UNIFIED DATA LAKE ARCHITECTURE

A Unified Data Lake Architecture for Integrating Enterprise Content Platforms (OpenText, IBM FileNet, DocuWare) in Saudi Organizations

Mohammad Hussain

Published November 2026

Abstract

Large Saudi organisations often run more than one enterprise content management platform, so contracts, correspondence, scanned forms and records sit in separate repositories with different metadata, security and retention rules. This fragmentation limits search, analytics and compliance just as national digital transformation and data protection rules raise expectations. This paper proposes a unified data lake architecture that brings content from OpenText, IBM FileNet and DocuWare repositories into one governed analytical environment, while the source platforms remain the systems of record. The study follows a design science approach supported by an integrative review of 30 journal studies (2020–2025) on data lakes and lakehouses, metadata and semantic management, data catalogues, personal data handling, document understanding, records management and federated data architectures. Nine design requirements are derived, and the resulting architecture comprises platform connectors, four storage zones, a canonical content model with a knowledge graph, an enrichment service for text extraction and classification, a governance service that synchronises retention, legal holds and access rights with the source platforms, and an access layer for search and analytics. An analytical evaluation finds strong published support for zoned storage, metadata-driven management and lakehouse query performance, mixed support for automated enrichment and personal data discovery, and limited evidence on retention synchronisation and on the specific interfaces of the three platforms. The evaluation draws only on published studies; no system was built or tested, no organisational data were used, and product-specific claims require confirmation against vendor documentation.

1 Introduction

Enterprise content management (ECM) platforms were adopted to bring order to documents: to capture them, classify them, control who may see them and keep them for as long as rules require. In many Saudi ministries, banks, utilities and holding companies, the result is not one platform but several. A records department may use one product for correspondence and archives, a finance unit another for invoices, and a business unit a third for scanned forms. OpenText, IBM FileNet and DocuWare are three commercial platforms of this kind, each with its own repository model, metadata schema, security model and programming interfaces.

Each repository may work well on its own, yet the organisation as a whole cannot easily answer questions that cross them: which contracts with a supplier are about to expire, which records contain personal data, or how long a process takes from application to decision. Users also behave in ways that policies do not anticipate. Interviews with 35 knowledge workers in one research organisation revealed gaps between content management policy and the way people actually create, store and retrieve their work (Mathieu, 2022), and implementations of document management systems often

fail or are used only at a basic level, with users still printing documents (Sternad Zabukovšek et al., 2023).

Data lakes offer a way to bring heterogeneous content together. Unlike warehouses that impose a schema when data are written, they store data in original formats behind a common access interface (Hai et al., 2023), and the newer lakehouse combines this flexibility with the management features of warehouses (Harby and Zulkernine, 2025). Most data lake research, however, deals with structured records, logs and sensor streams. Little of it addresses governed business documents whose legal status, retention period and access rights are defined in an ECM platform. Research on ECM and records management, for its part, rarely considers analytical architectures. This paper links the two literatures and proposes a unified architecture for Saudi organisations that run several ECM platforms.

2 Research aim, objectives and questions

The study aims to design and analytically evaluate a data lake architecture that unifies content from multiple ECM platforms for search, analytics and compliance, without displacing those platforms as systems of record. Four objectives follow: (1) to characterise the integration problem created by multiple ECM platforms in Saudi organisations; (2) to derive design requirements from peer-reviewed research on data lakes, metadata, governance, document understanding and records management; (3) to design an architecture that satisfies those requirements for OpenText, IBM FileNet and DocuWare sources; and (4) to assess, criterion by criterion, how well published evidence supports the design.

Three questions follow from this aim. RQ1: What requirements must an architecture meet to integrate content from several ECM platforms while preserving their governance? RQ2: Which data lake mechanisms does the literature support for ingesting, describing, enriching and protecting such content? RQ3: How can these mechanisms be organised into a unified architecture, and where is the supporting evidence strong or weak?

3 Method: design science supported by an integrative review

Design science is appropriate because the goal is a purposeful artefact, an architecture, whose quality can be argued from requirements and evidence. The study moved through problem identification, requirement derivation, design and analytical evaluation. An integrative review supplied the evidence base, since the relevant research is spread across database systems, information systems, records management and document analysis.

Searches in November 2026 used an open catalogue of scholarly works, and each candidate was checked against its registration metadata. Query terms combined architecture concepts (data lake, lakehouse, data mesh, data fabric, data catalogue, metadata management, knowledge graph) with content concepts (enterprise content management, document management system, electronic records, optical character recognition, unstructured documents) and governance concepts (personal data, compliance, data quality, e-government). We included English-language journal articles, peer-reviewed, issued in final form between 2020 and 2025, whose abstracts described an architecture, a method, an empirical study or a structured review relevant to unifying governed content. Conference proceedings, editorials, short commentaries and papers without abstracts were left out.

Two anchor studies guided the review: the survey of data lake concepts and architectures by Azzabi et al. (2024) and the overview of data warehouses and data lakes in enterprise data management by Nambiar and Mundra (2022). The 30 retained articles include six on data lake and lakehouse architectures, five on ingestion and metadata, six on catalogues, privacy, quality and compliance, seven on ECM, records management and the Saudi context, two on document understanding and four on platforms and federated architectures.

Evaluation grades were assigned on a three-point scale. A criterion is graded supportive when at least two studies demonstrate a mechanism and report evidence of its effect, mixed when mechanisms are documented, but evidence is narrow or trade-offs are significant, and limited when the corpus provides little relevant evidence. Claims about each study were checked against its published abstract and text. We did not build, test or deploy any system and used no organisational data. No peer-reviewed study in the corpus evaluates OpenText, IBM FileNet or DocuWare directly, so statements about these platforms are confined to their general role as ECM repositories.

4 Problem context

Three features of the problem shape the design: the diversity of the sources, the governance that attaches to their content, and the regulatory setting in Saudi Arabia.

4.1 Multiple content platforms

Each ECM platform organises content in its own way. Documents carry platform-specific properties, classes or index fields; folders and cabinets impose different hierarchies; versions and renditions are handled differently; and access control lists refer to the platform's own users and groups. Where platforms were introduced by different units at different times, the same kind of document can be described in three incompatible ways. The literature explains why such estates persist. ECM investments often deliver limited value, and a study of seven large organisations found that value depended on how ECM use was aligned with agile business processes (Arshad et al., 2022). Maturity matters too: organisations that manage the life cycle of their document management systems cope better with digital transformation (Sternad Zabukovšek et al., 2023). Replacing all platforms with one is therefore rarely realistic, and integration has to work with the estate as it is.

4.2 Records and governance

Content in an ECM platform is often a record, with legal and evidential value, a retention schedule and rules for disposal. Records research shows that technology alone does not resolve these obligations. In South African national government departments, records practitioners and chief information officers reported that, despite cloud computing, government still struggled with paper-based records and lacked a cloud of its own for managing and disposing of them (Shibambu and Marutha, 2022). Adoption of electronic records management depends on technological, organisational and environmental factors; in the Yemeni oil and gas sector, experts confirmed 12 of 20 factors from the literature and added three more (Hawash et al., 2020). An analytical copy of records that ignored retention or disposal would create a second, ungoverned archive.

4.3 The Saudi setting and design requirements

In Saudi Arabia, ECM has been linked to transparency and accountability. A survey of 395 employees in the public financial sector validated a model in which system, information and service quality,

integration of artificial intelligence tools, management support, training, change management, policy and legislation, and social influence shaped the intention to adopt ECM in support of anti-corruption efforts (Alshammari et al., 2025). A review of Saudi e-government data security stressed that breaches of user data can have severe consequences and examined how well security policies deal with threats and vulnerabilities (Alharbi et al., 2021). Organisations must also comply with national personal data protection and data management rules, which require that personal data be identified, protected and retained only as long as necessary. Table 1 sets out the nine design requirements derived from this context and the literature, answering RQ1.

Table 1. Design requirements for unifying content from multiple ECM platforms.

No.	Requirement	Rationale	Main sources
R1	Unified ingestion from each platform, bulk and incremental	Platforms differ in models, metadata and interfaces	Sreepathy et al. (2024)
R2	Zoned storage with a governed serving layer	Raw copies must be kept apart from curated content	Azzabi et al. (2024); Harby and Zulkernine (2025)
R3	Canonical content model and dynamic metadata	The same document type is described differently in each platform	Boukraa et al. (2025); Yu et al. (2022)
R4	Semantic links and a searchable catalogue	Users must find related content across sources	Hoseini et al. (2024); Jahnke and Otto (2023)
R5	Text extraction and classification of scanned content	Much content exists only as images	Mahadevkar et al. (2024); Nguyen et al. (2022)
R6	Discovery and protection of personal data	National data protection rules apply	Oreščanin et al. (2024); Alharbi et al. (2021)
R7	Retention, legal hold and access synchronisation	The lake must not become an ungoverned archive	Shibambu and Marutha (2022); Sen and Jindal (2024)
R8	Lineage and quality documentation	Analytical results must be traceable to records	Eder and Shekhovtsov (2021)
R9	Scalable, cost-aware operation	Content volumes and query loads grow	Bauer et al. (2021); Kuschewski et al. (2023)

5 Evidence on data lake mechanisms

This section reviews the mechanisms available for each requirement, from architecture and ingestion through metadata, governance and enrichment to the choice between central and federated designs.

5.1 Data lake and lakehouse architectures

Reviews agree that data lakes have matured from simple repositories into layered architectures. Azzabi et al. (2024) classified the main types of data lake architecture proposed over time, drew them in a consistent notation and compared lakes with warehouses. Nambiar and Mundra (2022) described the definitions, characteristics and design considerations of both and set out open research directions. Hai et al. (2023) classified data lake systems by the functions they provide and noted that a coherent picture of challenges and solutions was still missing.

The lakehouse has become the leading variant. Harby and Zulkernine (2025) compared lake, warehouse and lakehouse systems, proposed an advanced lakehouse architecture and measured the ana-

lytical query performance of an HDFS data lake, a Hive warehouse and a Delta lakehouse. Combining a literature review with expert interviews, Janssen et al. (2024) built a model of the strengths and weaknesses of the three generations of storage architecture and showed where the lakehouse adds value. For heterogeneous institutional content et al. (2025) used design science to develop a lakehouse for research data in institutional repositories with five layers, covering collection, storage, processing, management and services, and refined it through interviews with professionals.

5.2 Ingestion and metadata

Ingestion from diverse sources is a known difficulty. Sreepathy et al. (2024) proposed a data ingestion service with unified connectors to databases, file systems, cloud storage and programming interfaces, a module that transforms structured, semi-structured and unstructured data, and a metadata module for search and enrichment; measured ingestion took about 148 microseconds per record for structured data and 234 microseconds per record for semi-structured data. Metadata generation can be automated: Yu et al. (2022) annotated raw industrial data as linked data, used online clustering and entity discovery from short texts, and built semantic metadata, validated on a real data set and an energy case.

Metadata models determine whether a lake remains usable or decays into a swamp. Boukraa et al. (2025) argued that static metadata models fail to cover governance, visibility and uniform treatment of concepts, and demonstrated a dynamic graph-based metadata system for exploring NoSQL data lakes. Yang et al. (2025) compared traditional and AI-driven metadata management across open-source tools, commercial products and research projects, and proposed a framework that uses AI to automate metadata generation and strengthen governance.

5.3 Semantic layers and data catalogues

Semantic data management links metadata to knowledge graphs so that data from heterogeneous sources can be integrated and queried meaningfully. Hoseini et al. (2024) classified such approaches into basic semantic management, semantic enrichment of metadata in data lakes and ontology-based data access, and argued that big data and semantic web technologies need to be integrated more closely. Catalogues make the resulting metadata usable. Jahnke and Otto (2023) distinguished seven classes of catalogue application, four used inside the enterprise and three across enterprises, and analysed 51 catalogue offerings, noting that supported sources, federation and access provisioning remain unclear. In an empirical comparison of three catalogue tools on 100 documents, Cherradi et al. (2023) reported accuracy, precision and recall of 96% for the best-performing tool in retrieving relevant information from a heterogeneous lake.

5.4 Personal data, quality and compliance

Personal data in content repositories pose a specific challenge. Oreščanin et al. (2024) found no standard data lake architecture for discovering personally identifiable information, matching records about the same person across sources or removing expired personal data, and proposed a metadata model to support these functions. Compliance can itself be made machine-readable: Joshi et al. (2020) built a knowledge graph of data protection regulations, threats and controls for cloud data and validated it against the privacy policies of several cloud providers.

Quality documentation is part of governance. Using design science et al. (2021) developed a conceptual model for federated medical data lakes that records metadata and quality annotations while observing strict privacy rules. Records management also has a place in modern data gov-

ernance: in an analysis of Indian data governance policy and an urban data exchange et al. (2024) concluded that digital records management must be integrated into data policy if public data are to be accurate, current and accessible.

5.5 Understanding unstructured documents

Much ECM content consists of scanned images and documents whose value lies in text. Mahadevkar et al. (2024) reviewed AI-based information extraction from unstructured documents and found that methods based on static patterns or rules often fail on complex layouts, that public data sets are of low quality and task-specific, and that AI techniques show promise for both printed and handwritten text. Recognition errors propagate: Nguyen et al. (2022) showed that noisy output from optical character recognition (OCR) harms information retrieval and natural language processing, and reviewed post-OCR correction methods, metrics and resources. Neither review concerns Arabic documents specifically, a gap relevant to Saudi archives.

5.6 Central and federated designs

Architectures differ in how much they centralise. In a review of 114 practitioner sources, Goedegebuure et al. (2025) characterised data mesh by four principles: data as a product, domain ownership, a self-serve platform and federated computational governance. Comparing 189 contributions et al. (2025) found that data platforms and fabrics offer centralised control, whereas meshes and data spaces prioritise flexibility. Operational experience points to practical constraints. Reporting three years of running a large analytics platform on a private cloud, Bauer et al. (2021) described how technical choices evolved with actual user workloads, and Kuschewski et al. (2023) showed that open columnar formats are not optimised for remotely accessed lakes and proposed a format with lightweight encodings for faster decompression. In answer to RQ2, the literature supports zoned lake or lakehouse storage, unified connectors, dynamic and semantic metadata, catalogues, explicit models for personal data and quality, AI-based enrichment with error correction, and a choice of centralisation that matches the organisation's governance.

6 Proposed unified architecture

Figure 1 shows the proposed architecture. Its central principle is that each ECM platform remains the system of record. The lake holds governed copies for search, analytics and compliance; it never becomes the authoritative store, and any change to a record's status in a source platform is reflected in the lake.

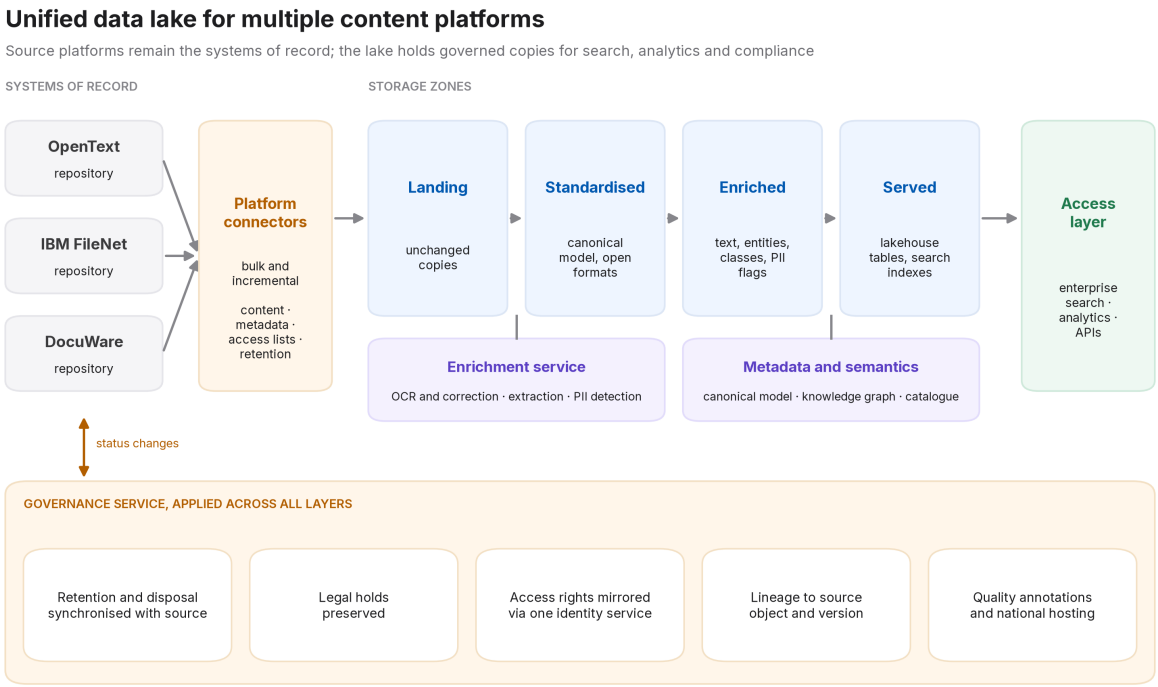


Figure 1. Unified data lake architecture for content from multiple ECM platforms.

6.1 Layers and services

Platform connectors form the first layer. For each of OpenText, IBM FileNet and DocuWare, a connector extracts content files, version information, metadata, classifications, access control entries and retention attributes, first in an initial bulk load and then incrementally as changes occur. Following the unified ingestion approach of Sreepathy et al. (2024), connectors share a common contract so that downstream services do not depend on any one product. The specific interfaces each platform exposes are not examined in the corpus and must be confirmed against vendor documentation and tested.

Storage is organised in four zones. The landing zone holds unchanged copies with their source metadata. The standardised zone maps metadata to a canonical content model and converts files to open, searchable formats. The enriched zone adds extracted text, entities, classifications and personal data flags. The served zone exposes lakehouse tables and search indexes. Zoning follows the layered architectures surveyed by Azzabi et al. (2024) and Hai et al. (2023), and the lakehouse serving layer reflects the evidence of Harby and Zulkernine (2025) and Janssen et al. (2024).

The metadata and semantic layer holds the canonical content model and a knowledge graph that links documents to parties, contracts, cases, records series and source locations. It uses the dynamic modelling of Boukraa et al. (2025) and the semantic enrichment approaches reviewed by Hoseini et al. (2024), and it feeds a catalogue of the kind classified by Jahnke and Otto (2023). The enrichment service performs OCR with post-correction (Nguyen et al., 2022), information extraction and classification (Mahadevkar et al., 2024), and personal data detection based on the metadata model of Oreščanin et al. (2024).

The governance service is what distinguishes the design from a generic data lake. It synchronises retention schedules, disposal decisions and legal holds from each source, so that content disposed of in a platform is removed from every zone, and content on hold is preserved. It mirrors source access

rights, mapping each platform's users and groups to a single identity service, so that the lake never shows a user content they could not see at the source. Lineage is recorded from every served item back to its source object and version, quality annotations are applied (Eder and Shekhovtsov, 2021), and data stay within approved national hosting. Compliance rules can be represented in a knowledge graph of regulations and controls (Joshi et al., 2020). The access layer offers enterprise search, analytical queries and programming interfaces, all subject to the same rights and audit.

6.2 Content journey

Figure 2 follows a single document. A contract scanned into one platform is extracted by its connector with its metadata and access list, stored unchanged in the landing zone, mapped to the canonical model, converted to text, enriched with parties, dates and personal data flags, and published to the search index and lakehouse tables. When its retention period ends and the source platform disposes of it, the governance service removes every copy and derived item and records the action. For organisations whose departments own distinct content domains, the same components can be operated in a federated way, with domain teams publishing content products under shared governance, as data mesh principles suggest (Goedegebuure et al., 2025). In answer to RQ3, the design organises the supported mechanisms into connectors, zones, a semantic layer, enrichment, governance and access, with records authority retained by the source platforms.

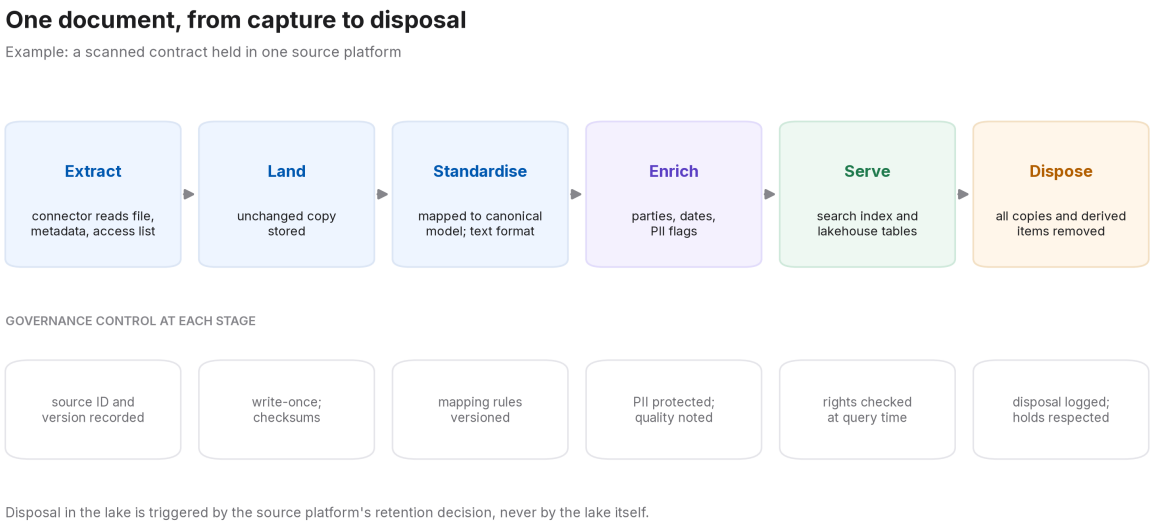


Figure 2. Journey of a document through the storage zones, with governance controls applied at each stage.

7 Analytical evaluation

Table 2 assesses the architecture against the nine requirements of Table 1 and, separately, against the need for product-specific integration. Each grade reflects how much the literature backs a design choice; none reflects observed system behaviour.

Table 2. Analytical evaluation of the unified architecture against the design requirements and product-specific integration.

Criterion	Design response	Evidence rating	Residual risk
R2 Zoned storage	Landing, standardised, enriched and served zones; lakehouse tables	Supportive	Storage duplication and cost
R3 Metadata	Canonical model; dynamic graph-based metadata	Supportive	Mapping errors between platforms
R4 Discoverability	Knowledge graph and enterprise catalogue	Supportive	Catalogue upkeep effort
R1 Ingestion	Common connector contract for all three platforms	Mixed	Not shown for ECM sources
R5 Enrichment	OCR with post-correction; extraction and classification	Mixed	Arabic and handwritten text untested
R6 Personal data	Detection flags and metadata model for personal data	Mixed	No standard architecture
R9 Scale and cost	Open columnar formats; workload-driven tuning	Mixed	Document-heavy loads unstudied
R7 Retention and access	Governance service mirroring source decisions	Limited	Inconsistency creates legal exposure
R8 Lineage and quality	Lineage to source object and version; quality annotations	Mixed	Quality models only conceptual
Platform-specific integration	Connectors checked against vendor interfaces	Limited	No peer-reviewed product evidence

Three criteria are graded supportive. Zoned storage and the lakehouse serving layer are backed by several surveys and an experimental comparison (Azzabi et al., 2024; Hai et al., 2023; Harby and Zulkernine, 2025). Metadata-driven management, including dynamic and semantic models, is supported by working systems and evaluations (Boukraa et al., 2025; Hoseini et al., 2024; Yu et al., 2022). Discoverability through catalogues has both a typology and an empirical comparison with high retrieval accuracy (Cherradi et al., 2023; Jahnke and Otto, 2023).

Five criteria are graded mixed. Unified ingestion has been demonstrated, but with general sources rather than ECM platforms (Sreepathy et al., 2024). Automated enrichment is promising but sensitive to layouts, recognition errors and the lack of suitable training data (Mahadevkar et al., 2024; Nguyen et al., 2022). Personal data handling has a proposed metadata model but no standard architecture (Oreščanin et al., 2024). Scalable operation is informed by operational reports and storage format research (Bauer et al., 2021; Kuschewski et al., 2023), but not for document-heavy workloads. Lineage and quality documentation rest on a conceptual model developed for medical data lakes (Eder and Shekhovtsov, 2021) rather than on evaluated implementations.

Two criteria are graded limited. One is retention and access synchronisation with the source platforms. Records research establishes why disposal and authority matter (Sen and Jindal, 2024; Shibambu and Marutha, 2022), but no study in the corpus evaluates mechanisms for keeping a lake consistent with the retention and access decisions of ECM platforms. The other is product-specific integration with OpenText, IBM FileNet and DocuWare, for which the corpus contains no peer-reviewed evidence.

8 Discussion, implications and limitations

The design shows that most of the technical components needed to unify ECM content already exist in data lake research. What is missing is the bridge between analytical data management and records governance. Data lake studies treat governance mainly as metadata, quality and privacy; records studies treat it as authority, retention and disposal. The architecture brings the two together by making the governance service, rather than the storage layer, the centre of the design.

For Saudi organisations, three implications follow. First, integration should begin with a canonical content model agreed between records managers and data teams, since the mapping of metadata from each platform determines the value of everything downstream. Second, retention, legal hold and access synchronisation should be tested before any content is exposed, because errors there create legal rather than merely technical risks. Third, enrichment of Arabic and bilingual documents requires local evaluation, since the reviewed studies give no evidence for Arabic text. Adoption research suggests that training, change management and management support will matter as much as technology (Alshammari et al., 2025).

Researchers could usefully study how analytical copies of records are governed in practice, evaluate enrichment on Arabic administrative documents and compare central and federated operating models in organisations with several ECM platforms.

The limitations are significant. The review covers 30 journal articles chosen by a single researcher and omits conference literature and vendor documentation, both of which contain relevant technical detail. Statements about OpenText, IBM FileNet and DocuWare are limited to their general role, and connector designs remain assumptions until checked against each product's interfaces. No organisation was studied, no system was built or measured, and no data were collected, so the evaluation indicates the strength of published support rather than proven performance. Saudi evidence in the corpus is limited to adoption and security studies.

9 Conclusion

Organisations that operate OpenText, IBM FileNet and DocuWare side by side hold valuable content that cannot be searched, analysed or governed as a whole. This paper proposed a unified data lake architecture in which platform connectors feed four storage zones, a canonical model and knowledge graph describe the content, an enrichment service extracts its meaning, and a governance service keeps retention, legal holds and access rights aligned with the source platforms. Published evidence strongly supports the storage, metadata and catalogue components, supports ingestion, enrichment, personal data handling and scaling with qualifications, and is weak on retention synchronisation and product-specific integration. Building and testing the architecture in a Saudi organisation, with Arabic content and real retention rules, is the next step.

References

[1] Alharbi AS, Halikias G, Rajarajan M, Yamin M (2021). A review of effectiveness of Saudi e-government data security management. *International Journal of Information Technology*, 13(2), 573–579. https://doi.org/10.1007/s41870-021-00611-3

[2] Alshammari BO, Asma' Mokhtar U, Binti Abdul Manap N (2025). A model for adopting an enterprise content management system for anti-corruption efforts. *IEEE Access*, 13, 195031–195049. https://doi.org/10.1109/access.2025.3633072

[3] Arshad NI, Bosua R, Milton S, Mahmood AK, Zainal-Abidin AI, Ariffin MM, et al (2022). A sustainable enterprise content management technologies use framework supporting agile business processes. *Knowledge Management Research & Practice*, 20(1), 123–140. https://doi.org/10.1080/14778238.2021.1973352

[4] Azzabi S, Alfughi Z, Ouda A (2024). Data lakes: A survey of concepts and architectures. *Computers*, 13(7), Article 183. https://doi.org/10.3390/computers13070183

[5] Bauer D, Froese F, Garcés-Erice L, Giblin C, Labbi A, Nagy ZA, et al (2021). Building and operating a large-scale enterprise data analytics platform. *Big Data Research*, 23, Article 100181. https://doi.org/10.1016/j.bdr.2020.100181

[6] Boukraa D, Bouraoui M, Grine C, Ouahab R (2025). Megale: A metadata-driven graph-based system for data lake exploration. *International Journal of Information Technology & Decision Making*, 24(01), 259–295. https://doi.org/10.1142/s0219622024500135

[7] Cherradi M, Bouhafer F, EL Haddadi A (2023). Data lake governance using IBM-Watson Knowledge Catalog. *Scientific African*, 21, Article e01854. https://doi.org/10.1016/j.sciaf.2023.e01854

[8] Eder J, Shekhovtsov VA (2021). Data quality for federated medical data lakes. *International Journal of Web Information Systems*, 17(5), 407–426. https://doi.org/10.1108/ijwis-03-2021-0026

[9] Gieß A, Hutterer A (2025). The future of data management: A delimitation of data platforms, data spaces, data meshes, and data fabrics. *Information Systems and e-Business Management*, 23(4), 971–997. https://doi.org/10.1007/s10257-025-00707-4

[10] Goedegebuure A, Kumara I, Driessen S, Van Den Heuvel WJ, Monsieur G, Tamburri DA, et al (2025). Data mesh: A systematic gray literature review. *ACM Computing Surveys*, 57(1), 1–36. https://doi.org/10.1145/3687301

[11] Hai R, Koutras C, Quix C, Jarke M (2023). Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 12571–12590. https://doi.org/10.1109/tkde.2023.3270101

[12] Harby AA, Zulkernine F (2025). Data lakehouse: A survey and experimental study. *Information Systems*, 127, Article 102460. https://doi.org/10.1016/j.is.2024.102460

[13] Hawash B, Mokhtar UA, Yusof ZM, Mukred M (2020). The adoption of electronic records management system (ERMS) in the Yemeni oil and gas sector. *Records Management Journal*, 30(1), 1–22. https://doi.org/10.1108/rmj-03-2019-0010

[14] He Z, Fang W (2025). Research data management in institutional repositories: An architectural approach using data lakehouses. *Digital Library Perspectives*, 41(1), 145–178. https://doi.org/10.1108/dlp-02-2024-0022

[15] Hoseini S, Theissen-Lipp J, Quix C (2024). A survey on semantic data management as intersection of ontology-based data access, semantic modeling and data lakes. *Journal of Web Semantics*, 81, Article 100819. https://doi.org/10.1016/j.websem.2024.100819

[16] Jahnke N, Otto B (2023). Data catalogs in the enterprise: Applications and integration. *Datenbank-Spektrum*, 23(2), 89–96. https://doi.org/10.1007/s13222-023-00445-2

[17] Janssen N, Ilayperuma T, Jayasinghe J, Bukhsh F, Daneva M (2024). The evolution of data storage architectures: Examining the secure value of the data lakehouse. *Journal of Data, Information and Management*, 6(4), 309–334. https://doi.org/10.1007/s42488-024-00132-1

[18] Joshi KP, Elluri L, Nagar A (2020). An integrated knowledge graph to automate cloud data compliance. *IEEE Access*, 8, 148541–148555. https://doi.org/10.1109/access.2020.3008964

[19] Kuschewski M, Sauerwein D, Alhomssi A, Leis V (2023). BtrBlocks: Efficient columnar compression for data lakes. *Proceedings of the ACM on Management of Data*, 1(2), 1–26. https://doi.org/10.1145/3589263

[20] Mahadevkar SV, Patil S, Kotecha K, Soong LW, Choudhury T (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, 11(1), Article 92. https://doi.org/10.1186/s40537-024-00948-z

[21] Mathieu C (2022). Defining knowledge workers' creation, description, and storage practices as impact on enterprise content management strategy. *Journal of the Association for Information Science and Technology*, 73(3), 472–484. https://doi.org/10.1002/asi.24563

[22] Nambiar A, Mundra D (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*, 6(4), Article 132. https://doi.org/10.3390/bdcc6040132

[23] Nguyen TTH, Jatowt A, Coustaty M, Doucet A (2022). Survey of post-OCR processing approaches. *ACM Computing Surveys*, 54(6), 1–37. https://doi.org/10.1145/3453476

[24] Oreščanin D, Hlupić T, Vrdoljak B (2024). Managing personal identifiable information in data lakes. *IEEE Access*, 12, 32164–32180. https://doi.org/10.1109/access.2024.3365042

[25] Sen S, Jindal T (2024). What does a record have to do with it?. *Re-situating records management within Indian public data governance and policy*. Data & Policy, 6, Article e38. https://doi.org/10.1017/dap.2024.30

[26] Shibambu A, Marutha NS (2022). A framework for management of digital records on the cloud in the public sector of South Africa. *Information Discovery and Delivery*, 50(2), 165–175. https://doi.org/10.1108/idd-10-2020-0128

[27] Sreepathy HV, Dinesh Rao B, Kumar Jaysubramanian M, Deepak Rao B (2024). Data ingestions as a service (DIaaS): A unified interface for heterogeneous data ingestion, transformation, and metadata management for data lake. *IEEE Access*, 12, 156131–156145. https://doi.org/10.1109/access.2024.3479736

[28] Sternad Zabukovšek S, Jordan S, Bobek S (2023). Managing document management systems' life cycle in relation to an organization's maturity for digital transformation. *Sustainability*, 15(21), Article 15212. https://doi.org/10.3390/su152115212

[29] Yang W, Fu R, Amin MB, Kang B (2025). The impact of modern AI in metadata management. *Human-Centric Intelligent Systems*, 5(3), 323–350. https://doi.org/10.1007/s44230-025-00106-5

[30] Yu H, Cai H, Liu Z, Xu B, Jiang L (2022). An automated metadata generation method for data lake of industrial WoT applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(8), 5235–5248. https://doi.org/10.1109/tsmc.2021.3119871