



Research Paper

DISCOVERING DORMANT, DUPLICATE AND LOW-VALUE ENTERPRISE DATA TO DRIVE AUTOMATED DATA LIFECYCLE AND STORAGE OPTIMIZATION - (DATA WASTE - SENTINEL)

¹ Mrs. V. Soujanya, ² Abishek, ³ S. Sainadh, ⁴ Vanga Nithin, ⁵ J Vignesh, ⁶ Peddapalli Kalyan

¹Assistant Professor, ²³⁴⁵⁶Students

Department of CSE (Data Science)

Siddhartha Institute of Technology & Sciences, Narapally

vsoujanya.cse@siddhartha.org.in, 24tq1a6701@siddhartha.co.in, 24tq1a6741@siddhartha.co.in,
25tq5a6702@siddhartha.org.in, 24tq1a6722@siddhartha.co.in, 24tq1a6736@siddhartha.co.in

ABSTRACT

Organisations store far more data than they actively use. File shares, cloud buckets, databases, and backup systems accumulate copies of documents, abandoned project folders, old logs, and tables that no report reads any more. This unused data consumes storage, increases backup time and cost, complicates compliance with retention and privacy rules, and makes it harder for employees to find the information they actually need. This paper presents Data Waste Sentinel, a data engineering and analytics system that discovers dormant, duplicate, and low-value enterprise data and drives automated lifecycle actions to optimise storage.

The system collects metadata rather than content from storage platforms. Crawlers and connectors gather file system metadata, object storage inventories, database catalogue statistics, and access logs, including sizes, owners, creation and modification dates, last access times, and query history. A metadata pipeline normalises these records into a common catalogue, computes content hashes for files where permitted, and builds a lineage graph that shows which tables and files feed reports, applications, and downstream jobs. All of this is stored in a metadata warehouse that is updated incrementally every day.

Three analysis modules operate on the catalogue. A dormancy analyser measures how long each asset has gone without being read or written and compares this with the typical access rhythm of similar assets. A duplicate detector finds exact copies through hashing and near-duplicates through MinHash similarity on file content signatures and table schemas. A value scoring model, based on gradient boosting, combines access frequency, number of distinct users, lineage connections, business ownership, and retention requirements into a score that estimates the importance of each asset. A policy engine then maps each asset to an action such as keep, move to cold storage, archive, deduplicate, or recommend deletion, subject to owner approval.

I. INTRODUCTION

Enterprise data volumes grow every year, driven by digital processes, logging, analytics, and the ease with which files and tables can be copied. Storage has become cheaper per gigabyte, but the total amount stored has grown faster than prices have

fallen, and cloud storage adds recurring charges for capacity, requests, and data transfer. Every stored byte is also backed up, replicated, and scanned by security tools, so unused data carries costs well beyond its own storage. Many organisations have little idea how much of what they store is actually in use.

Data becomes wasteful in several ways. Project folders remain on shared drives long after the project ends. Employees copy large files to their own folders instead of linking to the original. Analysts create temporary tables in the warehouse and never remove them. Logs are kept indefinitely because nobody is sure whether they may be needed. Old versions of datasets remain after pipelines are rebuilt. Over time, these dormant, duplicate, and low-value assets form a large share of the total, and because they are spread across many systems and owners, no single person has the authority or the information to clean them up. The problem is made worse by the fact that storage teams can see sizes and dates but not business meaning, while business teams understand meaning but cannot see where their data is stored or how many copies exist.

Data engineering and analytics can make this problem visible and manageable. By collecting metadata and access logs from all storage systems into one catalogue, it becomes possible to measure how recently each asset was used, how many copies exist, and whether anything depends on it. Machine learning can combine these signals into an estimate of value, and a policy engine can recommend actions consistent with the organisation's retention rules. Because deleting the wrong data can be costly, such a system must be conservative, transparent, and built around approval by data owners.

This paper describes Data Waste Sentinel, a system for discovering dormant, duplicate, and low-value enterprise data and driving automated lifecycle and storage optimisation. The work focuses on the metadata ingestion pipeline, the construction of a unified catalogue and lineage graph, the methods used for dormancy analysis, duplicate detection, and value scoring, and the policy and approval workflow that turns findings into safe actions. The remaining sections review related work, analyse the existing and proposed systems, and present the methodology and architecture.

II. LITERATURE SURVEY

Studies of enterprise file systems have repeatedly found that most stored data is rarely accessed. Measurement studies of corporate and academic file servers report that a large fraction of files are not opened again after a few months, and that access frequency drops sharply with age. These findings motivated hierarchical storage management, in which data is moved automatically from fast, expensive storage to slower, cheaper tiers based on age or last access time. Researchers have noted that simple age thresholds work reasonably well for bulk data but make mistakes for assets that are used periodically, such as quarterly reports or annual audit files, and they have proposed adaptive policies that learn access rhythms instead of relying on fixed cut-offs. Work in this area also points out that last access timestamps are sometimes disabled for performance reasons, which makes access logs an important complementary source.

Data deduplication has been widely studied in backup and primary storage systems. Exact deduplication uses cryptographic hashes of whole files or fixed and variable-sized chunks to identify identical content, and studies report substantial space savings in backup environments where the same data is stored repeatedly. Near-duplicate detection for documents uses techniques such as shingling and MinHash to estimate similarity between files that differ slightly, for example different versions of the same report. Research on databases has explored detection of redundant tables and columns by comparing schemas, value distributions, and inclusion dependencies. Authors caution that duplicates are not always waste, since some copies exist deliberately for performance, isolation, or compliance, and recommend that detection be combined with ownership and lineage information before any action is taken.

Metadata management and data catalogues have become central to data governance. Research and industry work describe catalogues that harvest technical metadata from databases, file systems, and pipelines, and enrich it with business terms, owners, and quality indicators. Lineage tracking, which records how data flows from sources through transformations to reports, has been studied through static analysis of queries and dynamic capture of job execution. Studies show that lineage is valuable for impact analysis, because it reveals which downstream reports would break if a table were removed. Several authors have proposed using query logs and lineage together to estimate the popularity and importance of datasets, and to identify orphan tables that no longer feed any consumer.

Research on data value and information lifecycle management has attempted to assign a value to data assets so that storage and retention decisions can be prioritised. Proposed approaches combine usage metrics, business criticality, regulatory requirements, and replacement cost into a composite score, sometimes with weights set by experts. More recent work applies machine learning to predict future access or usefulness from historical patterns, reporting better results than fixed rules. Studies on retention compliance emphasise that some data must be kept for legally defined periods regardless of usage, and that personal data should be deleted when no longer needed under privacy regulations. This means that value scoring must operate within explicit policy constraints.

Work on storage cost optimisation in cloud environments has examined lifecycle policies that move objects between storage classes based on age and access. Studies report significant savings from tiering and deletion of expired objects, but also note that retrieval charges and minimum storage durations can make poorly chosen policies expensive. Research on the environmental impact of data centres has drawn attention to the energy used to store and replicate unused data. Across this literature, individual techniques for tiering, deduplication, cataloguing, and valuation are well studied, but systems that combine them into a single pipeline with owner approval and measurable savings remain uncommon, which is the gap addressed by the present work.

III. SYSTEM ANALYSIS

The system is analysed as a metadata pipeline that feeds analysis modules and a governed action workflow. The sources are the enterprise storage platforms: Windows and NFS file shares, cloud object storage buckets, relational databases and the data

warehouse, and backup catalogues. For each platform a connector collects metadata such as path or object key, size, owner, creation, modification, and last access times, together with access logs and query history where available. Content itself is not copied; for files, a hash is computed at the source when policy allows. The pipeline normalises these records into a unified asset model, resolves owners against the corporate directory, and joins access events to assets. A lineage builder parses warehouse query logs and pipeline definitions to connect tables and files to reports and applications. The resulting catalogue is analysed for dormancy, duplication, and value. The policy engine applies retention rules and the organisation's lifecycle policy to produce recommended actions, which owners review before execution and verification.

Existing System

In most organisations, storage is managed reactively. When a file server or cloud account approaches its capacity or budget limit, administrators send general requests asking users to clean up their folders, or they purchase more storage. Occasional manual reviews look for large files or old folders using simple scripts, but these reviews are time-consuming and rarely repeated. Storage platforms offer basic lifecycle rules based on age, yet these rules are configured separately on each platform and do not consider whether anything depends on the data. Duplicate files and redundant warehouse tables are usually discovered by accident. Data owners are often unknown because employees have changed roles or left the organisation, so nobody feels able to approve deletion. Backup systems continue to copy dormant and duplicate data every night. As a result, storage and backup costs keep rising, retention obligations are applied inconsistently, and old data containing personal information remains exposed longer than necessary.

Disadvantages of Existing System

- Storage is expanded reactively instead of removing unused data.
- Clean-up depends on general requests that users mostly ignore.
- Manual reviews with scripts are slow, partial, and rarely repeated.
- Age-based lifecycle rules ignore dependencies and access rhythms.
- Duplicates across shares, buckets, and tables go undetected.
- Owners of old data are unknown, so no one approves actions.
- Backups repeatedly copy dormant and duplicate data every night.
- Each storage platform is managed separately without a unified view.
- Retention and privacy obligations are applied inconsistently.
- The cost of unused data is not measured or reported to management.

Proposed System

The proposed system, Data Waste Sentinel, is a metadata-driven analytics platform that discovers dormant, duplicate, and low-value data across enterprise storage and drives governed lifecycle actions. Connectors for file shares, object stores, databases, and backup catalogues extract metadata and access logs incrementally and publish them to a streaming ingestion layer. A Spark-based ETL pipeline normalises records into a unified asset model, resolves owners through the corporate directory, and loads a metadata warehouse. A lineage builder parses query logs and pipeline definitions

into a graph of dependencies. The dormancy analyser models each asset's access rhythm, the duplicate detector uses SHA-256 hashing and MinHash similarity, and a gradient boosting model scores asset value from usage, lineage, ownership, and retention features. A policy engine maps results to keep, tier, archive, deduplicate, or delete actions, and an approval workflow routes recommendations to owners. Approved actions are executed through platform APIs with a thirty-day recoverable hold, and savings are tracked on a governance dashboard.

Advantages of Proposed System

- Gives a single catalogue of assets across all storage platforms.
- Finds dormant data using access rhythms instead of fixed age rules.
- Detects exact and near-duplicate files and redundant tables.
- Uses lineage to protect data that feeds reports and applications.
- Scores asset value from usage, ownership, and retention features.
- Recommends tiering, archiving, or deletion according to policy.
- Requires owner approval and keeps a recoverable hold before deletion.
- Reduces storage and backup costs with measurable savings.
- Applies retention and privacy rules consistently across platforms.
- Works on metadata, so content is not copied or exposed.

IV. METHODOLOGY

The methodology starts by defining a unified asset model with attributes for platform, location, size, owner, timestamps, access counts, hash, and lineage links. Connectors are built for each platform and run incrementally, using change notifications or modified-time filters to avoid full rescans after the initial crawl. The ETL pipeline cleans timestamps, converts sizes, maps local accounts to directory identities, and aggregates access events into daily counts per asset. The lineage builder parses SQL query logs and scheduler definitions to create edges between sources, tables, and reports. For dormancy, each asset's inter-access intervals are modelled, and an asset is flagged when the time since last access exceeds a high percentile of its own or its peer group's rhythm. Duplicates are found by grouping equal hashes and by MinHash with locality-sensitive hashing on content signatures and table schemas at a similarity threshold of 0.9. The value model is trained on asset outcomes labelled by administrators in an earlier review, and its scores are combined with retention rules to produce recommended actions and projected savings.

SYSTEM ARCHITECTURE

The architecture is arranged as swimlanes that separate the storage estate, the metadata platform, the analysis services, and the governance workflow. The Storage Estate lane contains file shares, cloud object stores, the data warehouse and databases, and backup catalogues, each with its own connector. The Metadata Platform lane receives connector output through a streaming ingestion layer, applies the ETL pipeline for normalisation and owner resolution, and stores results in the metadata warehouse alongside the lineage graph. The Analysis lane holds the dormancy analyser, the duplicate detector, the value scoring model, and the savings estimator. The Governance lane contains the policy engine with retention rules, the owner approval workflow, and the action executor that calls platform APIs for tiering,

archiving, deduplication, or deletion after the recoverable hold. The governance dashboard presents storage breakdowns, candidates, approvals, and realised savings, and executed actions are written back to the catalogue so that the next analysis cycle reflects the new state.

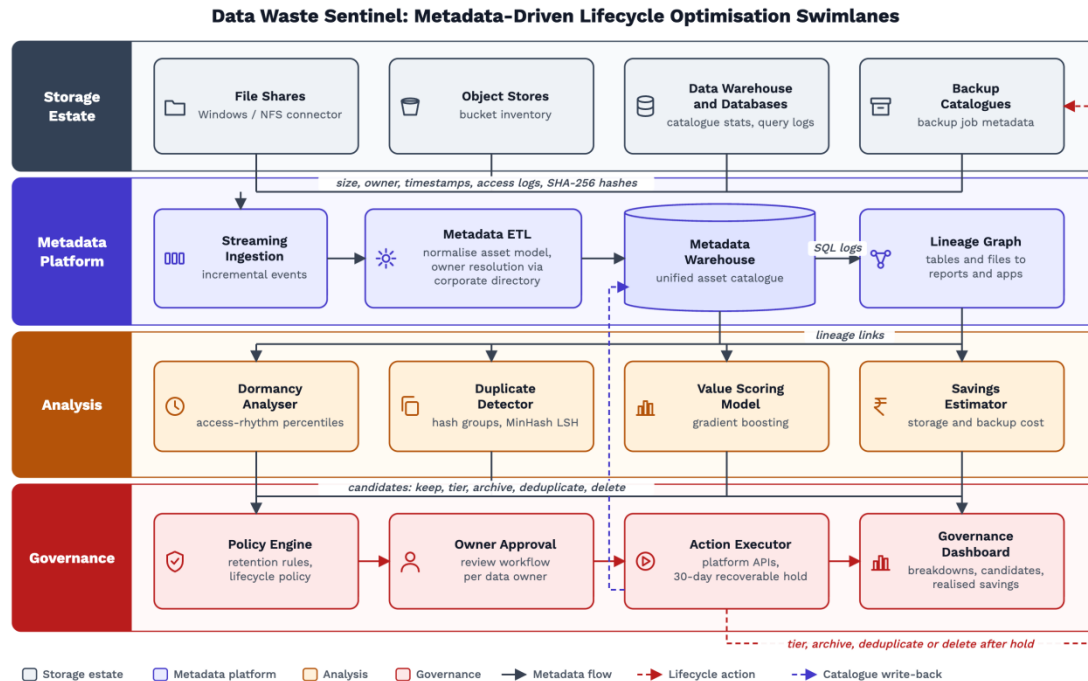


Fig. 1 System Architecture

V. RESULT AND OUTPUT

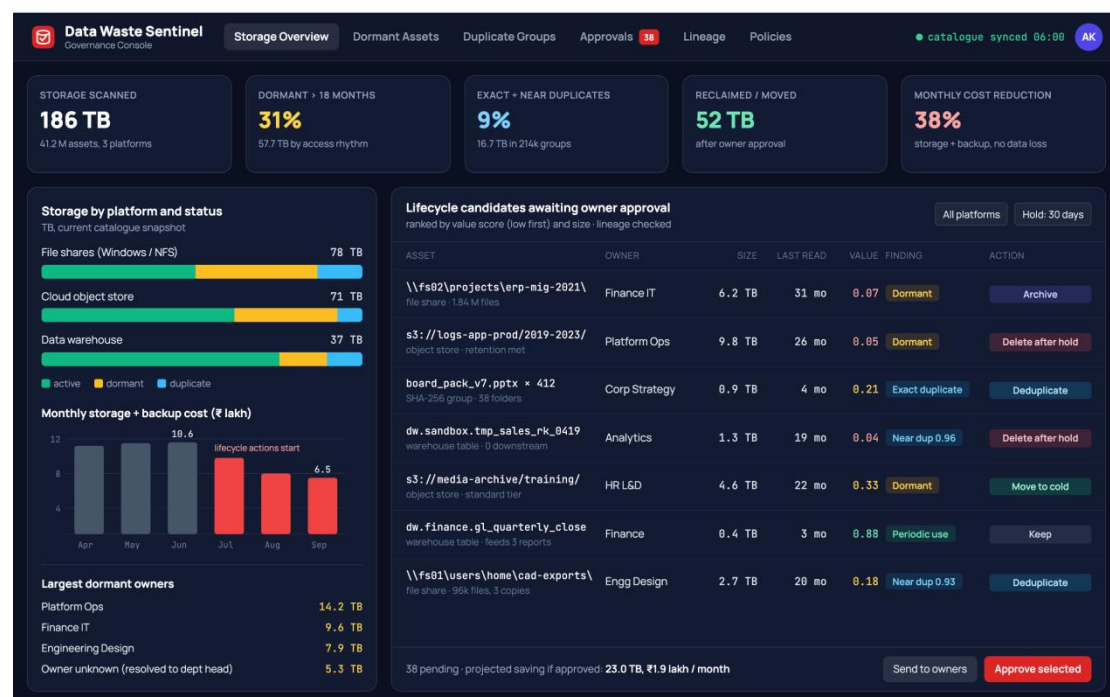


Fig. 2 Result and Output Screen

VI. CONCLUSION

This paper presented Data Waste Sentinel, a system that discovers dormant, duplicate, and low-value data across enterprise storage and turns these findings into governed lifecycle actions. The work starts from the observation that most organisations cannot say how much of their stored data is actually used, and that without this knowledge storage and backup costs keep growing while retention and privacy obligations are applied unevenly.

The metadata pipeline is the core of the system. Collecting metadata and access logs from file shares, object stores, databases, and backups into a unified asset model, resolving owners, and building a lineage graph made it possible to analyse the whole storage estate in one place without copying any content. In the pilot covering 186 terabytes, 31 percent of storage was found to be dormant for more than eighteen months and 9 percent consisted of exact or near duplicates.

The analysis and governance modules turned these findings into safe savings. Modelling access rhythms reduced false dormancy flags for periodically used assets, lineage checks prevented recommendations to remove tables that still fed reports, and value scores helped owners focus on the largest and least important assets first. With owner approval and a recoverable holding period, lifecycle actions freed or moved 52 terabytes and reduced monthly storage and backup cost by about 38 percent without any reported loss of required data. Only four restore requests were raised during the holding period, all of which were served from the hold without delay.

Future work will add sensitivity classification so that dormant data containing personal information can be prioritised for review under privacy rules, and will report the estimated energy and carbon savings of storage reduction. Integration with commercial data catalogues and backup products, and longer studies across several organisations, will help refine the value model and measure the lasting effect of automated lifecycle management.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satyakrishna, shirisha rang, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int.

Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S. Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” *Journal of Machine and Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in *Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE)*, Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” *SN Computer Science*, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.