degree centrality is an important measure for this application. For example, any measurement taken with instruments is prone to degradation and failure, and understanding when such data has high degree centrality indicating a large number of downstream intermediary tasks and business applications that rely on it is crucial. Such data sets can then be monitored for accuracy and precision; failure to do so would result in several intermediary and business processes being vulnerable. Alternatively, low degree centrality can help identify data types and intermediary processes that receive relatively little attention, with fewer connections. In terms of data, this could reveal specific data types that are underutilized but may have the potential to be further integrated to support certain intermediary actions. While not all underutilized data may prove useful, this insight provides the water company with an opportunity to explore creative ways to enhance operations and decision-making. By actively utilizing these datasets, the company can improve downstream intermediary tasks. Similarly, intermediary processes with very few edges may indicate that they are not widely integrated into business applications, possibly highlighting areas for improvement or rationalisation.

Extending this further, business applications themselves can also exhibit low centrality, indicating that they are poorly supported by upstream processes. This may reflect operational areas where decisions are made with limited evidence or weak foundations. Identifying such low-centrality business application nodes can therefore help water utilities uncover potential blind spots in their decision-making processes. These gaps may point to critical business functions that are either underserved by current data collection and processing strategies or where integration between operational data and strategic applications is lacking. Addressing these gaps could drive targeted improvements, enabling more robust, data-informed decision-making across the organisation.

3.3.2. Relationship attribute-based graph querying

An intentional advantage of our proposed three-layer approach is that it allows for the specification of a comprehensive set of edge attributes (similar to defining attributes at node levels). This enables running queries based on specified attribute values to extract various insights based on connectivity. For example, the attributes *current_usage* and *potential_usage* can be used to identify data types and intermediary processes with low *current_usage* but moderate to untapped *potential_usage*. This capability is especially valuable when complex systems, managed by various divisions of a company, are incorporated into a graph database. Such queries can help recognize specific data types and processes with significant potential to enhance operations further. In cases where the resulting queried subgraphs became visually dense, the query outputs were also examined in tabular form within Neo4j, which allowed the underlying node-relationship connections to be inspected more rigorously in a structured format.

3.3.3. Impact analysis through failure simulations

Impact analysis evaluates how disruptions or changes in one part of the network might ripple through the rest of the system. Graph databases such as Neo4j facilitate techniques like node or link removal simulations, allowing nodes or links to be temporarily excluded from queries to observe the resulting network paths and connectivity. This can typically be done virtually to simulate the conditions that arise if one or more data streams or critical intermediary processes fail. The modified graph can then be analysed to assess the impacts by closely examining selected portions of the resulting structure.

Such critical assessments can be conducted proactively to test the system against potential data losses and evaluate their impact on downstream processes using graph databases. This approach not only enhances preparedness but also allows for the implementation of precautionary measures in advance, ultimately increasing the resilience of the overall system and related operations within the company.

3.4. Application of the three-layer framework to Scottish Water

For this application, a targeted subset of attributes was selected from the conceptual framework presented above to align with Scottish Water's available datasets and operational needs. For nodes, in addition to a *name* attribute, this included attributes such as *category* (for data nodes), *type* (for intermediary processes), and *business_impact* (for business applications). For relationships, attributes such as *current_usage* and *potential_usage* were retained to capture the extent and opportunities of data and process utilization. This focused implementation represents a practical subset of the conceptual framework, enabling the graph database to remain both manageable and analytically useful for uncovering inefficiencies, dependencies, and opportunities within Scottish Water's DWDS.

4. RESULTS AND ANALYSIS

This section presents the results of applying the proposed three-layer framework to Scottish Water's DWDS. The framework was implemented in Neo4j, resulting in a complex graph structure comprising 127 nodes and 911 links. Figure 3 illustrates this network, showing the intricate web of connections between Data sources (60 nodes), Intermediary Processes (54 nodes), and Business Applications (13 nodes). The graph highlights the interdependencies within Scottish Water's DWDS data ecosystem, demonstrating how raw data flows into processes, is transformed into information, and subsequently informs operational actions and decisions. The dense structure of the network, with its overlapping relationships, reflects the complexity and interconnected nature of data utilisation across the company.

To extract meaningful insights from this graph, we employed three analytical approaches:

1. Node centrality measures – Identify the most and least influential data types and processes within the system.
2. Relationship attribute-based graph querying – Enable targeted analysis of data utilization patterns.
3. Impact analysis through failure simulations – Assess the consequences of disruptions in key data flows.

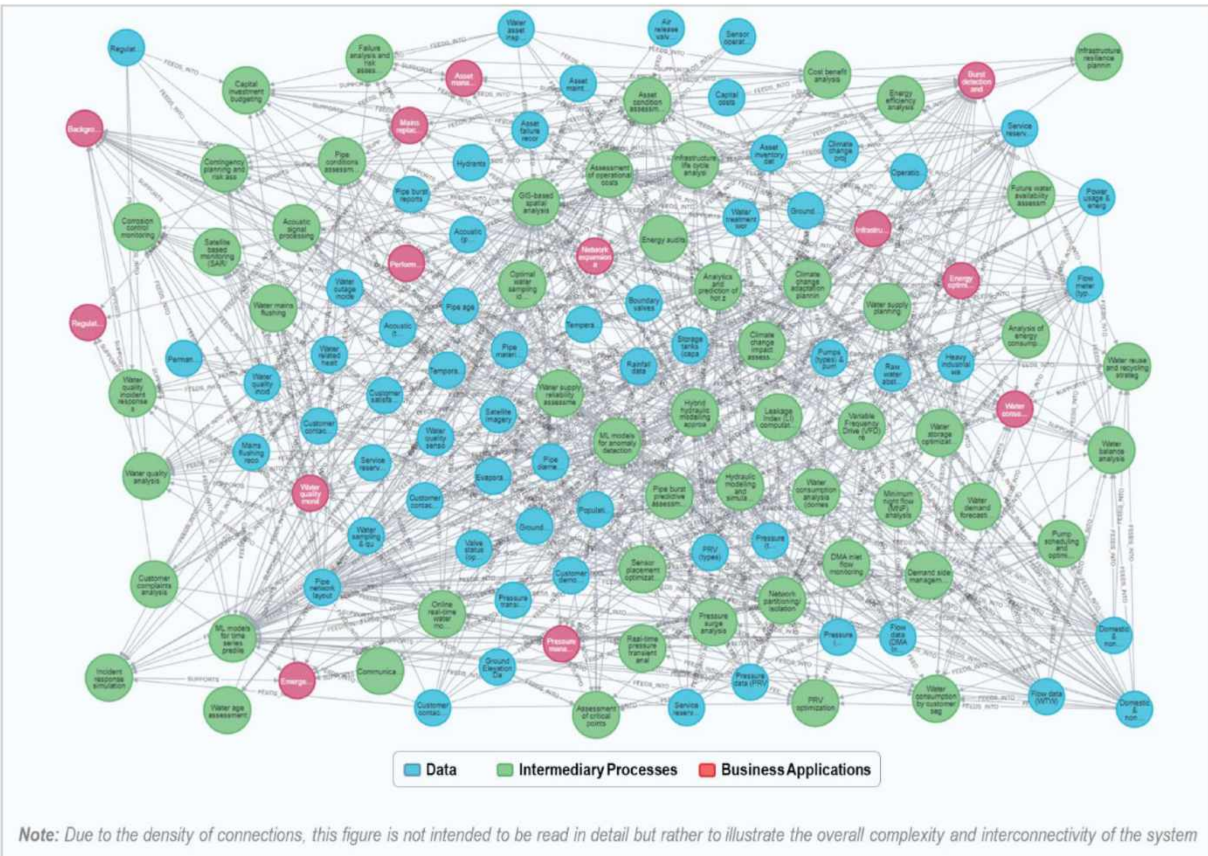


Figure 3 | Graph database created using Neo4j, representing data and process flows for Scottish Water's management of their DWDS.

By systematically applying these techniques, we uncover examples of underutilized data sources, inefficiencies in data flow, and critical dependencies that could support more effective data-driven decision-making. The following subsections provide a detailed analysis and discuss the key findings derived from these methodologies.

4.1. Insights based on node centrality measures

Centrality analysis was first used to identify the datasets and processes that exert the strongest influence on downstream activities. Degree centrality, in particular, provides a straightforward but powerful way to reveal which data nodes connect to the largest number of processes, and therefore contribute most significantly to business applications. Within the time-series data *category*, flow data emerged as one of the most important datasets in the system. As shown in Figure 4, flow data feeds into 21 intermediary processes, representing 38.9% of all intermediary processes considered in the main graph database. These processes, in turn, support 11 different business applications, emphasising the pivotal role of flow data across both operational analytics and decision-making functions. In a similar manner, pressure data also proved to be highly influential, contributing to 17 out of 54 intermediary processes (31.5%), which again link to 11 business applications.

The reliance of so many downstream processes on flow and pressure time series datasets highlights the need for high accuracy, completeness, and reliability in their collection and storage. Any inaccuracies or disruptions in flow or pressure data

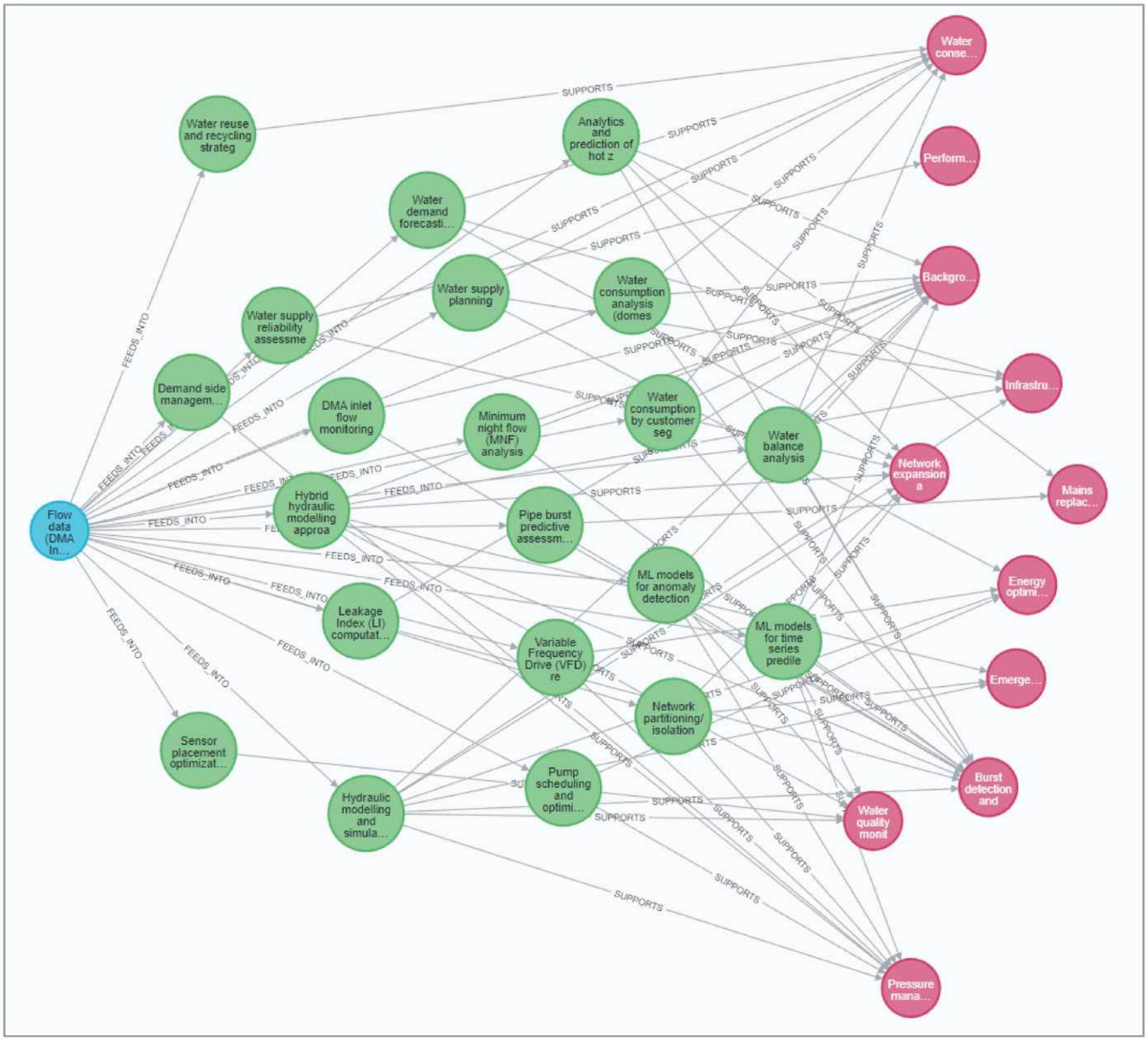


Figure 4 | Query result showing the DWDS flow data utilization and the corresponding process flow to applications.

would propagate widely through the system, reducing confidence in analytical outputs, undermining operational performance and decision making.

Beyond time-series datasets, static asset data was also examined using degree centrality. Here, the GIS pipe network layout was identified as the most critical dataset. As illustrated in Figure 5, this dataset underpins 48 of the 54 intermediary processes, making it by far the most connected data node in the system. This result is unsurprising, given that most processes rely directly or indirectly on accurate knowledge of asset location, connectivity, and spatial extent. However, the degree of centrality observed underscores just how fundamental this dataset is for nearly every aspect of DWDS management. Inaccuracies or errors in the network layout would therefore have far-reaching consequences, affecting the validity of many intermediary processes and the reliability of the business applications they support. Given its critical role and extensive connectivity, this is a dataset that warrants continued investment to ensure its accuracy, completeness, and overall veracity, as even minor inconsistencies can propagate widely through the decision-support chain.

Applying degree centrality in an inverse manner also revealed datasets connected to relatively few downstream intermediary processes, such as satellite imagery. These results provide insights into where opportunities may lie for expanding data utilisation and are easy to visualise in the graph database.

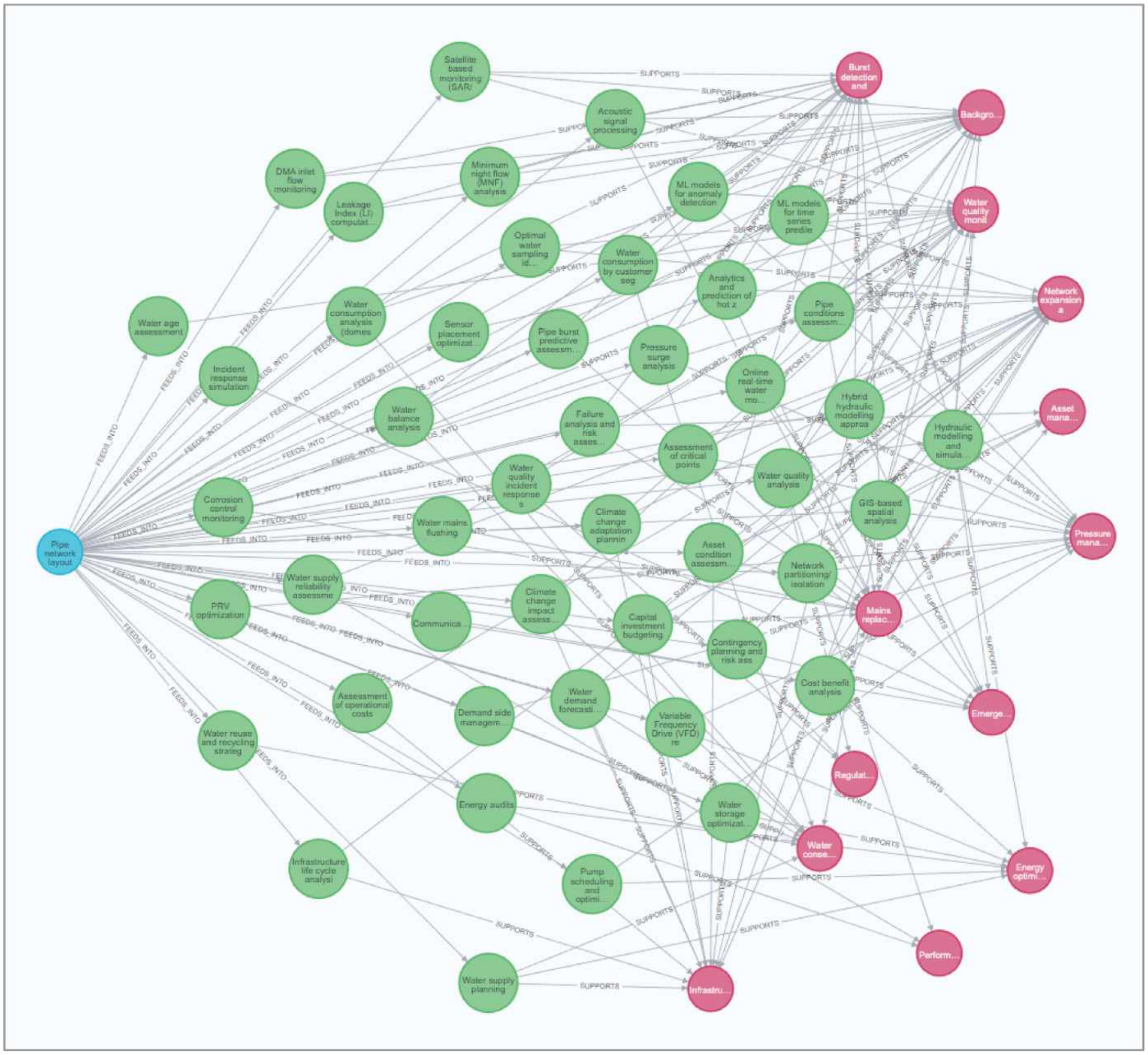


Figure 5 | Query result showing the pipe network layout data utilization and the corresponding process flow to applications.

4.2. Identification of opportunities using relationship-attribute-based graph querying

In this section, we demonstrate the insights that can be derived by applying graph querying methods to the three-layer framework, with a particular focus on the relationship attributes between nodes. The attributes *current_usage* and *potential_usage* were queried for relationships extending from Data nodes to Intermediary Processes (*FEEDS_INTO*) and from Intermediary Processes to Business Applications (*SUPPORTS*).

Taking the business application of burst detection and localization as an example, Figure 6 illustrates the complete graph network from Data through Intermediary Processes that connect to this specific application. Extracting meaningful insights in this case requires human input to guide the process through a series of iterative, targeted queries. For instance, the initial query was refined to identify intermediary processes with greater potential for adoption by filtering the graph to show only those links where *current_usage* was set to Low and *potential_usage* was High or Moderate, specifically for the connections between intermediary processes and business applications. This iterative workflow is analogous to the Human-in-the-Loop (HITL) concept commonly employed in computer science, where humans actively participate in the feedback loop of a model, algorithm, or process to refine outcomes and ensure context-specific relevance.

Figure 7 shows the refined graph network generated by this query. This type of analysis makes it possible to quickly recognize Intermediary Process nodes that meet the specified conditions. Four Intermediary Process nodes (green nodes) are highlighted as underutilised: satellite-based monitoring, machine learning anomaly detection, machine learning time series prediction, and pipe condition assessment. These represent processes where opportunities exist to integrate them more fully into decision-making workflows than is currently the case.

Relationship-attribute-based graph querying was also applied to highlight specific types of data with greater potential for use than their current levels of adoption suggest. Again, considering burst detection and localization as the application of interest, we investigated types of Data connected at Low levels of *current_usage* to Intermediary Processes supporting this business application. The resulting graph shown in Figure 8 highlights such Data types. Manual inspection of the resulting graph revealed examples including water quality data, power and energy usage data, and satellite imagery. While

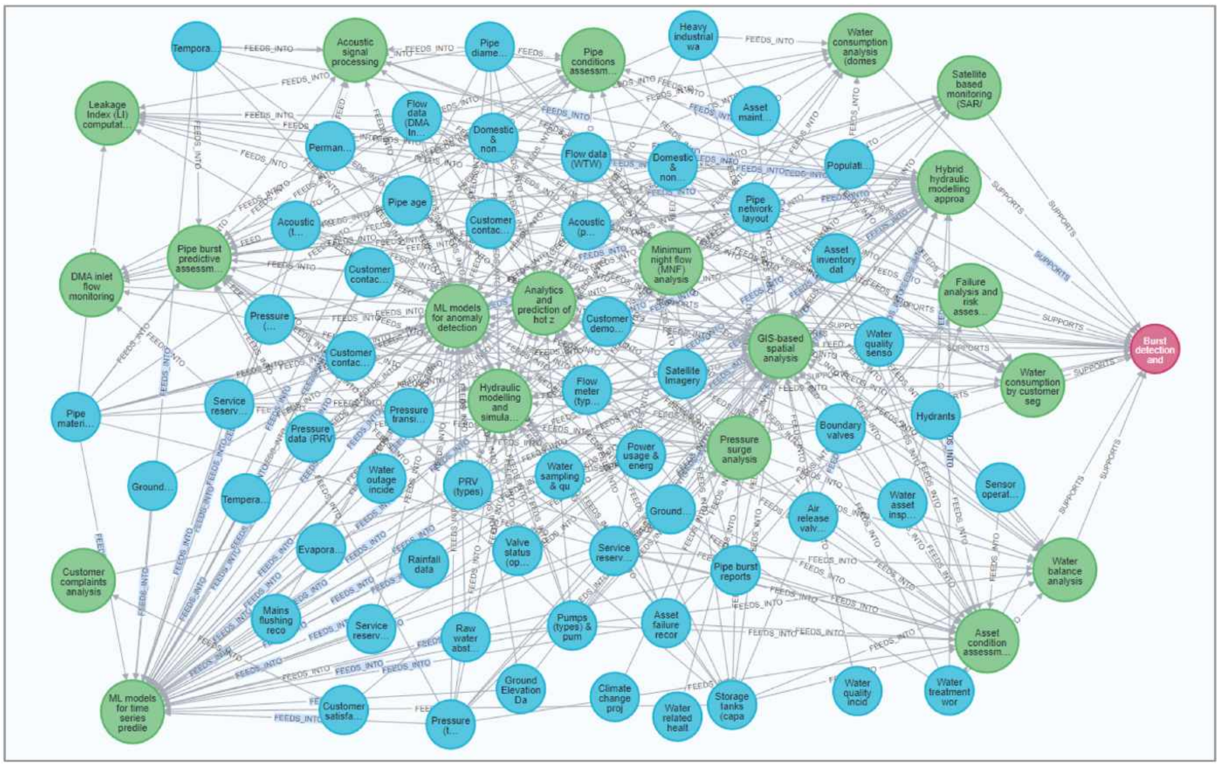


Figure 6 | Query result showing the network of all Intermediary Processes and Data types that are connected to burst detection and localization business application.

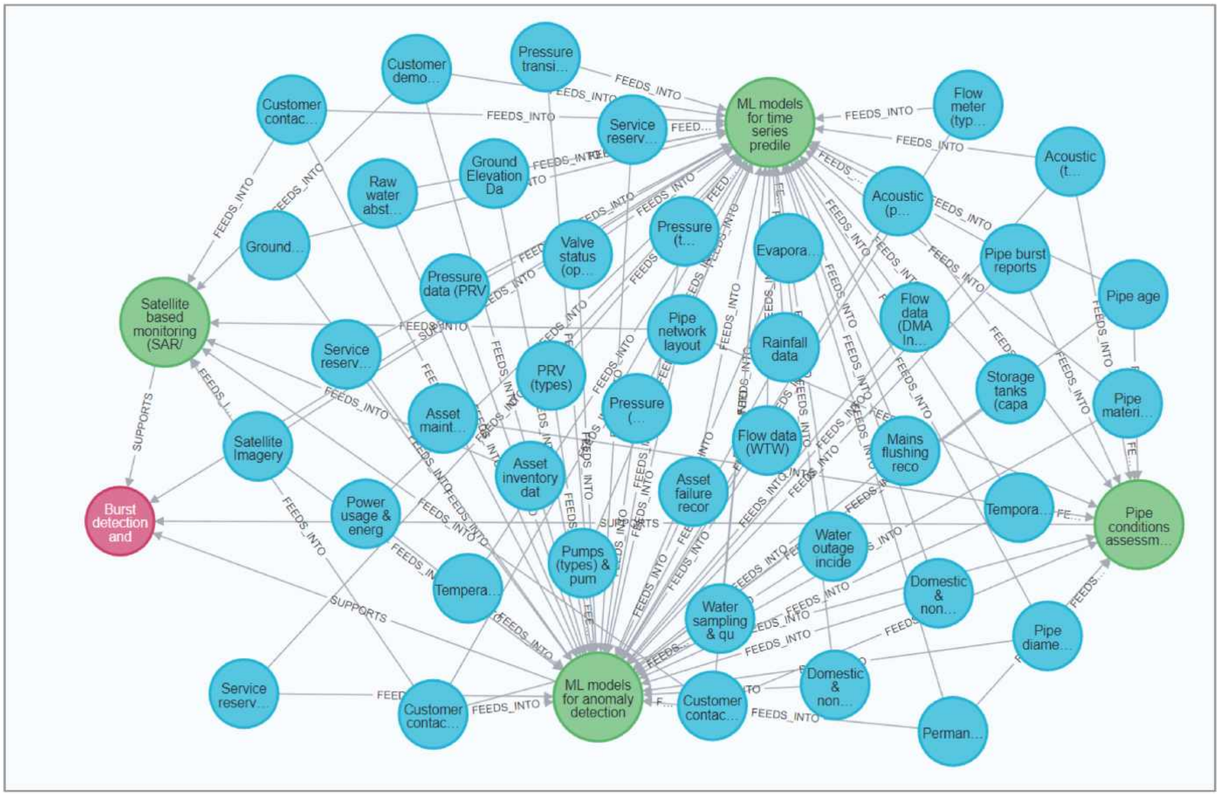


Figure 7 | Query result showing the filtered graph network of burst detection and localization business application, network where *current_usage* is Low and *potential_usage* is High or Moderate for relationships between Intermediary Processes and Business Applications.

currently underutilized, these datasets have significant potential for broader integration, particularly given the advances in data science and machine learning where their applicability has been widely demonstrated. The Intermediary Processes these datasets connect to (as shown in Figure 8) further emphasize how such data could support enhanced analytical workflows and decision-making in burst detection and related applications.

4.3. Impact analysis through simulation of failure modes

Graph databases enable various types of hypothetical simulations to be conducted virtually to assess the impact of failures in specific data types or processes within the network. For instance, if pressure data were to become temporarily unavailable, an impact assessment could help determine where a data failure would result in a process breakdown for processes that are: (i) solely dependent on pressure data (ii) dependent on multiple data inputs but critically dependent on pressure data, and (iii) functional without pressure data if other available data inputs can compensate for the missing pressure data, albeit with potentially reduced accuracy or confidence.

Considering the case where pressure data is assumed to be temporarily unavailable, and using this as the basis for an impact analysis, the results of this simulation are summarized in Figure 9. Figure 9(a) shows the intermediary processes that directly depend on pressure data. These are the first elements of data analysis to be affected in the event of such a failure. Extending this view to business applications, Figure 9(b) highlights the downstream applications that would in turn be impacted (such as burst detection and localization and pressure management), since their functioning relies on the intermediary processes identified earlier.

By conceptually removing the data linkages from the pressure data node, the resulting loss of connectivity can be observed in the overall network structure, illustrating how disruptions in a single critical data source can cascade through multiple intermediary process layers. In this particular case study, no intermediary processes were found to rely exclusively on pressure data as their sole input. Thus, while none became entirely non-functional, the analysis does show a reduction in the robustness of several processes.

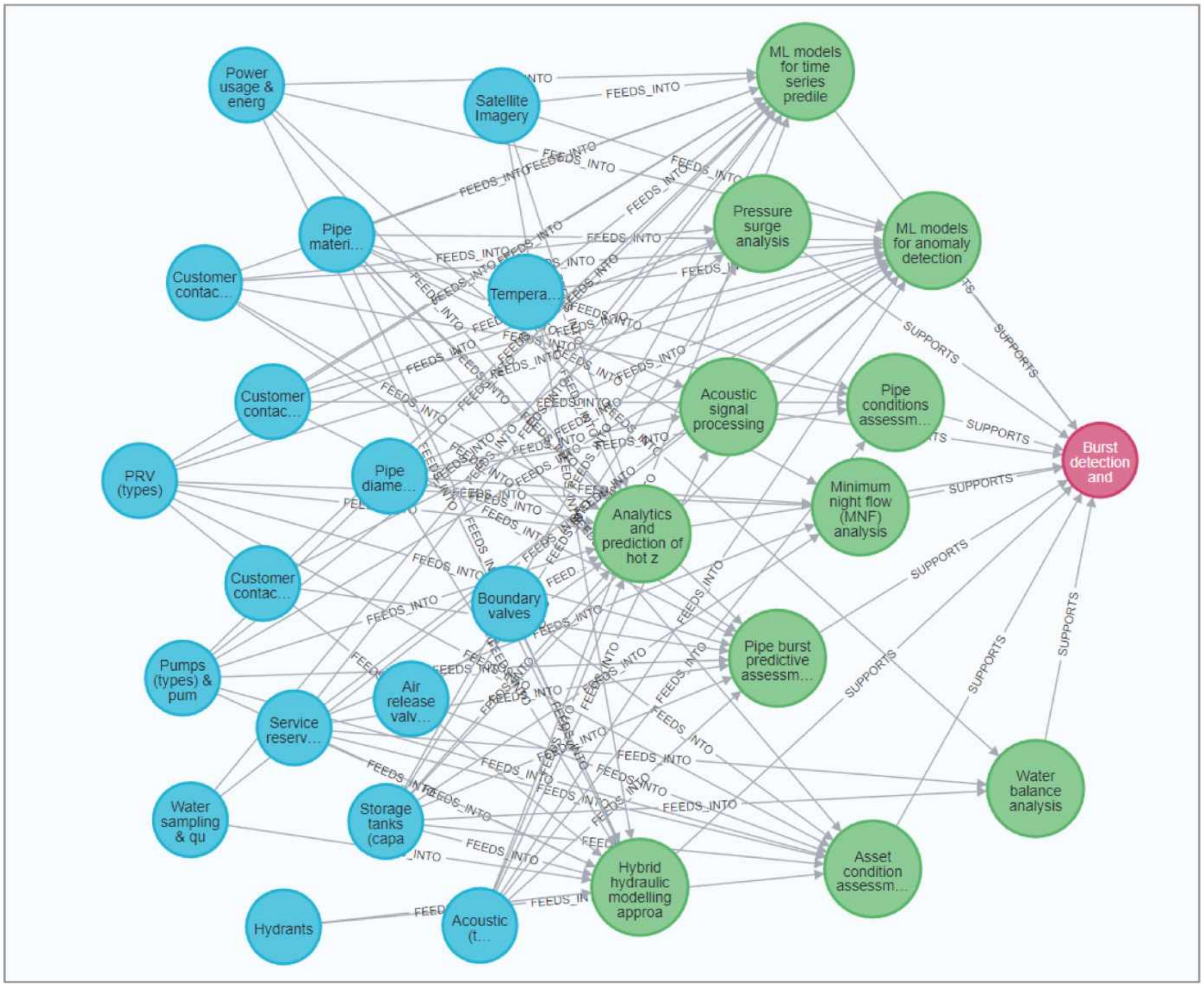


Figure 8 | Query result showing the filtered graph network of burst detection and localization business application, where the *current_usage* is considered to be Low for links between Data and Intermediary process nodes.

The degree of reliance on pressure data is further revealed by inspecting the relationship attributes, as depicted in Figure 9(c). For instance, processes such as minimum night flow (MNF) analysis are classified as having a Medium reliance on pressure data according to the original attribute value assignments. This indicates that while pressure data contributes significantly to the process, MNF analysis can still function at a basic level if that data becomes unavailable. However, MNF depends on flow measurements and the pipe network layout (both with a High reliance), so if all flow data were to fail alongside pressure data, MNF functionality would be severely compromised. This example illustrates how the simulation of failure modes provides insights not only into immediate disruptions but also into the relative criticality of different data types within the network.

5. DISCUSSION

As water utilities continue to generate vast amounts of data across different operational and business domains, the challenge of integrating these datasets into a unified decision-support framework has become increasingly apparent.

The development of the graph database presented in this study was undertaken to encapsulate the data and process flows of Scottish Water's DWDS. This was achieved through discussions with data managers and business personnel across Scottish Water. These discussions played a crucial role in identifying key datasets, defining intermediary processes, and understanding the pathways of how different business applications consume data. They were particularly important for the more subjective attributes such as level of usage and potential. One of the advantages of this graph-database-driven

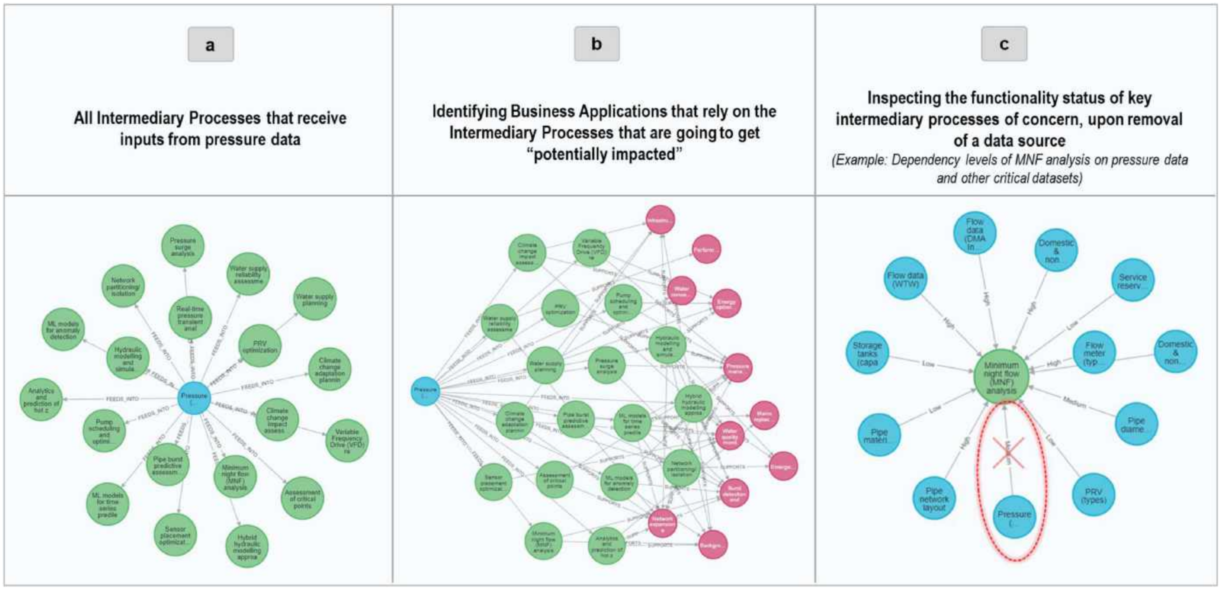


Figure 9 | Results pertaining to various stages of impact analysis process for the specific application of pressure data node removal.

approach is its adaptability: when disagreements emerge regarding the importance or ranking of certain data sources, the graph structure allows for quick updates and iterative refinements, making it a dynamic and continuously evolving decision-support tool.

The study revealed both strengths and opportunities in Scottish Water's data integration practices. Certain data types, such as flow, pressure, and network layout were found to be extensively utilized across multiple processes, highlighting their critical role in decision-making. Consequently, ensuring the highest possible accuracy for these datasets is essential due to their significance in the overall process chain.

Opportunities were also identified in both data and intermediary process domains, not only when examining the connectivity structure, but also when probing deeper into current and potential usage levels through relationship-based graph querying. For example, when considering burst detection and localization as the business application of interest, intermediary processes such as satellite-based monitoring and machine-learning-based anomaly detection emerged as areas where adoption remains limited. While many industries are rapidly pivoting towards these techniques, their application within water utilities is still at an early stage. Given the extensive research demonstrating the effectiveness of machine learning for anomaly detection (Zhou *et al.* 2019; Glynis *et al.* 2023; Sousa *et al.* 2023), water utilities stand to benefit greatly from integrating such advanced methods to enhance decision-making and operational efficiency.

In terms of data, relationship-based querying exposed the potential of certain types of data that can add value as direct or supplementary inputs into intermediary processes, ultimately strengthening business applications. For instance, satellite data was identified as having far greater potential for integration than is currently being realized. Numerous studies have demonstrated the applicability of satellite data, particularly Synthetic Aperture Radar (SAR) techniques, for burst and leak localization in water networks (Arsénio *et al.* 2015; Mazzarotto *et al.* 2023). Similarly, water quality data, while already perceived as useful for several operational tasks, was also identified as having considerable untapped potential, particularly as a supplementary input for enhancing burst and leak detection (Barros *et al.* 2023).

It should be emphasized, however, that a graph-database-driven approach such as this does not automatically highlight gaps and opportunities. Rather, it provides a structured environment in which experts with domain knowledge can interrogate the system, surface opportunities, and make decisions through iterative engagement. This highlights the importance of the Human-in-the-Loop (HITL) concept, which is highly relevant to this exercise. Analysts play a central role in refining queries, interpreting results, and steering the analysis toward meaningful outcomes. As with any expert-informed modelling process, inaccuracies or incomplete relationships may occasionally arise; however, the graph database structure allows these to be readily identified through analysis and corrected through iterative updates without requiring changes to the overall framework.

An additional value of the framework lies in its ability to simulate potential disruptions before they occur. Once a process is identified as potentially susceptible, relationship attribute-based queries can reveal whether most of its remaining inputs are tagged with Low *current_usage* levels. This becomes particularly important when simulating the concurrent failure of multiple data types. If only weak inputs remain, the process may become unreliable and either need to be terminated or risk passing unreliable outputs downstream. By enabling such 'in-advance' simulations, graph databases offer utilities the ability to anticipate vulnerabilities, prepare precautionary measures, and ensure that critical decision-support processes remain intact even under stressed conditions.

An important observation from this study is that while the three-layer framework provides a structured approach, its application is not without assumptions. One key aspect of this approach is the subjectivity involved in defining various attribute parameters, particularly those that influence how one perceives the current levels of usage and identifies potential areas for improved adoption of data types and intermediary processes. This may sometimes require expert judgment; however, the nature of this framework allows for quick and easy updates to these parameters when differences in opinion necessitate parameter overrides.

It is possible to argue that the connections, and indeed potential connections, must already be known; otherwise, they could not have been entered and captured within the framework. An example is the linkage from water quality data to burst-related business applications highlighted in this study. The potential to use water quality data has been demonstrated (Kumar *et al.* 2012; Barros *et al.* 2023) but this understanding may exist in silos. The power of the three-layer framework and graph-database-driven approach lies in its ability to map the complexity of interactions and interconnections, revealing the full extent of potential usage. In particular, by making underutilized connections for water quality data visible, the framework provides additional inputs to burst-related processes, enhancing the robustness and resilience of decision-making even when other key datasets are temporarily unavailable. Without the framework, these connections are not appreciated, remain siloed, and stay hidden.

Another important consideration is the scalability and adaptability of this framework. While this study applied the three-layer framework to Scottish Water DWDS, the same approach could be adapted for sewerage and drainage water infrastructure, or indeed to other utility types with different data infrastructures and business needs. The same structured discussions used in this study could be applied when implementing the framework in other utilities; however, the specific datasets, intermediary processes, and relationships represented in the graph would reflect the particular data structures and operational workflows of the organisation. The framework's flexibility lies in its ability to accommodate new data sources, evolving business applications, and shifting regulatory requirements over time. This adaptability ensures that water utilities can continuously refine their data strategies without requiring a complete overhaul of their existing systems. In addition, while the present framework focuses on cross-layer relationships that trace how data feeds intermediary processes and supports business applications (captured through the FEEDS_INTO and SUPPORTS relationship types), these categories were selected to maintain a clear and interpretable structure for analysing data–process–application pathways. The graph structure could also accommodate additional relationship types or intra-layer relationships (e.g., dependencies between datasets or interactions between analytical processes) if deeper modelling of system interdependencies were required.

While the three-layer framework highlights areas where data and processes are not fully utilized and provides opportunities to simulate the repercussions of losing critical datasets or processes on the connectivity chain, it does not automatically prescribe solutions. Instead, it offers utility companies a structured diagnostic approach to evaluate their data landscape and explore opportunities for better cross-departmental collaboration, improved data governance, and more data-driven decision-making.

Beyond the analytical techniques demonstrated in this study, the graph structure developed here could also support integration with graph-based machine learning approaches such as Graph Neural Networks (GNNs). Once datasets, intermediary processes, and business applications are represented as connected nodes and relationships, GNN models could potentially be used for tasks such as link prediction to identify previously unrecognised relationships between datasets and processes, or to learn structural patterns within the network that highlight datasets whose connectivity differs from expected usage patterns. Such approaches could complement the analytical techniques used in this study and provide additional insights for improving data utilisation strategies in water utilities.

Overall, this study demonstrates that a graph-based approach can significantly improve how water utilities manage and analyse their data. However, its successful implementation depends on continuous engagement with domain experts, periodic updates to reflect new insights, and an openness to adapting the framework based on evolving industry needs. Future research could explore more advanced graph analytics techniques and comparative studies across multiple utilities to further enhance decision-making capabilities in water utility management.

6. CONCLUSIONS

The graph-database-driven three-layer framework proposed and demonstrated here can enable water utilities to understand and manage their extremely complex data structures and flows. It was designed to integrate, structure, and analyse water utility data flows and utilisation within a unified platform. The systematic approach reveals underutilized data sources, inefficiencies in analytical processes, and opportunities for improved integration. It can also identify missed opportunities in how data and processes are utilized for operational processes and decision-making. This was demonstrated for Scottish Water drinking water distribution systems but is applicable to any linear infrastructure sector.

Our findings confirm that traditional siloed data structures hinder the full potential of water utility data, while a graph-based approach enhances visibility, interconnectedness, and analytical depth. By mapping data sources, intermediary processes, and business applications into a single, structured framework, this study provides a practical methodology for improving data-driven decision-making and operational efficiency.

One of the key strengths of the three-layer framework is its flexibility and adaptability, allowing iterative updates based on evolving business needs. Its effectiveness depends on accurate data representation and organizational alignment. The framework does not prescribe rigid solutions but serves as a decision-support tool, providing actionable insights that require expert interpretation. As the industry moves towards digital transformation, leveraging such graph databases will be crucial in creating more efficient, interconnected, and data-driven management systems.

ACKNOWLEDGEMENTS

This research was supported by an Engineering and Physical Sciences Research Council (EPSRC) studentship as part of the Centre for Doctoral Training in Water Infrastructure and Resilience (EPSRC EP/S023666/1), with additional support from Scottish Water as the industrial partner. We would like to express our gratitude to the members of various departments at Scottish Water for providing valuable insights and information about data and technical processes, which greatly enhanced our understanding of the business functions and operations.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Al-Mansouri, A. (2021) Graph databases for fraud detection: a fresh look at financial security, *International Journal of Digital Innovation*, **2** (1), 1–9.

Arsénio, A. M., Dheenathayalan, P., Hanssen, R., Vreeburg, J. & Rietveld, L. (2015) Pipe failure predictions in drinking water systems using satellite observations, *Structure and Infrastructure Engineering*, **11** (8), 1102–1111.

Barros, D., Almeida, I., Zanfei, A., Meirelles, G., Luvizotto, E. & Brentan, B. (2023) An investigation on the effect of leakages on the water quality parameters in distribution networks, *Water*, **15**, 324. https://doi.org/10.3390/w15020324.

Bollen, E., Hendrix, R., Kuijpers, B. & Vaisman, A. (2021) Towards the Internet of Water: using graph databases for hydrological analysis on the Flemish river system, *Transactions in GIS*, **25**, 2907–2938. https://doi.org/10.1111/tgis.12801.

Bollen, E., Hendrix, R. & Kuijpers, B. (2024) Managing data of sensor-equipped transportation networks using graph databases, *Geoscientific Instrumentation, Methods and Data Systems*, **13**, 353–371. https://doi.org/10.5194/gi-13-353-2024.

Carriço, N., Ferreira, B., Barreira, R., Antunes, A., Grueau, C., Mendes, A., Covas, D., Monteiro, L., Santos, J. & Brito, I. S. (2020) Data integration for infrastructure asset management in small to medium-sized water utilities, *Water Science and Technology*, **82**, 2737–2744. https://doi.org/10.2166/wst.2020.377.

Cheng, J., Ke, Y. & Ng, W. (2009) Efficient query processing on graph databases, *ACM Trans. Database Syst.*, **34**, 2:1–2:48. https://doi.org/10.1145/1508857.1508859.

Chicho, B. & Mohammed, A. O. (2023) An empirical comparison of Neo_4J and tigergraph databases for network centrality, *Science Journal of University of Zakho*, **11**, 190–201. https://doi.org/10.25271/sjuoz.2023.11.2.1068.

Dell, R. K. (2005) Breaking organizational silos: removing barriers to exceptional performance, *Journal AWWA*, **97**, 34–36. https://doi.org/10.1002/j.1551-8833.2005.tb10902.x.

Glynis, K., Kapelan, Z., Bakker, M. & Taormina, R. (2023) Leveraging transfer learning in LSTM neural networks for data-efficient burst detection in water distribution systems, *Water Resour Manage*, **37**, 5953–5972. https://doi.org/10.1007/s11269-023-03637-3.

Grueau, C., Antunes, A., Ferreira, B., Gonçalves, M., Gomes, J. & Carriço, N. (2019). 'Towards an Integrated Platform for Decision Support in Water Utility Management', *Presented at the IADIS International Conference Information Systems 2019*, pp. 235–240.

Halfawy, M. R. (2008) Integration of municipal infrastructure asset management processes: challenges and solutions, *Journal of Computing in Civil Engineering*, **22**, 216–229. https://doi.org/10.1061/(ASCE)0887-3801(2008)22:3(216).

Henderson, R. (2020) Using graph databases to detect financial fraud, *Computer Fraud & Security*, **2020**, 6–10. https://doi.org/10.1016/S1361-3723(20)30073-7.

Javaherian Pour, E., Atazadeh, B., Rajabifard, A., Sabri, S. & Norris, D. (2025) A graph data model for CityGML utility network ADE: a case study on water utilities, *ISPRS International Journal of Geo-Information*, **14**, 493. https://doi.org/10.3390/ijgi14120493.

Javier Zorzano Mier, F. & Iglesias, C. Á. (2023) Applications of knowledge graphs in telecommunication systems management, *IEEE Internet Computing*, **27**, 29–34. https://doi.org/10.1109/MIC.2023.3253305.

Kumar, P. & Huang, H. H. (2020) Graphone: a data store for real-time analytics on evolving graphs, *ACM Trans. Storage*, **15**, 29. https://doi.org/10.1145/3364180.

Kumar, J., Sreepathi, S., Brill, E. D., Ranjithan, R. & Mahinthakumar, G. (2012) Detection of Leaks in Water Distribution System Using Routine Water Quality Measurements. In: Richard N., Palmer (ed.) *World Environmental and Water Resources Congress 2010: Challenges of Change*, Providence, Rhode Island, USA:American Society of Civil Engineers (ASCE), pp. 4185–4192. https://doi.org/10.1061/41114(371)426.

Kyritsakas, G., Boxall, J. B. & Speight, V. L. (2023) A Big data framework for actionable information to manage drinking water quality, *AQUA – Water Infrastructure, Ecosystems and Society*, **72**, 701–720. https://doi.org/10.2166/aqua.2023.218.

Lawand, S., Hegde, G., Lawand, D., Parveen, N. & Rohatgi, D. (2024) A proposal for route optimization in supply chain management using graph database (GDMS), *AIP Conference Proceedings*, **3214**, 020040. https://doi.org/10.1063/5.0239095.

Lehotay-Kéry, P. & Kiss, A. (2020) Process, analyze and visualize telecommunication network configuration data in graph database, *Vietnam J. Comp. Sci.*, **07**, 65–76. https://doi.org/10.1142/S2196888820500037.

Mazzarotto, G., Tessari, G., Pizzaia, P. & Salandin, P. (2023) Identifying pipeline leak positions potentially connected to soil deformations through SAR data analysis, *Journal of Infrastructure Systems*, **29**, 04023017. https://doi.org/10.1061/JITSE4.ISENG-2189.

Medhi, S. & Baruah, H. K. (2017) Relational database and graph database: a comparative analysis, *Journal of Process Management and New Technologies*, **5**, 1–9. https://doi.org/10.5937/jouproman5-13553.

Park, Y., Shankar, M., Park, B.-H. & Ghosh, J. (2014). 'Graph databases for large-scale healthcare systems: a framework for efficient data management and data services', *2014 IEEE 30th International Conference on Data Engineering Workshops. Presented at the 2014 IEEE 30th International Conference on Data Engineering Workshops*, pp. 12–19. https://doi.org/10.1109/ICDEW.2014.6818295.

Platfoot, R. A., (2020) Predictive Analytics Combining Multi-stream Data Sources. In: Liyanage, J. P., Amadi-Echendu, J. & Mathew, J. (eds.) *Engineering Assets and Public Infrastructures in the Age of Digitalization*. Cham: Springer International Publishing, pp. 697–703. https://doi.org/10.1007/978-3-030-48021-9_77.

Ponce Romero, J. M., Hallett, S. H. & Jude, S. (2017) Leveraging Big data tools and technologies: addressing the challenges of the water quality sector, *Sustainability*, **9**, 2160. https://doi.org/10.3390/su9122160.

Qiao, J., Shen, K., Xiao, W., Tang, J., Chen, Y. & Xu, J. (2024) Integrating graph data models in advanced water resource management: a new paradigm for complex hydraulic systems, *Water*, **17**, 3. https://doi.org/10.3390/w17010003.

Rauch, E., Bataleblu, A. A., Golser, M., Emer, A., Matt, D. T., (2024) Potential of Graph Database Visualization of the Supplier Network to Increase Resilience in Multi-tier Supply Chains. In: Trojanowska, J., Kujawińska, A., Pavlenko, I. & Husar, J. (eds.) *Advances in Manufacturing IV*, Cham Springer Nature Switzerland, pp. 125–139. https://doi.org/10.1007/978-3-031-56444-4_10.

Robinson, I., Webber, J. & Eifrem, E. (2015) *Graph Databases*, 2nd edn. Sebastopol, CA, USA: O'Reilly Media, Inc. Available at: https://www.oreilly.com/library/view/graph-databases-2nd/9781491930885/ (accessed 3.16.25).

Rose, A. (2015). 'On the uses of graph databases in water distribution systems', *Presented at the International Conference on Water Management Modeling*.

Rowley, J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy, *Journal of Information Science*, **33**, 163–180. https://doi.org/10.1177/0165551506070706.

Scottish Water. (2024) 2023/ 2024 Annual Report and Accounts Performance and Prospects. Scottish Water.

Sousa, D. P., Du, R., Mairton Barros da Silva Jr, J., , Cavalcante, C. C. & Fischione, C. (2023) Leakage detection in water distribution networks using machine-learning strategies, *Water Supply*, **23**, 1115–1126. https://doi.org/10.2166/ws.2023.054.

Turhan, S. N. (2023). 'Leveraging Graph Databases for Enhanced Healthcare Data Management: a Performance Comparison Study', *2023 IEEE International Conference on Big Data (BigData). Presented at the 2023 IEEE International Conference on Big Data (BigData)*, pp. 5007–5013. https://doi.org/10.1109/BigData59044.2023.10386297.

Vaquero, L. M., Cuadrado, F. & Ripeanu, M. (2014) Systems for Near Real-Time Analysis of Large-Scale Dynamic Graphs. https://doi.org/10.48550/arXiv.1410.1903.

Wallis-Lage, C. (2020) Report reflects critical role of data and planning in water infrastructure resilience, *Journal AWWA*, **112**, 38–46. https://doi.org/10.1002/awwa.1462.

Weintraut, Z., Newhart, K., Thompson, K. & Roostaei, J. (2022) Are you ready for big data?, *Journal AWWA*, **114**, 78–82. https://doi.org/10.1002/awwa.2021.

Zhang, P., Wang, T. & Yan, J. (2022) Pagerank centrality and algorithms for weighted, directed networks, *Physica A: Statistical Mechanics and its Applications*, **586**, 126438. https://doi.org/10.1016/j.physa.2021.126438.

Zhou, X., Tang, Z., Xu, W., Meng, F., Chu, X., Xin, K. & Fu, G. (2019) Deep learning identifies accurate burst locations in water distribution networks, *Water Research*, **166**, 115058. https://doi.org/10.1016/j.watres.2019.115058.

First received 12 October 2025; accepted in revised form 26 May 2026. Available online 18 June 2026