

BMAN73250

Data Architecture for a Conservation “Lakehouse”

A dissertation submitted to The University of Manchester for the degree of Master of Science in the
Faculty of Humanities

Alliance Manchester Business School

MSc Business Analytics: Operational Research and Risk Analysis

2026

Student ID:

Table of Contents

List of Figures.....	4
List of Tables	5
List of Abbreviations	6
Abstract.....	7
Declaration.....	8
Intellectual Property Statement.....	8
1. Introduction.....	9
1.1 Background	9
1.2 Problem Statement and Research Aim.....	9
1.3 Research Questions	9
1.4 Research Objectives	10
1.5 Scope of the Study	10
2. Literature Review	10
2.1 Evolution of Data Architectures: From Data Warehouses and Data Lakes to Lakehouse	10
2.2 Lakehouse Implementation: Medallion Architecture and ETL/ELT Pipelines	11
2.3 IoT and Environmental Sensor Data Integration	12
2.4 Spatial Data Infrastructure for Geospatial Lakehouses.....	13
2.5 Metadata Management and Data Governance	13
2.6 Research Gap in Conservation Data Integration.....	14
3. Research Methodology	16
3.1 Research Approach: Design Science Research.....	16
3.2 Research Process and Requirements Analysis.....	18
3.3 Design Framework and Principles	20
3.4 Evaluation Criteria	20
4. Conservation Lakehouse Architecture.....	22
4.1 Current State Assessment and Requirements.....	22
4.2 Data Sources Overview.....	23
4.3 Lakehouse Data Pipeline Design	25
4.4 Environmental IoT Data Pipeline Example	27
4.5 Metadata, Governance and Security Framework.....	28
4.6 Sample Data Catalogue	30
4.7 Output and Consumption: Prototype Dashboard and User Views.....	31
4.8 Final Lakehouse Architecture Blueprint	33
5. Evaluation and Discussion	35
5.1 Evaluation of the Proposed Architecture	35

5.2 Benefits and Limitations	36
5.3 Recommendations and Future Roadmap	38
6. Conclusion.....	39
6.1 Summary of Findings and Contributions	39
6.2 Practical Implications and Future Research.....	39
References	40
Appendices.....	43
Appendix A. Full sample data catalogue.	44
Appendix B. Suggested refresh frequencies by data domain.....	47
Appendix C. Prototype repository structure.	48
Word Counts: 8,739	

List of Figures

Figure 1. Synthesis of the literature review and the identified research gap.	15
Figure 2. Design Science Research Framework, adapted from Hevner et al. (2004).....	17
Figure 3. DSRM process model adapted for this research.....	19
Figure 4. Existing architecture provided by The Wildlife Trusts.	22
Figure 5. Data-centric taxonomy of digital reserve datasets.....	23
Figure 6. Prototype lakehouse data pipeline.	25
Figure 7. Environmental IoT data across the Databricks Medallion layers.....	27
Figure 8. Example environmental summary returned by the API.	28
Figure 9. Digital Reserve ELT pipeline in Databricks.....	29
Figure 10. Section-level data lineage for the Habitat & Environment view.....	29
Figure 11. Overview dashboard of the Digital Reserve prototype.	32
Figure 12. Final Lakehouse architecture blueprint of the Digital Reserve prototype.....	34
Figure 13. Layered benefit model derived from the prototype evaluation.	37

List of Tables

Table 1. Application of Hevner’s design science research guidelines in this study.	17
Table 2. Requirements analysis for the proposed conservation lakehouse.....	18
Table 3. Mapping of DSRM activities to the conservation lakehouse study.....	19
Table 4. Evaluation matrix.....	21
Table 5. Summary of digital reserve data characteristics.	24
Table 6. Sample data catalogue of representative dashboard views in prototype.....	31
Table 7. Evaluation summary of the prototype.....	35
Table 8. Page-level summary of linked Gold outputs.....	36
Table 9. Roadmap from prototype implementation to future production.	38

List of Abbreviations

Abbreviation	Full term
ACID	Atomicity, Consistency, Isolation and Durability
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
BI	Business Intelligence
CNN	Convolutional Neural Network
CRM	Customer Relationship Management
CSV	Comma-Separated Values
DSR	Design Science Research
DSRM	Design Science Research Methodology
EBV	Essential Biodiversity Variable
ELT	Extract, Load, Transform
ETL	Extract, Transform, Load
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Regulation
GIS	Geographic Information System
IoT	Internet of Things
LiDAR	Light Detection and Ranging
LoRaWAN	Long Range Wide Area Network
REST API	Representational State Transfer API
RSSI	Received Signal Strength Indicator
SDI	Spatial Data Infrastructure
SNR	Signal-to-Noise Ratio
SQL	Structured Query Language
UC	Unity Catalog

Abstract

This study investigates how a conservation lakehouse architecture can support integrating diverse, heterogeneous conservation data. As environmental sensors, soundscapes, biodiversity, and geographical location data increase, existing data sources are becoming more fragmented, inconsistent, and difficult to manage and use at the front end. To address these issues, a Design Science Research approach is used to design and implement a conservation-focused lakehouse model using Databricks. The model adopts Medallion Architecture to support data processing, standardisation, and analysis. Databricks notebooks and pipelines process the heterogeneous data, converting it into analysis-ready data for dashboards, while PostgreSQL and REST APIs support front-end displays. The main outputs include an architecture blueprint, a sample data catalogue, Gold-layer data products with data quality and lineage information, and a role-oriented prototype dashboard. The evaluation results indicate that the prototype can support heterogeneous data integration, repeatable data processing, and the delivery of Gold-layer outputs to multiple dashboard views. However, the prototype remains constrained by the scope, sample data, platform permissions, and the incomplete deployment of governance mechanisms. Overall, the prototype demonstrates the feasibility of the lakehouse architecture as a future data platform foundation for the Digital Reserve and provides a reference point for future deployment and governance enhancements.

Declaration

I declare that this dissertation is my own original work, except where referenced clearly to the contrary. No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Intellectual Property Statement

i. The author of this dissertation (including any appendices and/or schedules to this dissertation) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of works in the dissertation, for example graphs and tables (“Reproductions”), which may be described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy, in any relevant Dissertation restriction declarations deposited in the University Library, and The University Library’s regulations.

1. Introduction

1.1 Background

The Wildlife Trusts¹ is a conservation organisation representing an association of 46 independent trusts, which collectively oversee the management of over 2,600 protected areas (The Wildlife Trusts, n.d.). The organisation has been involved in species conservation, habitat protection and public engagement for a considerable period of time. In recent years, as digital conservation has become an increasingly important part of conservation practice, The Wildlife Trusts has also been promoting the development of Digital Reserves. The objective of this study is to utilise digital solutions to enhance the integration, analysis and application of conservation data. In this context, conservation work is no longer solely dependent on traditional surveys and records, but is gradually combining sensors, environmental monitoring equipment, communication networks, and other digital data sources (Stephenson and Stengel, 2020). This has resulted in a wider variety of data in conservation areas.

1.2 Problem Statement and Research Aim

The main challenge of this project is to design a future data platform built on an existing data lake architecture to support new types of data generated by sensors and communication networks. This dataset covers areas such as the Internet of Things (IoT), biodiversity, environmental monitoring, risk, human activity, and sensor connectivity. This may depend on the format, update frequency, application setting, and analysis requirements. The platform should preserve information on data sources, processing, and data quality for users to know where and how the data was generated. Without a consistent and scalable architecture, it would be difficult to integrate these heterogeneous data effectively and transform them into outputs that can be used for decision-making. This research aims to design and prototype a lakehouse data architecture to support the integration of heterogeneous conservation data, while exploring its feasibility as the foundation for the Digital Reserve's future data platform (Armbrust et al., 2021).

1.3 Research Questions

This study concentrates on three research questions relating to data integration challenges, architectural design, governance and security design. Accordingly, the following research questions are posed in this paper:

1. What data integration challenges does The Wildlife Trusts face when incorporating sensor and environmental data into its existing data architecture?

¹ The Wildlife Trusts official website, available at: <https://www.wildlifetrusts.org/>

2. How can The Wildlife Trusts' existing lakehouse architecture be extended using a medallion framework to support the scalable integration of current data?
3. How should metadata management, data governance, and security controls be designed to support scalable conservation analytics within the extended lakehouse architecture?

1.4 Research Objectives

The objectives of this study are as follows:

1. **Requirements and challenge analysis:**

Analyse the main requirements and challenges that The Wildlife Trusts face when integrating sensor, environmental and other conservation data.

2. **Architectural design:**

Design a scalable data architecture prototype utilising the lakehouse and Medallion frameworks, and incorporating Databricks.

3. **Data pipeline and governance design:**

Design ETL and ELT processes to support geospatial, survey, sensor, administrative and other relevant data. Integrate metadata management, data governance and security considerations into the overall design.

4. **Blueprint and catalogue development:**

Create an architectural blueprint and a sample data catalogue to show how the proposed architecture will support subsequent analysis and dashboard applications.

1.5 Scope of the Study

This research focuses on designing a lakehouse-based prototype conservation data architecture to support the integration, processing, and subsequent utilisation of The Wildlife Trusts' diverse and heterogeneous data. It covers the data architecture design, the Medallion workflow, and metadata and governance considerations. It also demonstrates a sample data catalogue and prototype dashboard outputs. However, it does not cover full-scale, enterprise-level deployment or advanced governance features. Therefore, this study takes a design- and prototype-driven approach to validate the technical feasibility of the conservation lakehouse architecture.

2. Literature Review

2.1 Evolution of Data Architectures: From Data Warehouses and Data Lakes to Lakehouse

Data storage architectures have evolved from data warehouses and data lakes to data lakehouses as data volumes have grown rapidly (Janssen et al., 2024). Data warehouses offer robust governance,

querying and ACID transaction capabilities; however, their ‘schema-on-write’ design is primarily suited to structured data, and they offer lower flexibility and cost-effectiveness when processing large volumes of unstructured data such as images and audio. While a data lake can store data in a variety of formats at a reduced cost, inadequate governance and transactional mechanisms can compromise data quality and even result in a 'data swamp'.

Armbrust et al. (2021) point out that a data lakehouse combines the governance capabilities of a data warehouse with the storage flexibility of a data lake, utilising open formats such as Parquet to support data engineering, SQL analytics, business intelligence, and machine learning on a single platform. Mazumdar et al. (2023) further explain that data lake table technologies such as Delta Lake, Apache Iceberg, and Apache Hudi can provide ACID transactions, version control, schema management, and metadata management on top of cloud storage. Coupled with compatible engines, they support SQL, streaming, and machine learning workloads, thereby reducing data replication across systems. However, Schneider et al. (2024) note that no consensus exists on the definition, technical requirements, and level of tool support for lakehouses. Their suitability should be assessed individually in each case.

2.2 Lakehouse Implementation: Medallion Architecture and ETL/ELT Pipelines

The Medallion architecture typically organises data into three tiers: Bronze, Silver and Gold². Bronze retains raw data, Silver performs data cleaning and standardisation, and Gold provides analysed data (Databricks, n.d.). To assist in maintaining consistency and version control across the entire hierarchical process, Delta Lake can be used as the underlying storage mechanism. Armbrust et al. (2020) describe Delta Lake as an open-source ACID table storage layer that provides snapshot isolation, time travel, and the ability to modify existing data via transaction logs and Parquet checkpoints. This mechanism maintains consistency across a single data table during concurrent reads and writes while supporting queries on historical versions, thereby enhancing the reliability and traceability of data pipelines.

Jain et al. (2023) compared Delta Lake, Apache Hudi, and Apache Iceberg, noting differences in how they manage transactions, metadata and data updates. Delta Lake's transaction log, checkpointing features, and batch/stream processing integration enable continuous sensor data writes while maintaining data consistency; it can therefore be considered a candidate solution for the Medallion lakehouse storage layer. However, Camacho-Rodríguez et al. (2024) note that actual performance

² See Databricks (n.d.) for the Medallion architecture as a layered lakehouse design pattern: <https://docs.databricks.com/aws/en/lakehouse/medallion>

still depends on the table format, processing engine, update patterns and optimisation strategies; consequently, no single format is optimal for all scenarios.

Research by Schneider et al. (2025) and Abouzaid et al. (2025) indicates that the lakehouse already has the technical foundation for industrial deployment, stream processing and cross-environment implementation; however, data governance, platform configuration and performance still require further evaluation. Existing research has largely focused on manufacturing, automotive telemetry, and enterprise analytics. This leaves a gap in end-to-end Medallion lakehouse implementations and evaluations concentrating on multimodal conservation IoT data.

2.3 IoT and Environmental Sensor Data Integration

Al-Fuqaha et al. (2015) defined the Internet of Things (IoT) as a technological paradigm that provides physical objects with sensing, communication, and computing capabilities. They proposed six core elements: identification, sensing, communication, services, computing, and semantics. In wildlife monitoring contexts, an opportunistic dual-radio architecture (Ayele et al., 2018) can address issues such as the dispersed nature of sensing nodes and intermittent network connectivity. Given these characteristics, this study suggests that The Wildlife Trusts should adopt an ‘offline-first’ and ‘store-and-forward’ data acquisition model, whereby raw data is first stored temporarily on the device and then written to the Bronze Layer once connectivity is restored.

The multimodal wildlife monitoring dataset proposed by Kline et al. (2025), which integrates drone imagery, camera-trap images, and bioacoustic recordings, demonstrates the high heterogeneity of modern monitoring data. Accordingly, this study stores raw data and its source information in the Bronze Layer, whilst performing missing-value handling, outlier detection, format standardisation, and data cleaning in the Silver Layer. From the Silver to Gold stages, BirdNET can be applied to species identification and feature extraction from AudioMoth recordings (Kahl et al., 2021), whilst research by Tabak et al. (2019) demonstrates that CNNs can support animal detection and species classification in camera trap images.

In addition to centralised storage, the integration of environmental and biodiversity data must also preserve the data context, sources and processing history (Vargas-Solar et al., 2024). Consequently, The Wildlife Trusts’ data platform should possess capabilities for multimodal storage, data governance, source tracing and processing history management. A review of existing research indicates that whilst the relevant literature has explored the Internet of Things (IoT), wildlife monitoring, multimodal data management and scientific data lakes separately, there is currently a scarcity of research integrating these aspects into the ‘Medallion Lakehouse’ architecture for conservation sensor data.

2.4 Spatial Data Infrastructure for Geospatial Lakehouses

Spatial Data Infrastructure (SDI) is a framework that combines institutional, organisational and technical elements, designed to assist organisations in discovering, accessing, exchanging and reusing spatial data. Rajabifard and Williamson (2001) indicated that the effectiveness of SDI depends not only on data, standards, metadata and technical services but is also influenced by governance, organisational collaboration and user awareness. Relevant research indicates that SDI can integrate dispersed biodiversity and environmental data and support spatial planning, environmental assessment and land management; however, it may still be constrained by data completeness, interoperability and system availability (Underwood et al., 2018; Zwirowicz-Rutkowska & Michalik, 2016).

Geospatial data comprises geometric types, coordinate reference systems and spatial relationships, and therefore requires specialised storage and query methods. Saeedan and Eldawy (2022) proposed Spatial Parquet, which enhances data retrieval efficiency through geometric encoding, compression and spatial indexing; Zhang et al. (2023) integrated Spatial SQL, spatial partitioning and data skipping into a lakehouse to improve the performance of large-scale spatial queries. Remote sensing data can also be acquired via satellites, aircraft or drones and is used to analyse land cover, habitat change and species distribution.

However, existing research largely focuses on a single technology or specific data types. Within the scope of this study's literature review, no research has been identified that, in the context of local conservation organisations in the UK, integrates camera traps, acoustic monitoring, environmental sensors and remote sensing data into a single lakehouse for end-to-end processing and evaluation. Heterogeneous data also faces challenges such as inconsistencies in spatio-temporal scales, classification standards, metadata and data quality.

2.5 Metadata Management and Data Governance

Wilkinson et al. (2016) proposed the FAIR principles, emphasising that data should include comprehensive metadata, standardised access methods, a common format, and clear licensing and source information to enhance data discoverability, accessibility, interoperability, and reusability. However, FAIR does not imply that all data must be made fully public; rather, it aims to enhance the value of data management and reuse whilst respecting legal, ethical and privacy considerations. As sensor data may contain location, behavioural, or device-identification information, it may also be subject to the GDPR. Even if data have been pseudonymised, appropriate access restrictions and security measures should be implemented as long as there remains a risk of re-identification (Mourby et al., 2019).

In a data lake, metadata describes the data's source, structure, format, quality, processing history, and usage, enabling different types of data to be searched, understood, and utilised correctly. Boukraa et al. (2024) point out that metadata management should encompass modelling, automatic extraction, storage, updating, and maintenance, and should integrate data catalogues, classification tags, search, and data lineage. In addition to supporting data exploration and integration, metadata can also record data responsibilities, access conditions and quality status, serving as the basis for authorisation decisions, quality checks, audit trails and lifecycle management.

A lakehouse, on the other hand, can incorporate metadata, access control, data lineage, auditing, quality and asset lifecycle management under a unified governance platform. For example, Unity Catalog supports a unified namespace and multi-tiered access control; Microsoft Purview focuses on data catalogues, sensitive data classification, and policy management, whilst the Unified Lineage System emphasises large-scale data transformation and tracking upstream and downstream dependencies. These examples demonstrate that modern catalogues have evolved from simple metadata repositories into a governance control layer that integrates permissions, lineage, auditing, data quality and policy enforcement (Ahmad et al., 2023; Chandra et al., 2025; Jacques-Silva et al., 2025).

However, existing research largely focuses on ethical principles, data standards or individual governance functions, and enterprise-level platforms may not fully address the needs of conservation research regarding sensitive species-location data, limited resources and the requirements for cross-organisational data sharing.

2.6 Research Gap in Conservation Data Integration

Pereira et al. (2013) introduced Essential Biodiversity Variables (EBVs) as a common conceptual framework for biodiversity monitoring, covering genetic composition, species populations, species traits, community composition, ecosystem structure and ecosystem function. EBVs are not a specific data format, but rather enable information generated by different monitoring programmes to be organised, compared and integrated more consistently. For this study, the EBVs can serve as a design reference for the 'Gold Layer' biodiversity data product, supporting the organisation of soundscape identification, camera trap and field survey results according to common taxonomic, spatial and temporal dimensions; environmental observations such as temperature, rainfall and water levels can serve as contextual information for interpreting changes in species distribution and activity.

Stephenson and Stengel (2020) emphasise that conservation information is scattered across different databases, organisations and monitoring programmes, which increases difficulties in searching for, accessing, ensuring compatibility and reusing the data. Binley et al. (2024) also emphasise that

existing conservation data is not yet being fully utilised. Such issues are also evident in The Wildlife Trusts' data management context, where soundscapes, imagery, species observations, environmental sensor data, and geospatial information are managed across different formats and approaches. The primary challenge is not a lack of data, but rather the absence of a consistent processing framework to support raw data preservation, standardisation, metadata management, quality checks and subsequent analysis.

Taken collectively, IoT, geospatial technologies, and data governance provide methods for heterogeneous information storage, sensor transmission, spatial analysis, and source tracing; however, few studies have integrated these capabilities into a prototype-validated conservation data solution. Consequently, The Wildlife Trusts' data requirements indicate that the lakehouse approach, centred on the Medallion Architecture, is suitable. Data can be systematically structured through a hierarchical workflow from heterogeneous conservation datasets to verified and analysed results. Any future conservation data management system should consider more than just data storage and processing. It should also consider metadata and processes, along with governance structures, to increase traceability, reusability, and scalability. Figure 1 summarises how the reviewed literature leads to the research gap this study addresses.

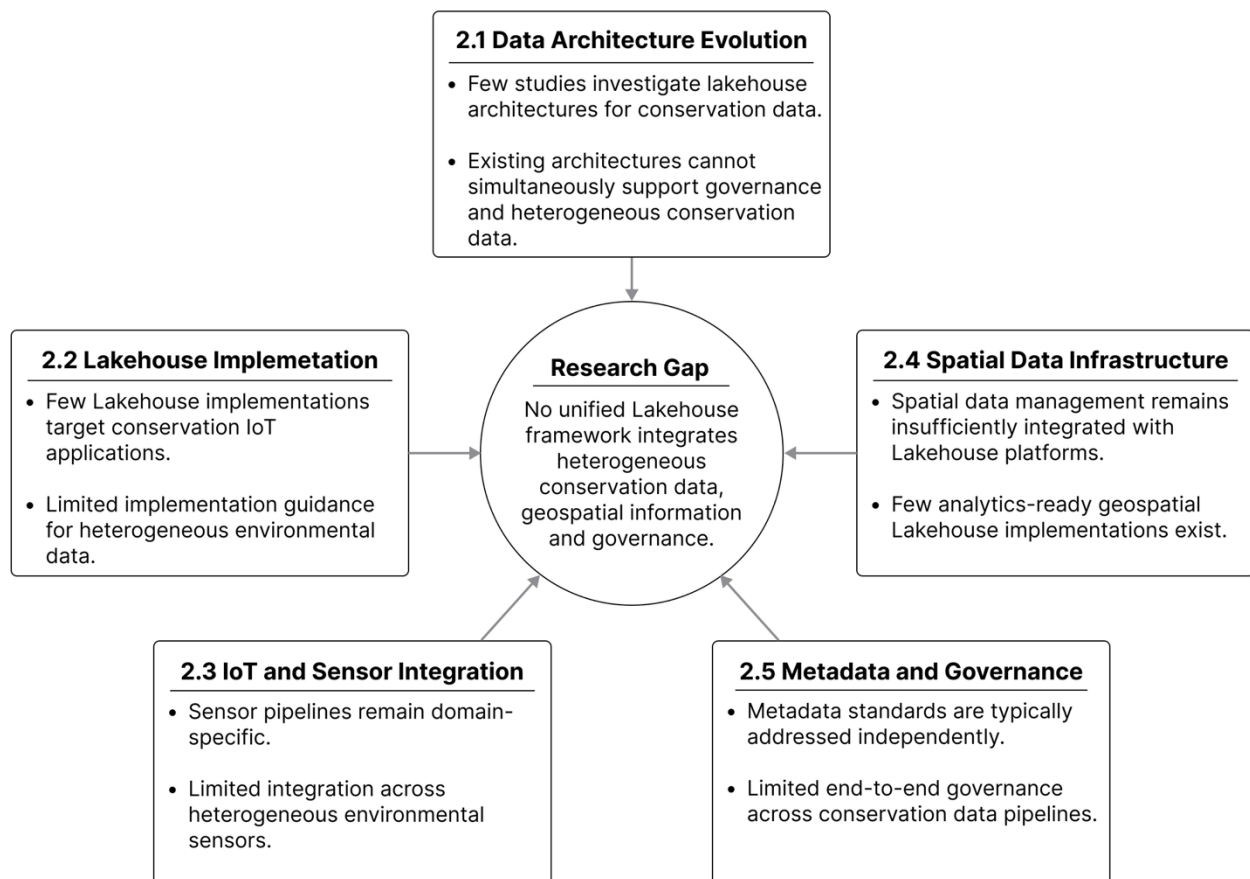


Figure 1. Synthesis of the literature review and the identified research gap.

3. Research Methodology

3.1 Research Approach: Design Science Research

The heterogeneous data gathered by The Wildlife Trusts consists of imagery, audio, time-series, meteorological, IoT, and spatial data. Therefore, developing a scalable data lake is necessary to improve data integration and governance for conservation decisions. This research focuses on designing and validating a sensor data integration solution within an existing technical environment, rather than simply describing the current situation. Since Design Science Research (DSR) centres on constructing and evaluating artefacts, it is more appropriate for this investigation than case studies or questionnaires.

The artefacts generated in this research include a conservation data lakehouse architecture framework and a Databricks functional prototype. The former delineates mechanisms for data ingestion, storage, processing, governance and analysis; the latter is used to validate the feasibility of the architecture, the ETL processes and the analytical outcomes. The present study adheres to the DSR paradigm proposed by Hevner et al. (2004) (see Table 1), starting from the practical environment and requirements of The Wildlife Trusts, building a knowledge base through relevant literature, and connecting practical needs with research knowledge through the construction and evaluation of artefacts.

In carrying out the project, this study followed the six-stage Design Science Research Methodology (DSRM) proposed by Peffers et al. (2007), comprising the identification and motivation stage, the definition of solution objectives, the design and development stage, the demonstration stage, the evaluation stage and the communication stage. Figure 2 shows the design science research framework used in this study.

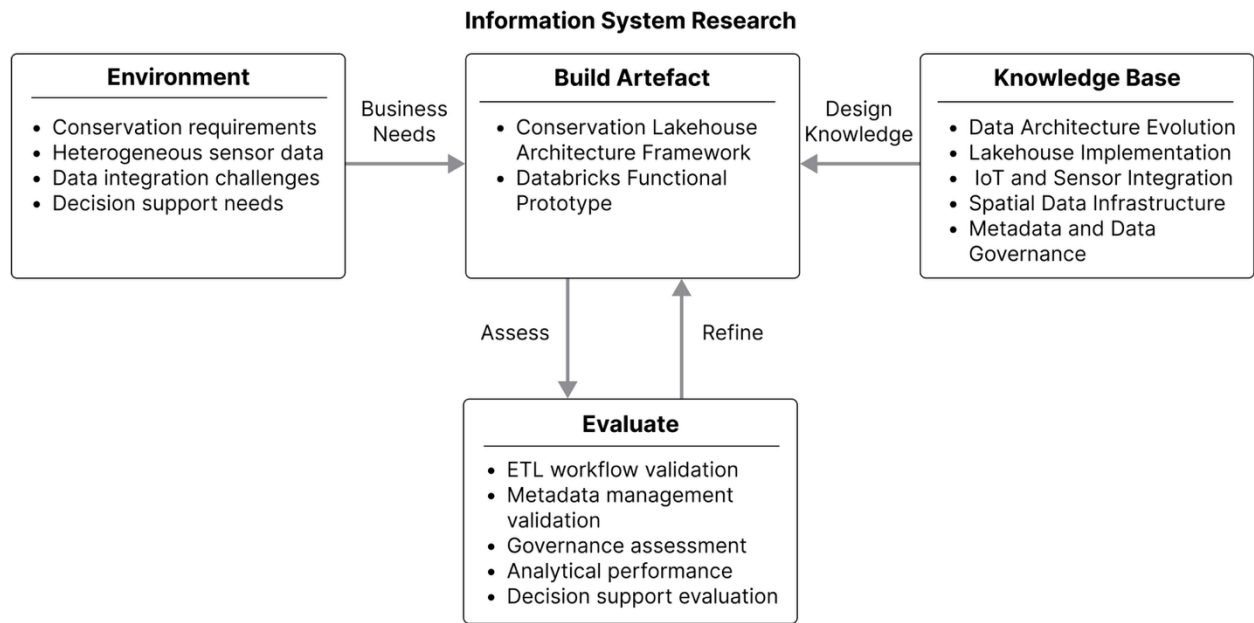


Figure 2. Design Science Research Framework, adapted from Hevner et al. (2004).

Table 1. Application of Hevner’s design science research guidelines in this study.

DSR Guideline	Manifestation of this study
1. Design as an Artifact	Construction of a conservation lakehouse architecture framework and implementation of a Databricks functional prototype as the artefact of the research.
2. Problem Relevance	Addressing the practical needs of the Wildlife Trusts regarding heterogeneous conservation data integration, data governance, scalability, and monitoring.
3. Design Evaluation	Validate the prototype's feasibility and effectiveness through ETL processes, Unity Catalog, Metadata, governance mechanisms, analytical results, and Pipeline testing.
4. Research Contributions	Propose a lakehouse architecture, data governance methodology, and system implementation recommendations for conservation data, serving as a reference for future conservation information platforms.
5. Research Rigor	Use relevant literature such as lakehouse, Medallion Architecture, FAIR Principles, and Metadata Governance as the basis for architecture design and evaluation.
6. Design as a Search Process	Compare different data ingestion, storage, processing, and governance solutions, and iteratively refine the architecture and prototype.
7. Communication of Research	Present the research findings through a paper, an architecture diagram, a dashboard, a metadata catalogue, and a Databricks prototype.

3.2 Research Process and Requirements Analysis

Conservation organisations require the integration of diverse data sources, including sensors, geospatial data, telemetry, imagery, and audio; however, significant differences in their formats, scale, update frequencies, and governance requirements increase the complexity of data acquisition, transformation, management, and analysis. Existing research has explored Lakehouse, IoT, geospatial technologies and data governance separately; however, there has been little effort to integrate these into a comprehensive framework suitable for conservation contexts and to validate its feasibility through functional prototypes. In particular, there remains a lack of concrete implementation examples regarding the hierarchical integration of heterogeneous data, metadata management and governance.

The scope of this research stems from the project brief, datasets, project meetings and existing platforms provided by The Wildlife Trusts. The client requested the design of a scalable data architecture based on the existing framework to support the acquisition and analysis of field sensor data and open data, incorporating ETL/ELT, metadata, data governance, data landing, automated extraction, data storage, cold data management, cost optimisation and information security considerations, whilst also producing an architecture blueprint and a sample data catalogue. The above requirements were further organised into functional and non-functional requirements (see Table 2), serving as the basis for architectural design and prototype development.

Table 2. Requirements analysis for the proposed conservation lakehouse.

Category	ID	Requirements
Functional	FR1	Ingest heterogeneous conservation datasets
	FR2	Support automated data landing and ingestion
	FR3	Implement ETL/ELT pipelines
	FR4	Maintain metadata and sample data catalogue
	FR5	Support analytics and AI change detection
Non-functional	NFR1	Scalability for large volumes of live data
	NFR2	Data governance
	NFR3	Cost optimisation
	NFR4	Cybersecurity
	NFR5	Supports multi-network connectivity (LoRaWAN, satellite, mobile)

This study followed the six-stage DSRM process outlined by Peffers et al. (2007), proceeding sequentially through problem identification, objective definition, design of the conservation lakehouse framework, prototyping using Databricks, sample data demonstration and evaluation of results. Figure 3 and Table 3 map the DSRM process to this research.

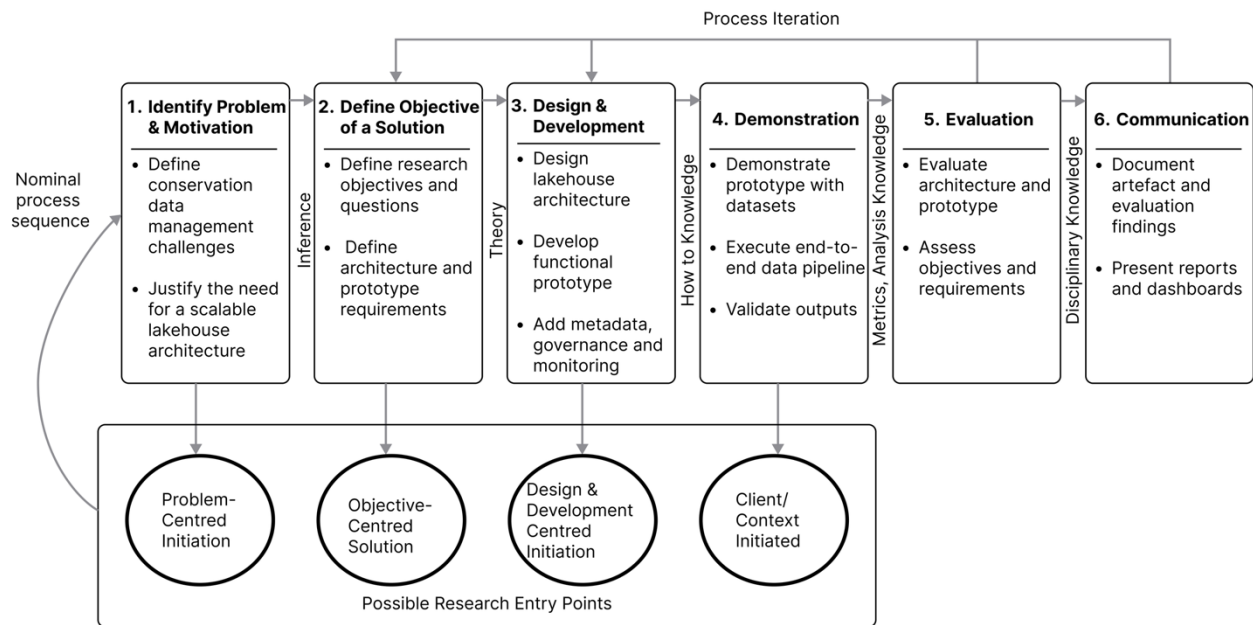


Figure 3. DSRM process model adapted for this research.

Table 3. Mapping of DSRM activities to the conservation lakehouse study.

DSRM activity	Application in this study
Problem identification and motivation	Conservation data integration, scalability and governance challenges
Define objectives of a solution	Requirements for an extensible medallion lakehouse
Design and development	Databricks, Delta Lake, PostgreSQL, API and dashboard artefact
Demonstration	Reserve Manager and Data Lead use cases
Evaluation	Delta validation, Gold inventory, build tests and requirements assessment
Communication	Dissertation, architecture blueprint and prototype demonstration

3.3 Design Framework and Principles

The design and development phases of the DSRM framework proposed by Peffers et al. (2007) first define the required functions and overall architecture of the conservation data lake, and then build a prototype system for demonstration and evaluation. To translate research objectives, stakeholder requirements and relevant literature into concrete architectural decisions, this study proposes five design principles:

1. Heterogeneous integration: Appropriate extraction and storage methods are adopted based on the characteristics of data sources such as sensors, CSV files, APIs, GIS, imagery, and audio, whilst retaining source and extraction metadata.
2. Medallion architecture: The adoption of a ‘Bronze, Silver and Gold’ tiered architecture, whereby raw data is preserved, data cleansing and quality checks are performed, and analytical data products are created.
3. Metadata and governance: The incorporation of source, schema, quality, access rights, audit trails and data lineage into end-to-end governance.
4. Scalability and extensibility: Separate the storage, transformation and service layers, and utilise Delta Lake to support data growth and schema evolution.
5. Binary–metadata separation: Store large image, audio, and LiDAR files in object storage or volumes; data tables should retain only URIs, timestamps, locations, and processing statuses, whilst maintaining data relationships and traceability through unique identifiers.

The above design principles collectively form the basis for prototyping and subsequent evaluation of the conservation data lake.

3.4 Evaluation Criteria

Based on the Demonstration and Evaluation phases proposed by Peffers et al., this study adopted a requirements-driven approach to validate the prototype. It established a correspondence between the research questions, functional requirements, the lakehouse architecture design and the evaluation criteria (see Table 4).

The evaluation was divided into five aspects. ‘Data Ingestion’ was used to examine whether the prototype could integrate environmental sensing, soundscapes, hydrology, geospatial data, biodiversity and open data, whilst retaining data sources, timestamps and relevant metadata. ‘Data Transformation’ assessed whether the ETL/ELT processes could complete data cleansing, format standardisation, and conversion between the Bronze, Silver and Gold Layers. ‘Metadata and Governance’ examines the data catalogue, data quality information, quality labels and governance

records. ‘Monitoring and Lineage’ assesses whether pipeline execution records, audit logs, version information and data lineage are sufficient to support data tracing, repeatability and monitoring. Finally, ‘Prototype Validation’ verifies the complete process from data ingestion, transformation and storage through to visualisation and analysis by an end-to-end demonstration.

The focus of this study’s evaluation is on confirming whether the proposed lakehouse architecture satisfies the requirements for conservation data integration, governance, and analysis, rather than measuring the predictive performance of individual analytical models.

Table 4. Evaluation matrix.

Evaluation Area	Pass Criterion	Evidence
Data Ingestion	At least four distinct data domains are successfully ingested into the Bronze Layer, with source and ingestion metadata retained.	Bronze datasets, source files and metadata fields
Automated Ingestion	The ingestion workflow can load configured source files into the Bronze Layer without manual restructuring of each dataset.	Notebook execution results and pipeline logs
Data Transformation	Bronze datasets are successfully cleaned, standardised and transformed into validated Silver and Gold datasets.	SQL validation, schemas and record-count checks
Metadata & Governance	Evaluated datasets contain catalogue entries, metadata records and data quality information.	Dataset Catalogue, metadata tables and Data Quality Dashboard
Monitoring & Lineage	Pipeline executions generate status, source, target and record-count information.	MLflow, Audit Log and Delta History

4. Conservation Lakehouse Architecture

4.1 Current State Assessment and Requirements

The Wildlife Trusts has provided a system architecture currently under planning (Figure 4). Based on the Azure cloud platform, it comprises key modules, including data sources, data ingestion, lakehouse storage and processing, and analytics and applications, utilising Databricks, Delta Lake, and the Medallion Architecture as the core of its data management. Furthermore, the client has provided a list of potential future data sources, including Digital Reserve Sensors, On-site Sensors, the LoRaWAN Network, and Open GIS Data. This data spans various domains such as biodiversity, environmental monitoring, hydrology, geospatial information, human activity and risk, and features diverse data formats, varying update frequencies and transmission methods, thereby increasing the complexity of data integration, governance and analysis.

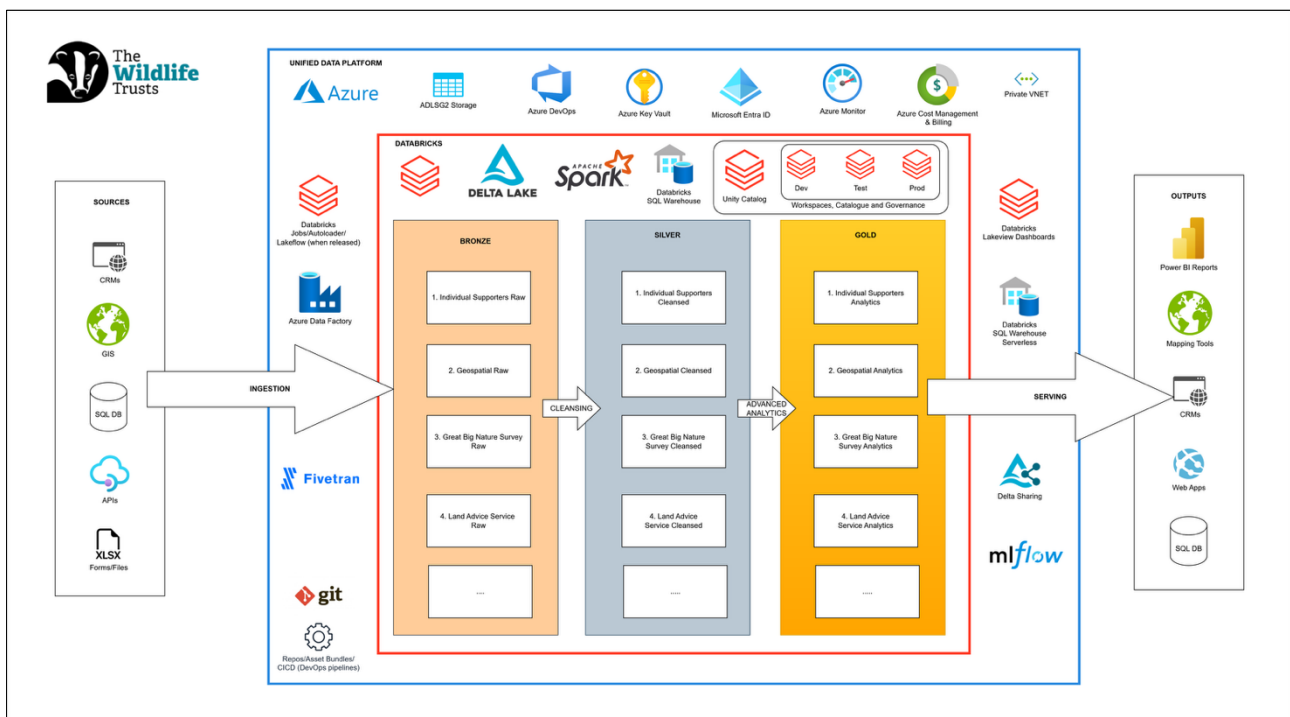


Figure 4. Existing architecture provided by The Wildlife Trusts.

Source: Client-provided project documentation.

As this project is still in the preliminary planning phase and the sensor network and data platform have not yet been formally established, the client is unable to provide actual data; instead, this study will utilise the aforementioned information to generate synthetic data to simulate future data collection and analysis processes; data egress and real-time streaming will be analysed through conceptual design, whilst cost recommendations will be provided solely from an architectural design perspective, without conducting precise estimates.

Given the existing architecture, data sources, and research constraints, the greatest challenge is not the sensor technology itself, but the lack of a data platform capable of effectively integrating heterogeneous data sources. Diverse data sources present different formats, update frequencies, data volumes, and transmission methods. Without an effective process for data acquisition, governance, and analysis, data integration across domains and the development of future digital conservation applications may be complicated. Therefore, this study has established a prototype centred on a lakehouse to validate the processes for integrating, transforming, and analysing heterogeneous sensor data. The feasibility of the data platform has been verified through dashboards, metadata, data quality and APIs, serving as a proof of concept for future applications integrating CRM, Digital Twin, AR and AI analytics.

4.2 Data Sources Overview

The data sources provided by the client primarily comprise project documentation, sensor technical specifications, communication protocols such as LoRaWAN, Digital Reserve Sensors, and relevant open data. To analyse the impact of different data types on the platform architecture, this study categorises them according to data type, update patterns, data format and acquisition methods into Acoustic Data, Imaging and Remote Sensing Data, Environmental Time-series Data, Spatial and Survey Data, External and Reference Data, and Connectivity Infrastructure, as shown in Figure 5. Each data category exhibits distinct differences in terms of format, update frequency, data volume, spatial characteristics and data egress mechanisms, thereby imposing different design requirements on the data platform.

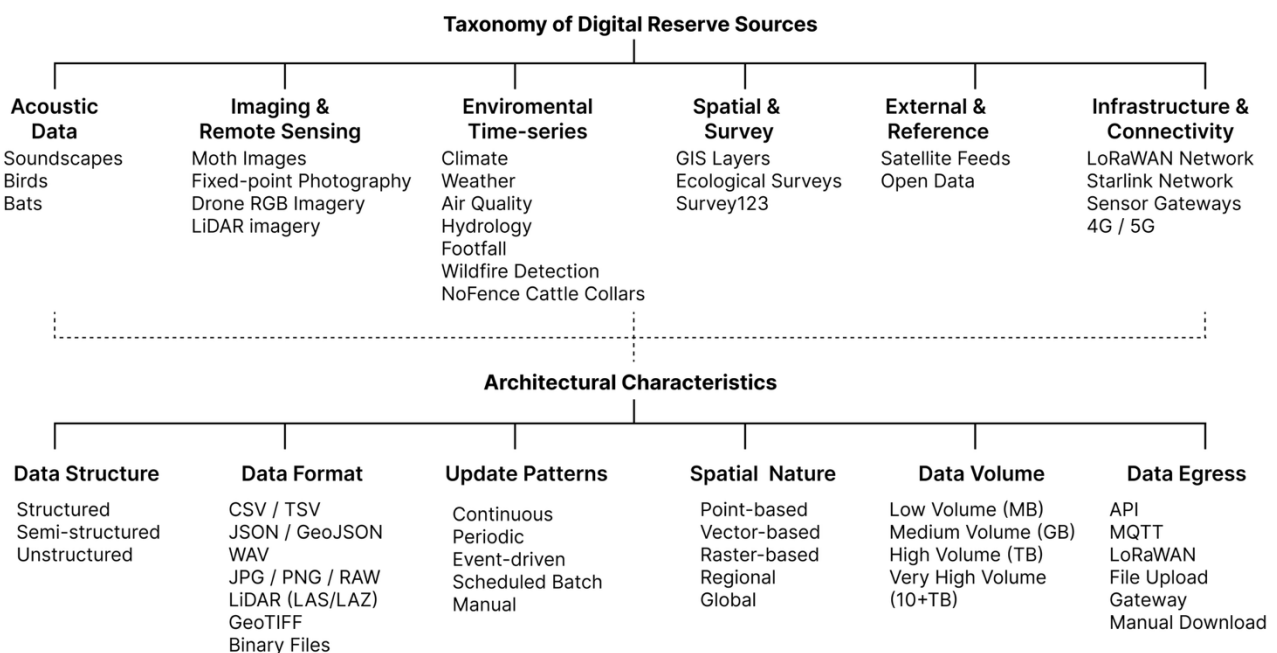


Figure 5. Data-centric taxonomy of digital reserve datasets.

Data such as soundscape recordings, images, drone aerial footage and LiDAR are primarily stored in formats such as WAV, JPEG, RAW, GeoTIFF and LAS/LAZ. As high-volume unstructured or semi-structured data, they are well suited to scalable object storage, with metadata used to manage information such as file location, device, time, and geographical coordinates. By contrast, data from environmental, hydrological, pedestrian flow, wildfire monitoring and NoFence Cattle Collars tend to be generated continuously and exhibit time-series characteristics; such data can be received via LoRaWAN gateways, REST APIs, file uploads or manual downloads. In this study, the relevant data flows are simulated in batch mode.

GIS data, ecological surveys, Survey123, satellite data and external open data are acquired via APIs, file downloads or periodic batch processing. These require format standardisation, spatial reference management and source documentation to support cross-domain data integration. As some data sources rely on third-party platforms, proprietary systems or commercial licences, the data platform must consider factors such as licensing restrictions, system compatibility and vendor lock-in.

Given these data characteristics, Digital Reserve encompasses large-scale binary data, time-series data, geospatial data, and external reference data and therefore requires support for data ingestion, storage, and processing strategies. Consequently, this study employs the lakehouse and Medallion architectures to develop a prototype to validate functionalities such as heterogeneous data integration, metadata, data quality, lineage, and analytics dashboards. The integration of comprehensive real-time streaming with actual sensor networks is identified as a future direction for expansion (see Table 5).

Table 5. Summary of digital reserve data characteristics.

Data Domain	Example Sources	Update Pattern	Architectural Implication
Acoustic Data	Soundscapes (Birds & Bats)	Annual	Large binary storage and metadata management
Imaging and Remote Sensing Data	Moth Images, Fixed-point Photography, Drone RGB & LiDAR	Annual / Periodic	Object storage and large-file management
Environmental Time-Series Data	Climate, Weather, Hydrology, Footfall, Wildfire, NoFence Collars	Continuous / Mixed	Time-series processing and support for batch and streaming
Spatial and Survey Data	Ecological Surveys, GIS Layers, Survey123	Periodic	Spatial reference management and geospatial analytics
External and Reference Data	Satellite Feeds, Open Data	Source-dependent	External integration and data standardisation
Connectivity Infrastructure	LoRaWAN, Starlink, 4G and Sensor Gateways	N/A	Supports data egress and future sensor-network integration

4.3 Lakehouse Data Pipeline Design

Based on the requirements analysis in Section 4.1 and the data source analysis in Section 4.2, this study has designed a Lakehouse Data Pipeline centred on the Medallion Architecture to support the ingestion, processing, governance and analysis of heterogeneous data. Figure 6 shows the proposed prototype lakehouse pipeline.

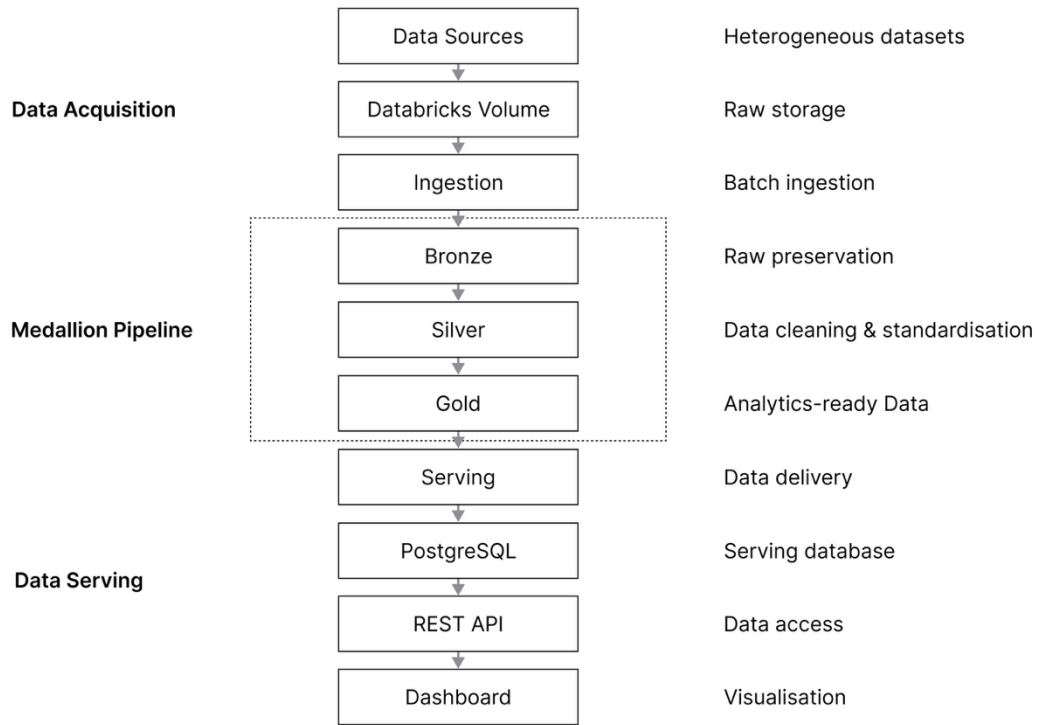


Figure 6. Prototype lakehouse data pipeline.

4.3.1 Ingestion Layer

The Ingestion Layer imports heterogeneous data from various sources into the lakehouse and serves as the data source for the subsequent Medallion Pipeline. Depending on the differences in data sources and formats, the prototype employs multiple ingestion methods, including file-based imports such as CSV, Excel and GeoJSON. Raw data is first uploaded and stored in the raw storage of the Databricks volume; after being categorised by data domain, it then enters the bronze layer. As this research is a proof of concept and real-time LoRaWAN sensor streaming has not yet been implemented, the current focus is on batch ingestion and API access, with real-time data ingestion identified as a future area for expansion.

4.3.2 Bronze Layer

The Bronze layer stores the raw data ingested by the ingestion layer. It preserves the data in an immutable format to ensure data integrity and traceability whilst providing the data required for subsequent reprocessing. Upon ingestion, the data undergoes standardisation of column naming

conventions to minimise schema discrepancies arising from different data sources, and metadata such as `ingestion\_time`, `source\_system` and `source\_file` is added to support data source tracking and governance. All data is stored as Delta Tables, without any data cleansing or transformation. The system also performs basic source validation on certain data sources to confirm successful ingestion, serving as the foundation for subsequent standardisation and quality processing in the Silver Layer.

4.3.3 Silver Layer

The Silver Layer transforms raw data from the Bronze Layer into standardised datasets. At this stage, duplicate records are first removed, and time formats, data types, and some field names are standardised to minimise schema discrepancies across data sources. For datasets on environment, hydrology, and connectivity, the validation rules for data quality have been implemented in the system, which employs a `quality\_flag` for determining missing values, outliers in timestamps, and measurements outside the set reasonable range. In addition, connectivity data are classified according to basic connectivity quality levels (Good, Moderate, Poor) using RSSI and SNR, serving as secondary data for further analysis. The processed data is stored in Delta tables, serving as a basis for the Gold Layer to create analytical data products.

4.3.4 Gold Layer

The Gold Layer converts standardised data from the Silver Layer into analytics-ready data products for analysis and decision-making. The system creates thematic summary tables for different data domains, including datasets such as environmental monitoring, hydrology, soundscapes and connectivity status. It generates statistical measures on a daily basis through aggregations, which include average, maximum, minimum, and number of data entries. To ensure dependable analytical results, the Gold Layer performs statistical calculations only on data with a `quality\_flag` set to `valid`. The system also generates Data Quality Summaries and Sensor Coverage Summaries to support data quality monitoring. Each Gold Dataset is annotated with metadata, including data standards, EBV classification, FAIR status, lineage, and creation time, whilst execution information is recorded via MLflow and pipeline audit to support data governance, traceability, and pipeline monitoring.

4.3.5 Serving Layer

The Serving Layer is responsible for converting the analytical data generated by the Gold Layer into data services that can be queried and visualised by the application layer. In the prototype, selected Gold data products are first exported and loaded into PostgreSQL as the serving database; data access is then provided via a REST API by the NestJS backend; and finally, the analytical results are presented via a React/Next.js dashboard. In this design, the lakehouse data processing flow is separated from the front-end application, providing access to Gold data through a common interface.

4.4 Environmental IoT Data Pipeline Example

The data collected from environmental IoT observations are used to demonstrate data processing in the lakehouse architecture and its availability to applications and dashboards.

4.4.1 Source Data and Ingestion

Since the documentation from The Wildlife Trusts mainly describes environmental monitoring requirements but does not contain actual sensor observations, this study uses the publicly available `_Environmental Sensor Telemetry Data_` dataset (Stafford, n.d.) as an alternative data source. This dataset includes device ID, timestamp, temperature, humidity, carbon monoxide, and smoke columns that have been loaded into Databricks Volume in batches for later processing.

4.4.2 Medallion Pipeline

The imported data is processed through the Medallion pipeline. Raw observations are stored with ingestion metadata. Later processing would remove duplicate records and standardise timestamps and data types; it also adds quality flags to unusual values. Finally, the Gold layer aggregates data by date and device, calculating metrics such as observation counts, temperature statistics, and average humidity, carbon monoxide, and smoke levels to produce data products ready for applications. Once complete, the Gold data is exported and loaded into PostgreSQL as the application's database. Figure 7 illustrates this transformation.

(a) Bronze: raw observations with ingestion metadata

	1.2 raw_timestamp	A _C device	1.2 temperature	1.2 humidity	1.2 co	1.2 smoke	🕒 ingestion_time	A _C source_system
1	1594512094.386	b8:27:eb:bf:9d:51	22.7	51	0.005	0.0204	2026-08-03T16:24:25.327+00:00	environmental_sensor
2	1594512098.074	b8:27:eb:bf:9d:51	22.6	50.9	0.005	0.0205	2026-08-03T16:24:25.327+00:00	environmental_sensor
3	1594512101.761	b8:27:eb:bf:9d:51	22.6	50.9	0.005	0.0204	2026-08-03T16:24:25.327+00:00	environmental_sensor

(b) Silver: standardised and validated observations

	🕒 event_timestamp	📅 date	A _C device	1.2 temperature	1.2 humidity	1.2 co	1.2 smoke	A _C quality_flag
1	2020-07-12T00:01:34.000+00:00	2020-07-12	b8:27:eb:bf:9d:51	22.7	51	0.005	0.0204	valid
2	2020-07-12T00:01:38.000+00:00	2020-07-12	b8:27:eb:bf:9d:51	22.6	50.9	0.005	0.0205	valid
3	2020-07-12T00:01:41.000+00:00	2020-07-12	b8:27:eb:bf:9d:51	22.6	50.9	0.005	0.0204	valid

(c) Gold: daily environmental aggregation

	📅 date	A _C device	🔢 record_count	1.2 avg_temperature	1.2 min_temperature	1.2 max_temperature	1.2 avg_humidity	1.2 avg_co	1.2 avg_smoke
1	2020-07-12	b8:27:eb:bf:9d:51	23406	21.8	21	22.7	51.4	0.005	0.0206

Figure 7. Environmental IoT data across the Databricks Medallion layers.

4.4.3 API Serving and Dashboard Visualisation

The REST API reads Gold data from PostgreSQL and provides summaries, daily statistics, device information, and recent observations via `/api/environmental/*``. The Next.js front-end retrieves the

required data via the API to display environmental summaries, device statistics and time-series trends, without directly accessing the database, thereby maintaining separation between the data supply layer and the user interface. Figure 8 provides an example API output used by the dashboard.

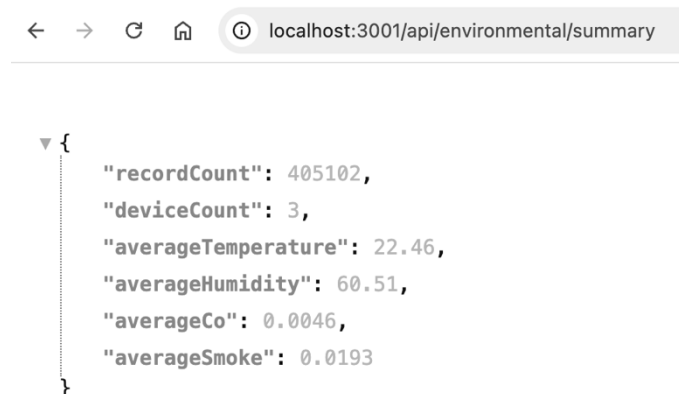


Figure 8. Example environmental summary returned by the API.

4.4.4 Outcome and Limitations

This case study illustrates the end-to-end process workflow for environmental data, from ingestion to dashboard visualisation. However, the prototype currently uses a public dataset rather than actual sensor observations. The data is processed in batches, and the pipeline still requires manual execution, as automated scheduling via Databricks Jobs has not yet been implemented. This dataset covers only three devices, and the carbon monoxide and smoke fields are representative sensor signals only, which should not be used to determine actual environmental alerts.

4.5 Metadata, Governance and Security Framework

The lakehouse prototype developed in this research focuses not only on data integration and output but also on the fundamental design of metadata, data governance, and security controls. Given the diverse sources of conservation data and the fact that data formats, update frequencies, and intended uses vary, it would be difficult to support subsequent analysis, tracking, and application without a consistent management framework. The following discussion explains how the prototype implements initial mechanisms for data traceability, quality management, and security protection through data-processing workflows and cloud platforms.

4.5.1 Metadata and Lineage



Figure 9. Digital Reserve ELT pipeline in Databricks.

The prototype incorporates metadata and data lineage into the lakehouse design to improve the traceability and transparency of data-processing workflows. During processing, the system retains or adds information such as source files, data domains, ingestion times, quality statuses and the creation time of Gold data, enabling users to identify data sources, processing stages and output results. As shown in Figure 9, the Databricks job sequentially executes Bronze data ingestion, Silver cleansing and standardisation, and Gold analysis and aggregation. It exports the Gold dataset to PostgreSQL for dashboard queries. This flowchart illustrates the execution order, dependencies and success status of each task whilst providing data lineage at both the application and task levels. However, as the current prototype primarily stores delta data via volume paths, automatic lineage tracking at the table and column levels—as supported by Unity Catalog—has not yet been implemented. Figure 10 provides a section-level lineage for the Habitat and Environment view.

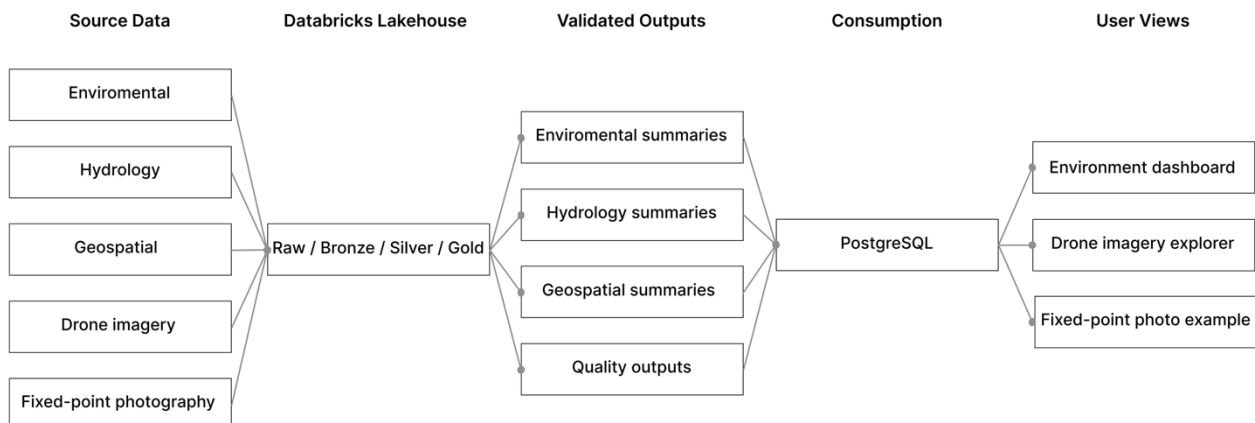


Figure 10. Section-level data lineage for the Habitat & Environment view.

4.5.2 Governance and Data Quality

Governance and data quality control in this prototype are primarily achieved through data standardisation, quality tagging and output validation. During the processing workflow, the system performs field cleaning, format harmonisation, duplicate removal, and type conversion across different datasets, while adding quality tags to certain data to help identify missing values, formatting

errors, or suspicious records. Taking the environmental, hydrology, and geospatial datasets as examples, quality control is reflected in the standardisation of time fields, numerical conversions and the verification of summary outputs, ensuring that the data used in the dashboard is highly consistent and interpretable. However, as a comprehensive data steward system, formal update policies and a comprehensive quality rule repository have not yet been established; this governance and quality mechanism remains a preliminary implementation at the prototype stage.

4.5.3 Security and Limitations

The security controls in this prototype primarily rely on the basic safeguards required for the prototyping phase, rather than on a full enterprise-level security governance framework. Data processing is primarily performed within the Databricks environment; raw and intermediate data are not exposed directly to front-end applications. The data queried by the dashboard is primarily sourced from the exported ‘Gold’ dataset and the corresponding tables in PostgreSQL, thereby reducing the need for direct access to raw data. Furthermore, database connection details are managed via usernames, passwords and secrets, thereby avoiding the direct inclusion of sensitive credentials in notebooks or front-end configuration files. However, granular role-based access control, formal audit logs, sensitivity classification and data retention policies have not yet been fully implemented; consequently, this security framework remains a preliminary implementation at the research prototype stage.

4.6 Sample Data Catalogue

Representative datasets used in the prototype are compiled here to illustrate the scope of data, subject types and output formats covered by the Digital Reserve Lakehouse prototype³. As this research involves prototype development, the data catalogue is not a comprehensive enterprise-level catalogue. Still, it is intended to demonstrate how the system integrates various types of conservation-related data and supports subsequent analysis and dashboard visualisation. The prototype currently covers data domains including environmental monitoring, hydrological monitoring, connectivity quality, soundscapes and biodiversity, as well as geospatial data and field surveys. Within the system, these data do not all exist at the same processing stage; rather, they correspond respectively to raw inputs, standardised analytical data, and final thematic outputs for users.

³ Sample datasets and source references were drawn from Kaggle, NASA FIRMS, the Environment Agency, the Office for National Statistics and data.gov.uk.

Table 6. Sample data catalogue of representative dashboard views in prototype.

Page	Subpage	Dataset	Prototype role	Example output
Biodiversity	Soundscape	Soundscape detections	Acoustic biodiversity evidence	gold_soundscape_daily
Habitat & Environment	Environmental	IoT telemetry	Environmental monitoring input	gold_environment_daily
Risks & Alerts	Wildfire Monitoring	Wildfire risk records	Risk monitoring input	gold_wildfire_overall
Site Management	Human Activity	Footfall records	Human pressure proxy	gold_human_activity_proxy
Data & System Health	Data Quality	Validation records	Quality assurance monitoring	gold_data_quality_summary
Innovation Lab	Scenario Lab	Environmental and hydrology summaries	Scenario exploration	gold_environment_daily

Through the organisation of the sample data catalogue, the datasets in the prototype are not merely technical data assets but can also be understood as information sources that support conservation decision-making, monitoring and interpretation, and dashboard visualisation. Table 6 summarises a representative data catalogue based on the main dashboard pages in the prototype. To maintain readability, we list only one subpage from each page as an example; a more comprehensive list of pages, subpages, and corresponding output data tables is provided in Appendix A.

4.7 Output and Consumption: Prototype Dashboard and User Views

The prototype's output and data provision layer explains how Gold data in the lakehouse is transformed via PostgreSQL, REST APIs, and the dashboard into a front-end interface that allows users to view, compare, and explore the data, and further discusses its interaction methods and prototype limitations.

4.7.1 Prototype Dashboard and User Views

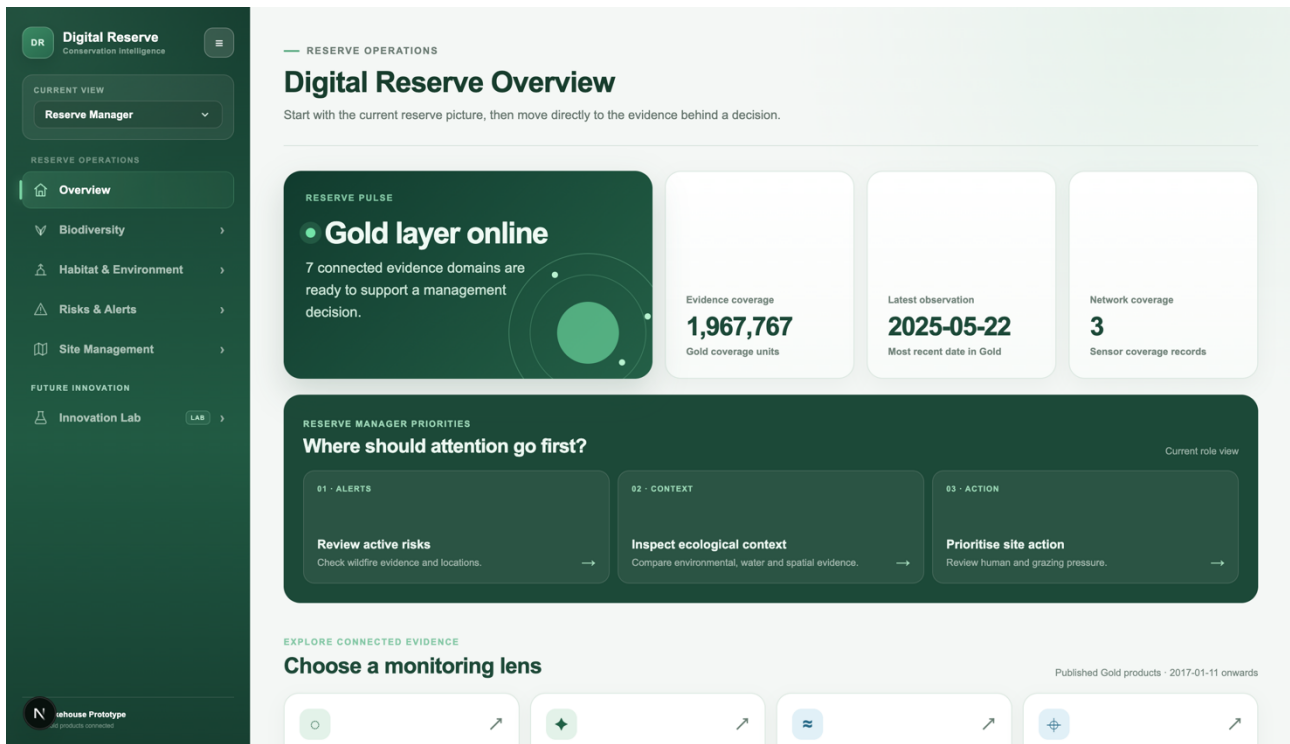


Figure 11. Overview dashboard of the Digital Reserve prototype.

As shown in Figure 11, the home page provides an overview that integrates data availability, monitoring scope, the latest observations, and topic navigation, which allows users to quickly view the overall status of the conservation site. This interface does not use raw data directly; instead, it queries tables published to PostgreSQL after Databricks processes the data, translating back-end analysis results into easier-to-understand, task-oriented views. The prototype currently includes two role-based views: ‘Reserve Manager’ and ‘Data Lead’, and organises thematic modules—such as Biodiversity, Habitat and Environment, Risks and Alerts, Site Management, Data and System Health, and Innovation Lab—via a sidebar.

4.7.2 Evidence Interaction and Experimental Interfaces

The prototype also attempts to transform different types of conservation evidence into interactive front-end interfaces. Users can view summary results on topics such as soundscape monitoring, hydrological changes, habitat conditions, risk alerts and data quality via thematic pages, and explore and compare them using cards, indicators, summary tables and thematic sub-pages. This means data consumption is no longer limited to displaying a single chart; instead, it organises multiple Gold outputs into task-oriented evidence views. In contrast, the Digital Twin, Scenario Lab, Field AR, and AI Intelligence Brief within the Innovation Lab are more exploratory, serving as experimental interfaces for future use.

4.7.3 Limitations of the Serving Layer

Although this dashboard demonstrates how lakehouse outputs are utilised by the front end, its serving layer still has notable prototypical limitations. Currently, not all back-end data products have been fully mapped to the user interface; some pages still focus primarily on representative outputs and demonstration scenarios. Certain modules within the Innovation Lab are best understood as conceptual demonstrations rather than formal decision-making tools. The role-based views also mainly reflect information organisation and reading priorities, rather than a fully implemented access control mechanism.

4.8 Final Lakehouse Architecture Blueprint

The lakehouse architecture blueprint ultimately proposed by this study is shown in Figure 12. Centred around Databricks, the overall architecture integrates diverse and heterogeneous conservation data sources, data processing workflows, data delivery components, and prototype dashboard outputs into a single end-to-end workflow, covering data ingestion, processing, publication, and front-end utilisation. Upon entering the system, raw data is first imported into a Databricks Volume, which serves as a unified storage and ingestion gateway.

The data then enters the Medallion process, where data retention, cleansing, standardisation and thematic transformation are completed within the same lakehouse workflow. Once processing is complete, ‘Gold’ data products are published to PostgreSQL and made available via a REST API for dashboard queries and visualisation. In other words, this blueprint does not allow dashboards to access raw data directly; instead, it uses pre-processed, task-oriented outputs as the foundation for the front-end interface. Compared to more linear models of data storage and front-end querying, this architecture more clearly delineates the roles and boundaries between data engineering, data provision and the user interface.

This blueprint also supports multiple themes, including biodiversity, habitat and environment, site management, risk monitoring and future-facing experimental modules, reflecting that the prototype focuses not only on data processing itself, but also on how data is utilised through different interfaces and use cases. Although this blueprint remains a research prototype rather than a complete enterprise-grade deployment, it has demonstrated the feasibility of a conservation-focused lakehouse at the prototype level. It provides a consistent, scalable architectural foundation for integrating diverse ecological data, analytical outputs, and decision-oriented interfaces.

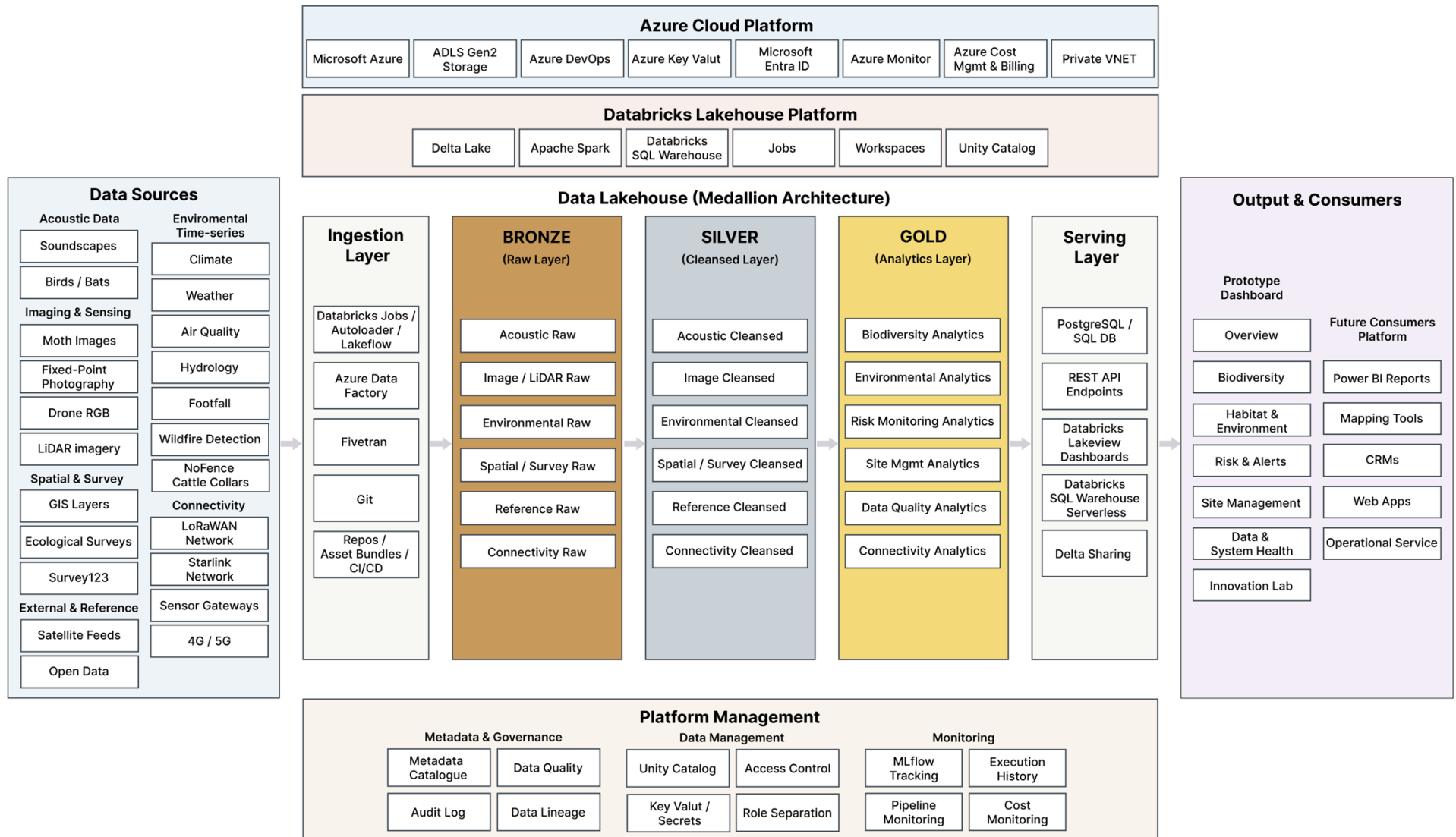


Figure 12. Final Lakehouse architecture blueprint of the Digital Reserve prototype.

5. Evaluation and Discussion

5.1 Evaluation of the Proposed Architecture

The proposed lakehouse prototype is evaluated across five dimensions, as set out in the 'Methodology' section. The prototype ingests data from multiple conservation domains into a Databricks workflow. After cleansing, standardising, and thematically transforming the data, the prototype produces 'Gold' outputs suitable for dashboards and the serving layer. The results demonstrate that the lakehouse architectural design is complete and validate the end-to-end process, from data loading and processing to front-end applications. This proves the technical feasibility of the prototype for integrating and applying conservation data.

Table 7. Evaluation summary of the prototype.

Criterion	Summary of Findings	Evidence	Result
Data Ingestion	Multiple conservation data domains were successfully ingested into the Bronze Layer, with source and ingestion metadata retained.	10 data domains integrated into the Bronze Layer	Achieved
Automated Ingestion	Configured source files could be repeatedly loaded into the Bronze Layer through reusable ingestion notebooks, although some setup remained prototype-specific.	Reusable ingestion notebooks executed across configured datasets	Partially Achieved
Data Transformation	Bronze datasets were cleaned, standardised and transformed into validated Silver and Gold datasets.	37 Gold tables and 6,159 Gold-layer records generated	Achieved
Metadata & Governance	Evaluated datasets included catalogue entries, metadata records and data quality information.	Dataset Catalogue, metadata records and Data Quality Dashboard implemented	Achieved
Monitoring & Lineage	Pipeline executions captured status, source, target and record-count information, while fine-grained lineage remained limited.	Execution records available through MLflow, Audit Log and Delta History	Partially Achieved

As shown in Table 7, the levels of completion for 'data ingestion' and 'data transformation' are relatively high, with data from different sources and formats already capable of being loaded, processed, and output via a consistent workflow. In contrast, the automated ingestion, metadata, governance, monitoring, and lineage features currently offer only basic functionality, including repeatable data processing workflows, basic metadata management, data quality checks, and workflow monitoring. Due to constraints related to platform permissions, deployment conditions and

project scope, these features have only been partially implemented in terms of automation, governance depth and field-level lineage. As a result, they have not yet met the standards required for full enterprise deployment. These functions remain at the prototype stage, with scope for further refinement and expansion.

Table 8. Page-level summary of linked Gold outputs.

Page	Linked Gold tables	Aggregated row count
All published Gold data	37	6,159
Overview	8	3,224
Biodiversity	10	88
Habitat & Environment	14	3,886
Risks & Alerts	8	2,086
Site Management	6	83
Data & System Health	3	1,431
Innovation Lab	4	1,429

Table 8 summarises that this study has published 37 Gold tables on the Databricks platform, totalling 6,159 Gold-layer records distributed across the relevant pages of each category.

5.2 Benefits and Limitations

Based on the assessment results in Section 5.1, the prototype's value lies primarily in its practical application. This discussion focuses on how the prototype supports integrating diverse conservation data and generating analytical results for the dashboard, serving as the foundation for a future production data platform. At the same time, there is still room for improvement in the prototype's automation, governance and production deployment. Figure 13 summarises the main benefits identified from the prototype evaluation.

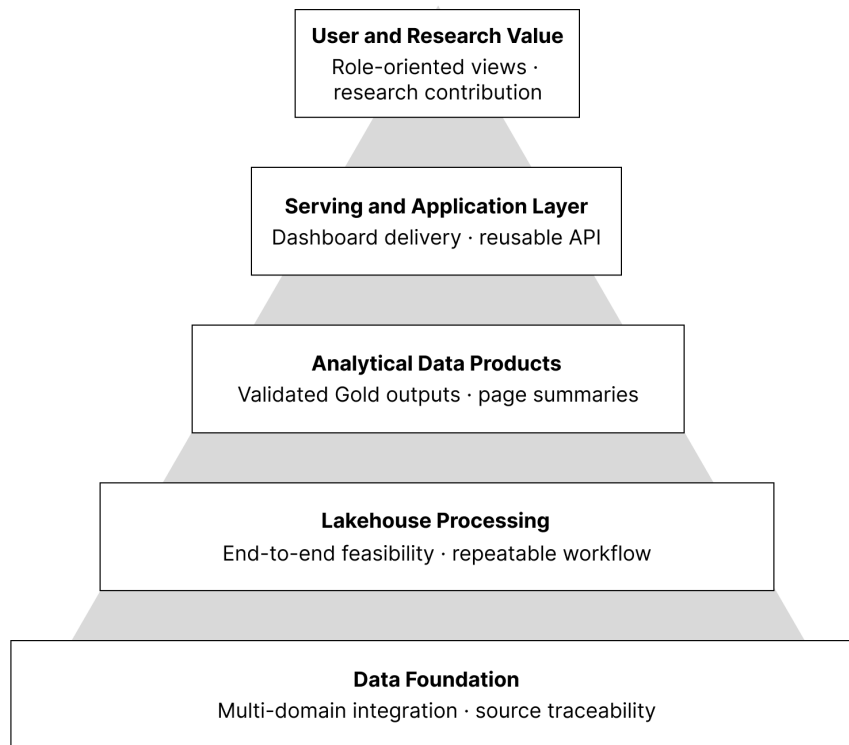


Figure 13. Layered benefit model derived from the prototype evaluation.

The main benefits of this prototype can be summarised as follows:

1. **End-to-end data workflow:** The prototype connects Databricks' data processing workflow, data provision via PostgreSQL and APIs, and dashboard visualisation capabilities to form a complete end-to-end data pipeline. The prototype provides initial evidence of the feasibility of this architecture.
2. **Integration of heterogeneous conservation data:** The prototype integrates various types of conservation data that address the requirement for heterogeneous data integration set out in the project brief. It also brings together different types of data in a single processing workflow.
3. **Dashboard-ready outputs:** The “Gold” process output is not only limited to backend tables but also available to dashboards via the data serving layer. This enables the data to be presented in the form of page-level summaries, metric cards and topic pages.
4. **User-orientated evidence views:** The dashboard organises information around the needs of users such as reserve managers and data leads. This helps transform data into evidence views that are easier to read, compare, and interpret, and are not confined to tables.
5. **Foundation for future development:** Although the system is still in the prototype stage, the overall architecture has identified areas for future development, including Azure deployment, scheduled workflows, formal governance mechanisms, and further integration of the BI layer.

This project is currently still at the prototype stage and has not yet become a fully operational platform. The prototype evaluation shows the system is ready for batch processing of multiple datasets, Databricks execution, Gold output creation, data cataloguing, auditing, and validation. If further deployment is required in the future, it will be necessary to strengthen work-process scheduling, real-time sensor data streaming, data quality monitoring, Unity Catalog management, and security controls. Integration with Azure, Power BI, CRM, etc., and existing systems will also be necessary. Therefore, these constraints can be regarded as future improvements, but this will not diminish the value of the current prototype execution.

5.3 Recommendations and Future Roadmap

According to the assessment, further development should focus on enhancing the existing prototype into an operational data platform. The roadmap is therefore organised into three maturity stages: prototype implementation, current development, and future production. Table 9 distinguishes between deliverables already completed, capabilities currently under development, and areas that will require further enhancement should the project proceed to full-scale deployment.

Table 9. Roadmap from prototype implementation to future production.

Roadmap stage	Key recommendations	Purpose
Prototype implementation	Consolidate the completed Databricks pipeline, Gold outputs, PostgreSQL/API serving layer, dashboard prototype, dataset catalogue, audit log and SQL validation.	Confirm the current baseline.
Current development	Improve dashboard filtering, multi-variable charts, generic API endpoints, latency checks, formal catalogue, lineage, monitoring and quality metrics.	Improve usability, observability and operational readiness.
Future production	Introduce Docker/CI/CD, queue-based sync, Azure deployment, live sensor streaming, large-scale audio/LiDAR pipelines, Unity Catalog governance and BI/CRM integration.	Support production-scale Digital Reserve use.

In the short term, there is no urgency to introduce many new features; instead, the focus should be on stabilising the completed prototype. Specifically, efforts should focus on the following: first, strengthening data validation; second, maintaining the data catalogue on an ongoing basis; and third, supplementing the mappings between 'data sources and output results'. The overarching aim of these three measures is to ensure that all dashboard pages utilise reliable Gold-layer outputs. Concurrently, the dashboard will be further enhanced by incorporating filters for species or data categories, multivariate environmental charts, monitoring of API response times, and natural-language

explanations of indicators. These adjustments will enable the front end not only to present data, but also to provide a view of conservation evidence that is easier to compare and interpret.

The next stage should align the prototype with operational requirements by improving process automation and scheduling, metadata management, data quality monitoring, and integration with geospatial data and ingestion tools. When the system moves to full deployment, additional thought will be needed on infrastructure, data updates, system integration, security governance, and maintenance resource allocation. The prototype shows some cost awareness by using batch processing, selectively publishing its 'Gold' tier outputs, and employing a lightweight PostgreSQL data provisioning layer. However, before full deployment, it will be necessary to redeploy this data-serving layer within The Wildlife Trusts' existing Azure architecture and then reevaluate it. Appendix B summarises the recommended update frequencies and scheduling methods for the various data domains, while Appendix C provides the prototype repository structure, showing how the implementation files and configuration components are organised.

6. Conclusion

6.1 Summary of Findings and Contributions

In this research, we explore how The Wildlife Trusts could use a lakehouse architecture to bring together conservation data from various sources and formats as part of developing the Digital Reserve data platform. The prototype results show that the Databricks and Medallion data pipeline provides a consistent way to process data, progressively turning raw data into analytical outputs that can then be used in dashboards. This shows that conservation data does not have to be scattered across several files or tables; it can be organised and turned into a summary of conservation evidence. This study makes a considerable contribution to the area because it operationalises rather abstract requirements outlined in the project brief into the lakehouse architecture prototype. The architectural design, data catalogue example, Databricks pipeline, PostgreSQL/API serving layer, and dashboard prototype all show how it can be done even at the prototype stage. Although the system is not yet ready for production use, it provides a clear starting point for building the Digital Reserve data platform.

6.2 Practical Implications and Future Research

From the Wildlife Trusts' perspective, this prototype could serve as a framework for developing a Digital Reserve data platform. This also demonstrates that data is collected at the point of generation and then fed into the lakehouse; through analysis of outputs, APIs, and dashboards, this processed data becomes conservation evidence that users can monitor and interpret. It also provides a useful reference for future cross-domain monitoring, dashboards and data governance planning, particularly

when organisations understand data engineering workflows and how they support front-end applications instead of just maintaining disparate data systems. Further research to validate this prototype could include The Wildlife Trusts applying it to real data over a longer period. This would help verify features such as real-time sensor streams, incremental updates, data catalogues, and role-based access control. User testing is also essential to determine whether conservationists, data administrators and other stakeholders truly understand the dashboard’s outputs.

References

- AbouZaid, A., Barclay, P.J., Chrysoulas, C. and Pitropakis, N. (2025) ‘Building a modern data platform based on the data lakehouse architecture and cloud-native ecosystem’, *Discover Applied Sciences*, 7, article 166.
- Ahmad, Y., Youssef, A., Kearns, M., Bhagavathula, R., Bhat, A., Jayachandran, P., Narayanan, S., Rao, S., Subramanian, S. and Varadarajan, K. (2023) ‘Microsoft Purview: A system for central governance of data’, in *Proceedings of the 2023 International Conference on Management of Data*. New York: Association for Computing Machinery.
- Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M. and Ayyash, M. (2015) ‘Internet of Things: A survey on enabling technologies, protocols, and applications’, *IEEE Communications Surveys & Tutorials*, 17(4), pp. 2347–2376.
- Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A., Świtakowski, M., Szafranski, M., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., Xin, R. and Zaharia, M. (2020) ‘Delta Lake: High-performance ACID table storage over cloud object stores’, *Proceedings of the VLDB Endowment*, 13(12), pp. 3411–3424.
- Armbrust, M., Ghodsi, A., Xin, R. and Zaharia, M. (2021) ‘Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics’, in *Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR 2021)*, 11–15 January 2021, online.
- Ayele, E.D., Das, K., Meratnia, N. and Havinga, P.J.M. (2018) ‘Towards a new opportunistic IoT network architecture for wildlife monitoring system’, *Procedia Computer Science*, 130, pp. 403–410.
- Binley, A.D., Vincent, J.G., Rytwinski, T., Soroye, P. and Bennett, J.R. (2024) ‘Making the most of existing data in conservation research’, *Perspectives in Ecology and Conservation*, 22(2), pp. 122–128.

- Boukraa, D., Bala, M. and Rizzi, S. (2024) ‘Metadata management in data lake environments: A survey’, *Journal of Library Metadata*, 24(4), pp. 215–274.
- Camacho-Rodríguez, J., Agrawal, A., Gruenheid, A., Gosalia, A., Petculescu, C., Aguilar-Saborit, J., Floratou, A., Curino, C. and Ramakrishnan, R. (2024) ‘LST-Bench: Benchmarking log-structured tables in the cloud’, *Proceedings of the ACM on Management of Data*, 2(1), article 59, pp. 1–26.
- Chandra, R., Chen, H., Matharu, R. et al. (2025) ‘Unity Catalog: Open and universal governance for the lakehouse and beyond’, in *Companion of the 2025 International Conference on Management of Data*. New York: Association for Computing Machinery.
- Data.gov.uk (n.d.) Weather data. Available at:
<https://www.data.gov.uk/collections/environment/weather> (Accessed: 3 September 2026).
- Databricks (n.d.) What is the medallion lakehouse architecture? Available at:
<https://docs.databricks.com/aws/en/lakehouse/medallion> (Accessed: 23 August 2026).
- Environment Agency (n.d.) Flood monitoring API. Available at:
<https://environment.data.gov.uk/flood-monitoring/doc/reference> (Accessed: 3 September 2026).
- Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) ‘Design science in information systems research’, *MIS Quarterly*, 28(1), pp. 75–105.
- Jacques-Silva, G., Kalyvianaki, E., Cohn-Gordon, K., Meguid, A., Nguyen, H., Ben-David, D., Nayak, C., Saravagi, V., Stasa, G., Papagiannis, I., Taieb, D., Tamirat, K., Wu, H., Xi, B., Zhang, T. and Zhou, Q. (2025) ‘Unified Lineage System: Tracking data provenance at scale’, in *Companion of the 2025 International Conference on Management of Data*. New York: Association for Computing Machinery, pp. 457–470.
- Jain, P., Kraft, P., Power, C., Das, T., Stoica, I. and Zaharia, M. (2023) ‘Analyzing and comparing lakehouse storage systems’, in *Proceedings of the 13th Annual Conference on Innovative Data Systems Research (CIDR 2023)*.
- Janssen, N., Singh, P., Klievink, B. and van der Voort, H. (2024) ‘The evolution of data storage architectures: Examining the secure value of the data lakehouse’, *Journal of Data, Information and Management*, 6, pp. 315–334.
- Kahl, S., Wood, C.M., Eibl, M. and Klinck, H. (2021) ‘BirdNET: A deep learning solution for avian diversity monitoring’, *Ecological Informatics*, 61, article 101236.

- Kline, J., Potlapally, A., Pillai, B., Wani, T., Katole, R., Patil, V., Covey, P., Subramoni, H., Berger-Wolf, T. and Stewart, C. (2025) ‘SmartWilds: Multimodal wildlife monitoring dataset’, arXiv preprint, arXiv:2509.18894.
- Mazumdar, D., Hughes, J. and Onofre, J.B. (2023) ‘The data lakehouse: Data warehousing and more’, arXiv preprint, arXiv:2310.08697.
- Mourby, M., Gowans, H., Aidinlis, S., Smith, H. and Kaye, J. (2019) ‘Governance of academic research data under the GDPR—lessons from the UK’, *International Data Privacy Law*, 9(3), pp. 192–206.
- NASA FIRMS (n.d.) Fire Information for Resource Management System: Web services. Available at: <https://firms.modaps.eosdis.nasa.gov/web-services/> (Accessed: 3 September 2026).
- Office for National Statistics (n.d.) UK retail footfall. Available at: <https://www.ons.gov.uk/economy/economicoutputandproductivity/output/datasets/ukretailfootfall> (Accessed: 3 September 2026).
- Peffer, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007) ‘A design science research methodology for information systems research’, *Journal of Management Information Systems*, 24(3), pp. 45–77.
- Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M. and Wegmann, M. (2013) ‘Essential biodiversity variables’, *Science*, 339(6117), pp. 277–278.
- Rajabifard, A. and Williamson, I.P. (2001) ‘Spatial data infrastructures: Concept, SDI hierarchy and future directions’, in *Proceedings of the Geomatics '80 Conference*, Tehran, Iran.
- Saeedan, M. and Eldawy, A. (2022) ‘Spatial Parquet: A column file format for geospatial data lakes’, in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. New York: Association for Computing Machinery.
- Schneider, J., Gröger, C., Schwarz, H. and Mitschang, B. (2024) ‘The lakehouse: State of the art on concepts and technologies’, *SN Computer Science*, 5, article 449.
- Schneider, J., Gröger, C., Lutsch, A., Schwarz, H. and Mitschang, B. (2025) ‘Architectures and implementations of data lakehouses: Case studies from industrial practice’, in Węcel, K. (ed.)

Business Information Systems. BIS 2025. Lecture Notes in Business Information Processing, volume 554. Cham: Springer, pp. 241–255.

Stafford, G. (n.d.) Environmental Sensor Telemetry Data. Kaggle. Available at: <https://www.kaggle.com/datasets/garystafford/environmental-sensor-data-132k> (Accessed: 3 September 2026).

Stephenson, P.J. and Stengel, C. (2020) ‘An inventory of biodiversity data sources for conservation monitoring’, *PLOS ONE*, 15(12), article e0242923.

Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., Vercauteren, K.C., Snow, N.P., Halseth, J.M., Di Salvo, P.A., Lewis, J.S., White, M.D., Teton, B., Beasley, J.C., Schlichting, P.E., Boughton, R.K., Wight, B., Newkirk, E.S., Ivan, J.S., Odell, E.A., Brook, R.K., Lukacs, P.M., Moeller, A.K., Mandeville, E.G., Clune, J. and Miller, R.S. (2019) ‘Machine learning to classify animal species in camera trap images: Applications in ecology’, *Methods in Ecology and Evolution*, 10(4), pp. 585–590.

The Wildlife Trusts (n.d.) About us. Available at: <https://www.wildlifetrusts.org/> (Accessed: 22 August 2026).

Underwood, E., Taylor, K. and Tucker, G. (2018) ‘The use of biodiversity data in spatial planning and impact assessment in Europe’, *Research Ideas and Outcomes*, 4, article e28045.

Vargas-Solar, G., Darmont, J., Adorjan, A., Espinosa-Oviedo, J.A., Hara, C., Loudcher, S., Motz, R., Musicante, M. and Zechinelli-Martini, J.-L. (2024) ‘Dataversifying natural sciences: Pioneering a data lake architecture for curated data-centric experiments in life and earth sciences’, in *Proceedings of the 8th International Workshop on Data Analytics Solutions for Real-Life Applications (DARLI-AP 2024)*, Paestum, Italy.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J. et al. (2016) ‘The FAIR guiding principles for scientific data management and stewardship’, *Scientific Data*, 3, article 160018.

Zhang, Y., Peng, B., Du, Y. and Su, J. (2023) ‘GeoLake: Bringing geospatial support to lakehouses’, *IEEE Access*, 11, pp. 143037–143049.

Zwirowicz-Rutkowska, A. and Michalik, A. (2016) ‘The use of spatial data infrastructure in environmental management: An evaluation of spatial data usability’, *Environmental Management*, 58, pp. 523–541.

Appendices

Appendix A. Full sample data catalogue.

Page	Subpage	Sample dataset	Prototype role	Major outputs
Overview	Overview	Cross-domain Gold summary records	Reserve-level operational overview	gold_environment_daily; gold_connectivity_daily; gold_hydrology_daily; gold_soundscape_daily; gold_wildfire_daily_risk; gold_footfall_site_type; gold_wildlife_trust_sites; gold_sensor_coverage
Biodiversity	Soundscape	Soundscape detections and deployment metadata	Acoustic biodiversity evidence	gold_soundscape_daily; gold_soundscape_species_summary; gold_soundscape_device_summary; gold_soundscape_deployment_summary
Biodiversity	Moth Monitoring	Moth image records	Image-based biodiversity evidence	gold_moth_overview; gold_moth_species_summary; gold_moth_dataset_summary
Biodiversity	Field Surveys	Survey and catalogue records	Ecological survey evidence	gold_field_survey; gold_field_survey_catalogue; gold_field_survey_type_summary
Habitat & Environment	Environmental	IoT telemetry	Environmental monitoring input	gold_environment_daily; gold_data_quality_summary; gold_sensor_coverage
Habitat & Environment	Hydrology	Hydrology observations	Water monitoring input	gold_hydrology_daily; gold_hydrology_rainfall_latest; gold_hydrology_water_level_latest
Habitat & Environment	Geospatial	Site, trust, survey and asset records	Spatial context for reserve interpretation	gold_wildlife_trust_sites; gold_wildlife_trust_regions; gold_field_survey; gold_field_survey_catalogue; gold_geospatial_asset; gold_geospatial_integrated
Habitat & Environment	Satellite Feeds	Remote-sensing feed references	Conceptual remote-sensing extension view	No single confirmed Gold table; prototype remote-sensing / conceptual view

Habitat & Environment	Drone Imagery	Orthophoto and LiDAR-linked asset records	Geospatial extension for habitat imagery review	gold_geospatial_asset; gold_geospatial_integrated
Habitat & Environment	Fixed-Point Photography	Fixed-point asset and comparison records	Worked example for repeat photography and change interpretation	gold_fixed_point_assets; gold_fixed_point_comparison_pairs
Risks & Alerts	Wildfire Monitoring	Wildfire risk records	Risk monitoring and alert evidence	gold_wildfire_overall; gold_wildfire_daily_risk; gold_wildfire_map_points; gold_wildfire_season_latitude; gold_wildfire_sensor
Risks & Alerts	Flood & High Water	Hydrology and open-reading records	Flood and water-level interpretation	gold_hydrology_daily; gold_hydrology_rainfall_latest; gold_hydrology_water_level_latest
Site Management	Human Activity	Footfall records and activity proxy records	Human pressure proxy	gold_human_activity_proxy; gold_footfall_region; gold_footfall_site_type
Site Management	Grazing & Livestock	Grazing management records	Habitat management evidence	gold_grazing_livestock_overview; gold_grazing_livestock_farm_summary; gold_grazing_livestock_field_summary
Data & System Health	Data Quality	Validation and readiness records	Quality assurance and table-readiness monitoring	gold_data_quality_summary; inventory of published gold_% tables
Data & System Health	Data Lineage	Lineage and catalogue metadata	Data traceability and governance interpretation	Lineage / metadata view; not one single Gold output
Data & System Health	Connectivity	Connectivity measurements	Network performance assurance	gold_connectivity_daily
Innovation Lab	Digital Twin	Cross-domain scenario data	Experimental future-facing operational concept	Prototype innovation view; no single confirmed Gold table

Innovation Lab	Scenario Lab	Environmental and hydrology summaries	Scenario-based exploratory interpretation	gold_environment_daily; gold_hydrology_daily
Innovation Lab	Field AR	Environmental and hydrology context data	Experimental field-facing interpretation layer	gold_environment_daily; gold_hydrology_daily
Innovation Lab	AI Intelligence Brief	Cross-domain summary records	Experimental AI-assisted interpretation	Prototype intelligence view; cross-domain dashboard evidence
Innovation Lab	Prototype Story	Cross-domain dashboard evidence	Narrative explanation of the prototype design	gold_environment_daily; gold_hydrology_daily; cross-domain dashboard evidence

Appendix B. Suggested refresh frequencies by data domain.

Data domain	Suggested frequency	Rationale
Environmental IoT	Hourly	Sensor readings may change frequently.
Connectivity	Hourly	Network status benefits from recent updates.
Hydrology	Daily; hourly for live monitoring	Dashboard use may not require minute-level updates.
Soundscape metadata	Daily	Usually produced through batch audio analysis.
Wildfire risk	Hourly or daily	Higher frequency may be needed during fire season.
Human activity	Monthly	Source data is usually updated less frequently.
Moth / biodiversity	Weekly or monthly	Survey and classification results are often batch-based.
Geospatial / assets	On change or monthly	Location and asset records are relatively stable.
Gold export to PostgreSQL	After each successful Gold run	Keeps the serving layer aligned with processed outputs.

Appendix C. Prototype repository structure.

```
digital_reserve_lakehouse_prototype/
├── apps/
│   ├── api/           REST API and PostgreSQL data serving
│   ├── web/           Next.js prototype dashboard
│   └── ai/             Experimental AI extension
├── notebooks/         Databricks workflow notebooks
│   ├── 00_setup_volume.py
│   ├── 01_bronze.py
│   ├── 02_silver.py
│   ├── 03_gold.py
│   └── 04_export_gold_to_neon.py
├── src/               Python pipeline modules
│   ├── bronze.py
│   ├── silver.py
│   ├── gold.py
│   ├── audit.py
│   ├── governance.py
│   ├── table_registry.py
│   └── *_pipeline.py
├── scripts/           Utility scripts for upload and database preparation
├── config/            Environment configuration
├── data/              Local raw and processed data
├── docs/              Supporting project notes
├── README.md
├── requirements.txt
├── package.json
└── render.yaml
```