International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 2, 234-243, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I2P127

Original Article

ML-Powered Data Catalogs and Metadata Inference for Self-Describing Big Data Ecosystems

Sivadeep Katangoori
Solutions Architect at Metanoia Solutions Inc., USA.

Received On: 26/04/2025 ***Revised On: 06/05/2025*** ***Accepted On: 21/05/2025*** ***Published on: 11/06/2025***

Abstract - *In today's fast-changing data world, businesses have too much data that is too different and spread out. Standard methods for managing data don't always work well for finding, understanding, and applying data. Modern data catalogues are more than simply static lists of items; they are also dynamic places to store information that provide structure, meaning, and easy access to large data ecosystems. These catalogues can really shine when ML is applied to figure things out smartly, even when the data sets are too messy, not well-documented, or not well-organized. Data catalogues that utilize ML look at these current data trends, user behaviors & contextual signals to automatically tag, categorize & link their data assets. This makes the data self-descriptive and makes governance easier. This solves a big problem: creating information by hand is hard work, prone to their mistakes, and typically doesn't work well with different teams & also technologies. Our methodology utilizes both supervised and unsupervised ML techniques to infer schema details, data lineage, semantic links, and use contexts, therefore reducing dependence on human input while improving their data discoverability & the dependability. This paper suggests a scalable architecture that integrates ML models directly into the cataloguing process, enabling actual time metadata enhancement as data is ingested into the system. Additionally, we have a feedback mechanism that constantly improves the quality of inferences via user interactions & validations. Initial findings indicate significant improvements in data discoverability, accelerated onboarding for data users, and improved compliance readiness, resulting from enhanced metadata completeness and traceability. This work shows how data catalogues that use machine learning can change from simple documentation tools into these smart systems that make self-service analytics easier, promote data democracy & support truly self-describing big data platforms. This makes organizations more agile and focused on getting insights.*

Keywords - *Ml-Powered Data Catalogs, Metadata Inference, Big Data Ecosystems, Self-Describing Data, Schema Detection, Data Discovery, Data Governance, Machine Learning, And Data Lineage, Semantic Tagging, Automated Metadata Management.*

1. Introduction

In today's hyper-connected digital world, businesses are producing & storing information at an incredible pace. The 3Vs of big data are the amount, variety, and speed of data. This includes clickstreams, IoT sensor logs, social media material & commercial databases. This growth has had a big effect on how data is stored, searched, and analyzed. This huge data environment has a lot of promise to encourage their innovation and provide businesses an advantage over their competitors. However, it also has a lot of problems, particularly when it comes to finding, trusting & using their information.

One of the biggest problems is figuring out how to deal with the growing number of data sources & formats. As businesses begin to use cloud-native architectures, decentralized data lakes & actual time streaming technologies, their data environments are becoming more and more separate. In this disconnected environment, it is very hard to find useful information, understand their context, and make sure they are more accurate. Without a systematic way to catalogue, describe, and manage data assets, the potential of big data is frequently not realized.

Metadata, which is information about data, is at the heart of these kinds of problems. Schema definitions, data lineage, access restrictions & data quality measurements are all examples of important metadata. It gives data engineers, analysts, and scientists a way to understand, access & trust datasets. Well-organized information makes things more clear, helps people follow the rules, and encourages better decision-making. In short, without reliable metadata, big data is just noise.

But in this changing world, traditional ways of managing information don't work. Many old data catalogues are static repositories that need people to add new information and are only updated once in a while. Because of this, catalogues are frequently out of date, incomplete, or don't make sense. Manual curation takes a lot of time & is prone to mistakes. It also can't grow as fast as information is being created right now. In this setting, finding the right dataset might be a big problem for productivity, much alone getting useful information from it.

This research supports a shift from static, human-curated data catalogues to dynamic, machine learning (ML)-enhanced data catalogues that can adapt to the evolving features of these data ecosystems. These sophisticated catalogues employ ML and AI to automatically figure out metadata, find changes to schemas, check the quality of material & provide their documentation.

This helps create self-describing data ecosystems, where data assets have built-in, always-updated context that makes them easier to find, understand, and trust.

Fig 1: ML-Driven Data Catalog and Metadata Inference Framework

Imagine a system where metadata grows at the same time as the base data, where each new dataset is automatically profiled and categorized, and where semantic links between datasets are detected and highlighted without any help from people. This isn't a distant dream; it's the latest reality made possible by ML. These smart technologies make it unnecessary for people to tag things, make information more accurate, and encourage better governance and compliance. In addition, they help businesses get more value out of their information by making sure it is easy to access & understand.

This study seeks to examine the revolutionary influence of ML on metadata management in extensive data environments. We will examine the architecture, techniques & practical implementations of ML-enhanced data catalogues, as well as the challenges and considerations associated with their huge scale deployment. Our goal is to provide both academics and professionals who want to create or use self-describing data systems a theoretical framework and useful information.

2. Background and Literature Review

2.1. Traditional Data Catalogs: Purpose, Architecture, and Limitations

Data catalogues have long been important for helping businesses understand, manage, and find data. Their main job has been to be an inventory system, giving the whole company a single place to keep its data assets. Think of them as the "library catalogue" of company data, showing people where data is, how to get to it, and what it is.

Most traditional catalogues are built on their frameworks that don't change much. They commonly use ETL/ELT processes to get metadata from different data sources and put it in these relational databases or graph-based storage systems. Their architecture includes pipelines for ingesting information, layers for storing information, and UI/UX parts that make it easier to search & find things.

These catalogues are helpful, but they typically contain big problems. At first, they require a lot of help from people to stay up to date; data stewards or engineers must monitor & annotate the material. Second, they typically have trouble in dynamic & decentralized scenarios, like cloud-native architectures, whsere the latest data is always being created by different teams and systems. They fundamentally lack the ability to independently infer meaning or relationships; they depend on others for accurate data representation.

2.2. Managing Metadata: Types and Problems

Metadata is just "data about data," and it comes in many forms:

- Technical metadata shows how things are set up, including the names of columns, the kinds of data, the formats of tables, and so on.
- Business metadata connects information to business concepts, as shown in the definition of a "customer" field in a business setting.
- Operational metadata keeps track of runtime & behavioral information, such as lineage, access frequency, recency & these usage patterns.

Managing several kinds of metadata across these different platforms is really very hard. There are several problems:

- Collection: It's hard to automatically get metadata from siloed or proprietary systems.
- Standardization: Metadata typically comes in many other formats. One source could call a field "cust_id," while another would call it "customer_identifier."
- Quality assurance: Without established techniques, metadata might become old, incomplete, or misleading.
- Governance: It is hard to keep metadata accurate & compliant throughout a huge firm.
- As the amount of data grows, it becomes impossible for people to handle it all, hence automation & intelligence are needed.

2.3. Self-Describing Data Ecosystems: What They Are and Why They Matter

In recent years, the idea of self-describing data has become increasingly common, notably in the areas of data lakes, data mesh & federated analytics. Self-describing data refers to datasets that include their own metadata—such as structure, semantics & provenance either embedded within the data or accompanying it.

SON and Avro files typically have schema requirements built into their structure. OpenAPI specs that describe endpoints, parameters, and replies are commonly part of these APIs. These self-descriptive features let clients understand & use the data without needing any other further documentation.

This is very important when data is federated or decentralized. In a data mesh, where several teams or departments maintain their own data domains, consumers can't depend on a single catalogue to always have the most up-to-date descriptions. The data has to be able to explain itself.

However, self-describing ecosystems still have a hard time understanding semantics. It's helpful to know that a column is called "txn_amt," but does that mean "total amount," "pre-tax amount," or "discounted amount"? This is when machine learning (ML) begins to show its value.

2.4. Machine Learning in Data Management: Techniques for Supervised, Unsupervised, and Natural Language Processing

ML methods may change how we get information & make inferences, particularly when traditional rule-based systems don't work.

- Learning under supervision: This means utilizing tagged data to train models to find more patterns. Based on past examples, a model may learn to sort data columns into "PII" (personally identifiable information) or "financial fields." But this means that a tagged dataset is needed, which may be hard to find in messy actual world data situations.
- Unsupervised Learning: When there are no labels, techniques like clustering & topic modelling are useful. For example, putting similar column names from different datasets into groups might show where standardization is possible. Autoencoders and dimensionality reduction may help you find more hidden patterns in high-dimensional metadata.
- Natural Language Processing (NLP): The most interesting area is using NLP for schema and semantic inference. Machine learning algorithms can figure out what data columns probably mean by looking at these field names, descriptions & usage logs, such as SQL queries or user comments. Named entity recognition (NER), topic modelling, and embeddings help link raw data to human understanding.

These machine learning methods all try to turn static data catalogues into dynamic knowledge systems that can automatically understand, link, and explain data.

2.5. Related Work: Current Systems and Their Shortcomings

Recently, a number of technologies have come up that try to combine intelligence with data cataloguing. Some well-known examples are:

- Amundsen, which Lyft built, is an open-source data discovery & metadata engine that uses graph-based storage. Amundsen can automatically trace lineage and rank search results, but its ML abilities for semantic inference are limited.
- DataHub was made by LinkedIn and is a metadata platform that may be expanded. It makes it easier to use active metadata, keep track of schema change & provide some machine learning-based metadata improvement by these features.
- Collibra is a proprietary tool that helps with data management & the governance. Collibra makes it easier to add information using rules & ML, but it frequently has to work with other systems to make these inferences.
- Alation: Known for its "behavioural data catalogue," Alation stands out because it uses user behaviour, such as searches, queries, and clicks, to rank and recommend datasets. It uses NLP to look at SQL queries & gives smart labels, but its ML framework focuses more on how users interact with the data than on figuring out structural details.

Despite these improvements, numerous problems still exist:

- Not fully automated: A lot of systems still require a lot of human curation, especially when it comes to getting semantic information out of them.
- Scalability challenges: ML models for inference often demonstrate inadequate scalability over several other datasets without considerable technological intervention.
- Contextual limitations: Current systems struggle to grasp domain-specific nuances. In healthcare, the word "score" may mean something very different than it does in their finance.
- Integration with self-describing formats: Limited systems may use embedded schema formats like Avro, Thrift, and Protobuf, or they can figure out the schema from these API specifications on their own.

3. ML Techniques for Metadata Inference

In today's world of big data lakes and decentralised systems, it's just as important to understand the data you have and how to use it as it is to have it. In this case, metadata is quite too important. Metadata is like a label on a box within a big repository. Without it, searching for, integrating, ensuring the quality of & governing data becomes exceedingly hard. However, manually developing & maintaining metadata for petabytes of data is unfeasible.

This is where ML steps in. Machine learning might make it possible for these systems to become "self-describing" by figuring out patterns, linkages, and semantic meanings straight from the data. This part will look at four important areas of machine learning: data profiling and pattern recognition, schema inference, semantic tagging, and

data lineage inference. It will also look at how to test these systems.

3.1. Data Profiling and Recognising Patterns

Before any smart metadata inference can happen, computers need to look at the data to understand its structure and content. This is the beginning of pattern recognition and data profiling.

- Statistical Analysis for Type Inference Machine learning models may begin by employing fundamental statistical summaries such as minimum, maximum, standard deviation, mean, and mode for each column within a dataset. After that, these summaries are utilized to figure out what kind of data there are. You may call a column that primarily has integer values and sometimes has nulls "Integer" or "Nullable Integer."
- Machine learning methods like clustering (k-means, DBSCAN) may also group similar columns from different datasets, finding similarities like "this looks like a ZIP code," "this looks like a timestamp," or "this could be a phone number." These clusters provide the basis for broad type inference.
- Finding outliers and strange things to check for quality Data anomalies, such numbers that are too high, columns that are full of NULLs, or patterns that don't follow the rules, might indicate that the data isn't good enough. Models like Isolation Forests or autoencoders, which are designed to recreate actual data distributions, may be able to find more values or records that don't fit expected patterns.

These techniques help find wrong information as well. They also contain information regarding confidence, which shows how reliable a dataset is based on how well it fits with established criteria.

3.2. Inferring and Aligning Schemas

After understanding column-level patterns, the next step is to figure out schemas that show how data entities are structured and how they relate to each other.

- Sequential Models for Predicting Schemas Recurrent Neural Networks (RNNs) and Transformers are two modern machine learning models that are used to model sequences of properties in datasets.
- This is useful for working with nested structures like JSON or Avro, where it's important to understand the order and hierarchy of fields.
  For example, a Transformer model may notice that { "user_id": ..., "event_type": ..., "timestamp": ... } commonly appears in the same order in several JSON documents. This means that the structure is normal.
  By looking at a large number of "normal" documents, these models may learn to expect missing fields or find more schemas that don't match.
- Schema Correspondence Embeddings After schemas are expected, the next step is to align them. Recognising that two datasets with different column names refer to the same concept.
- This is where vector embeddings really shine. Word2vec or BERT are examples of algorithms that put the information from each column (name, type, sample values) into a space with a defined number of dimensions. The system can figure out that a column called "cust_id" in one dataset is probably the same as "customer_identifier" in another by looking at the cosine similarity between vectors.
- Schema matching using embeddings makes it easier to automate data integration, remove duplicate datasets, and search across several sources at once.

3.3. Semantic Annotation and Classification

It is helpful to understand the shape and structure of a column, but it is far more important to understand what it means. Semantic metadata includes things like "this field contains geographic coordinates" and "this is Personally Identifiable Information (PII)."

3.3.1. Recognising Entities and Modelling Topics

- Latent Dirichlet Allocation (LDA) is a common way to represent topics in text. It may group columns in tabular data into groups like "financial metrics," "user demographics," or "network activity."
- Named Entity Recognition (NER) models that were created using text corpora may also be used to add notes to column values. NER may put strings like "New York" or "California" under the category of these geographical locations. This would make it easier to automatically designate a column as Geo::State.
- Models that have been trained before and transfer learning
- Instead of training models from scratch, many other systems now employ transfer learning. When it comes to metadata work, pre-trained models like BERT, GPT, or specialist models like SciBERT for scientific data need to be fine-tuned.
- These models have a deep understanding of semantics and can sort of information by context, even when it's not clear what it means. The word "apple" might mean a fruit in one situation and a technological corporation in another, depending on how the dataset is used.

3.4. Inference Based on Lineage and Use

Metadata includes not just the information itself but also how it was made, changed, and used. This is where machine learning methods for lineage analysis may be used.

3.4.1. Looking at Query Logs

Companies frequently save a lot of SQL logs, access logs, or pipeline execution traces. By looking at these logs, machine learning models can figure out how datasets are related to each other. If table B is always asked for after table A, or if table A's properties are utilised to make table B, there is a good chance that there is a lineage link.

Markov Models, LSTMs, and simple frequency-based rules are examples of sequence mining methods that can find hidden processes without the requirement for explicit documentation.

3.4.2. Modelling Lineage using Probability

Machine learning models generally utilise a probabilistic approach since logs might be noisy or incomplete. Bayesian Networks and Hidden Markov Models (HMMs) are two examples of methods that may provide probabilities to inferred relationships. This lets the system say things like, "there is a 70% chance that dataset X came from Y."

Probabilistic lineage is essential in modern data lakes, since ETL processes sometimes lack formal documentation or are executed ad-hoc.

3.5. Evaluation Metrics: Checking for Success

All of these ways of making inferences are strong, but how can we tell whether they work?

3.5.1. F1 Score, Recall, and Precision

Like regular classification tasks, you may use the following to test metadata inference:

- Correctness: How accurate are the tags or kinds that were guessed?
- Remember: How many real tags or kinds can you reliably guess?
- F1-score: A fair way to measure accuracy and recall by taking the harmonic mean of the two.

You may figure out these numbers by comparing ML-generated metadata to a "gold standard" dataset, which is usually put together or checked by professionals in the field.

3.5.2. Impact on Data Discoverability

Another way to check how effective something is is to look at how well it finds data. Are consumers finding relevant datasets more quickly because of better metadata? Are search engines giving greater weight to results that are more relevant?

A/B testing or techniques for getting input from users may be used to check these improvements.

3.5.3. Better data integrity and oversight

In the end, inferred metadata should assist improve the quality of data and make sure it is up to code. Can the system find datasets on its own that have low quality ratings or that could include sensitive Personally Identifiable Information (PII)?

Monitoring metadata-driven interventions like searches that are not allowed, issues that have been raised, or automated classifications shows that the system works.

4. Architecture of an ML-Powered Data Catalog

4.1. System Overview

Modern businesses deal with data from a wide range of their sources, including structured, semi-structured, and unstructured data. A well-designed ML-driven data catalogue uses machine learning models to automatically consume, characterize & enhance data with metadata, bringing order to this chaotic world. It gives users simple search, recommendations & governance tools so they can find and trust the data they use.

4.2. Getting and preparing data

This is the front of the system, when raw data begins to change into a format that is structured, easy to search, and smart.

4.2.1. Managing Different Data Sources

- Relational databases, spreadsheets, and CSV files all have structured information.
- You may get semi-structured formats like JSON, XML, Avro & Parquet via APIs or cloud storage.
- Unstructured data may be anything that isn't organized, such as PDFs, text files, log files, emails, photographs, and audio.
- Pluggable adapters or connectors connect all the sources, making sure that the latest data kinds or systems may be added without any problems.

4.2.2. Cleaning and Making Things the Same

Before any analysis can be done, the data must be:

- Looked at and confirmed: Syntax correctness and schema compliance.
- Normalization is making sure that column names, data formats (such date formats), and value representations are all the same.
- Removed duplicates and tokenized: Getting raw information ready for metadata inference.
- Also, at this stage, sensitive information is tagged and hidden to follow these privacy laws like HIPAA or GDPR.

4.3. Analyzing and Deducing Metadata

After the information has been cleaned up, metadata profiling begins. The goal is to understand the meaning of the data on its own.

4.3.1. Profiling Skills

- Column statistics include the minimum, maximum, mean, standard deviation, and null percentages.
- Schema identification: figuring out what kind of data types there are, like integer, float, category, and boolean.

- Value distributions: Understanding how numbers are spread out in categories or intervals.
- Key identification: knowing what the primary keys, foreign keys, or connect paths are.
- Using machine learning to make semantic inferences

The machine learning algorithms look at the data to make predictions about:

- Column semantics (like "email address," "social security number," and "Internet Protocol address")
- Possible categories include Personally Identifiable Information, financial data, geographical information, and others.
- How similar datasets are for deduplication or integration
- Links and relationships between columns

The system begins with pre-trained models and keeps improving them using actual world data and user changes.

4.4. Engine for Machine Learning

The ML Engine is the brain that makes metadata prediction, catalogue improvement, and smart recommendations possible.

- Instructions for the Pipeline Framework Pipelines: Use annotated data from curated datasets or user reviews.
- Validation Pipelines: Keep checking that projections are correct.
- Inference Pipelines: Send model outputs to the catalogue in actual time or in batches.

Column name embeddings (for example, utilizing Word2Vec or BERT-like models) are one of the things that make up Feature Engineering Features. Recognising patterns in value (like regular expressions, entropy, and frequency analysis)

- Schema form embeddings
- Contextual indications obtained from dataset metadata and provenance
- Ways to Get Feedback
- The algorithm learns more and more by keeping track of how users fix mistakes.
- Evaluating the effectiveness of query success rates (searches that provide beneficial results)

Incorporating governance requirements or external taxonomies (for instance, corporate glossaries)These feedback loops make the model work better, cut down on the amount of human metadata curation needed, and make the system work better with lexicons that are specialised to a certain field.

4.5. Layer for Storing Metadata

This layer makes sure that all derived metadata is reliable, versioned, and can be searched.
What the Metadata Repository Holds:

- Results of Profiling
- Implied designations
- User-made notes
- Relationships between datasets (lineage, joins, transformations)
- Versioning makes it easier to navigate across time for schema change.
- Keeping an eye on how the meaning of data evolves over time
- Putting things back to as they were before if the inference was wrong
- Working with Data Lakes and Data Warehouses
- The catalogue is very closely linked to data lakes like AWS S3, Azure Data Lake, and Google Cloud Storage.
- Data warehouses, such as Snowflake, BigQuery, and Redshift
- This makes sure that metadata matches the actual storage locations and permissions for data, which makes access secure and easy to scale.

4.6. Application Programming Interfaces and the User Interface/User Experience Layer

The top of the stack is where the interface that connects users to the system is located.
Search Interface Users may search by keywords, schemas, tags, or plain English (for example, "find all customer email datasets").

- Sort by type of metadata, owner, date, or privacy
- Look at previews of datasets that were made automatically

4.6.1. Notifications and Suggestions

- Smart Suggestions: related datasets, tables that can be combined, and other sources.
- Alerts for schema drift, PII identification, and access breaches.
- Customized Feeds: Alerts on datasets that the user or their team routinely accesses or changes.

4.6.2. Functions for working together

Giving comments and notes on datasets

- Asking for access
- Allowing or changing metadata predictions
- Data stewards have the power to certify or identify "golden datasets."

5. Case Study: Implementing ML-Powered Data Catalog in a Real-World Big Data Platform

5.1. Context and Environment

"RetailNova" is a global retail company that does business online, in stores & with logistics. The company's data footprint includes petabytes of raw logs, customer transactions, IoT telemetry from the supply chain, feeds from third-party vendors, and internal business reports. Each department works semi-autonomously & uses different

technologies. For example, Apache Kafka is used for actual time streaming, Hadoop and Spark for processing data, Hive and Presto for querying, and Snowflake for long-term storage and analytics.

In the past, RetailNova's data discovery process was a mess. There were more than 50,000 tables spread out throughout departments, many of which were not documented and had strange names like "tmp_cust_3421_final_copy_v3." Data analysts and engineers were wasting time on activities that were already done or looking at data that was already there. The metadata was not enough, was old, or was made by hand.

To fix the mess and create a self-describing big data ecosystem that would make it easier to find, control, and reuse data, the leadership team decided to use a machine learning-driven data catalogue.

5.2. Problems with Deployment

- Size and Speed of Data: Every day, RetailNova's data environment grew by several terabytes. Customer clickstreams, changes in inventory in actual time, and IoT signals from more than 10,000 delivery trucks were continually changing. Old-fashioned ways of managing metadata couldn't keep up with the speed at which data was coming in, thus some information was missing or out of date.
- No Schema or Documentation: A huge percentage of the data, especially logs and IoT feeds, was either semi-organized or not structured at all. There was not enough documentation for many other datasets. Schemas were created organically, maybe even many times a week, but there wasn't enough versioning or tracking.
- Data engineers who were against automation were not sure about it. One of the senior engineers asked, "Can machine learning really figure out what column names mean?" People were worried that automated metadata tagging may lead to a loss of control or a higher chance of misclassification. Organizational inertia continued since the culture had made handwritten documentation a part of it, even if it didn't work.

5.3. Execution Details

5.3.1. Used Machine Learning Models

The group employed both supervised and unsupervised machine learning methods:

- NER stands for Named Entity Recognition. Created models that can find personal information (such names and addresses), timestamps, geolocation, and product information from schema names and sample data.
- Clustering and Similarity Models: Used embeddings (via Word2Vec) to group similar datasets and find duplicates across silos.
- Topic Modelling: We used Latent Dirichlet Allocation (LDA) with column descriptions and query logs to automatically construct summaries and tags for datasets.
- Schema Drift Detection: Time-series models found changes in tables that are updated on a frequent basis, such IoT feeds, and automatically classified them.

5.3.2. Putting the system together

Apache Spark was part of RetailNova's data architecture. It was used to look at sample data from tables and create ML models.

- Kafka for the use of actual time schema evolution events.
- Apache Atlas is the main metadata platform, and it has plugins for machine learning inference.
- We checked Hive Metastore and Snowflake for structured datasets every night to see whether the schema had changed.
- ElasticSearch to make it easier for users to search and find things in the catalogue.

An individualised orchestrator (created on Airflow) started machine learning jobs on a regular basis and made sure that new or changed datasets were evaluated on their own.

5.4. Results and Benefits

- Better coverage of metadata
  Before adoption, around 12% of datasets had a lot of important information. After six months of use, the number rose to 87%, with ML models automatically creating 65% of the tags and descriptions. People who curate things went from inventing to validating, which cut down on the amount of work that needed to be done by hand.
- Improved query speed and less duplication
  Data analysts said that the time it took to find data was down by 40%. Teams that used to make extra reports using similar data sources started utilising the standardised datasets that the catalogue suggested. Audit records show that the number of duplicate queries went down by 28%. Governance teams may easily find old or useless datasets that need to be fixed.
- Using and Getting Feedback from Customers
  Adoption didn't happen right away, but as people saw the initial successes—like an analyst finding a key KPI stream that had never been seen before—information spread quickly. Users could check the quality of the information and suggest changes via the user interface, which created a feedback loop that made the machine learning models better. Weekly office hours were held to answer questions and make the move easier.

Some engineers liked having a person in charge, but they also liked being able to change tags made by machine learning or provide context directly via the user interface.

Over time, the catalogue became a dependable part of the data lifecycle.

6. Discussion and Future Directions

6.1. The Transformational Impact of ML-Powered Data Catalogs

As businesses deal with more and more organized and unstructured information, they need better tools to store, organize, and evaluate this information. Machine Learning (ML)-driven data catalogues have come into this field with a promise: to turn messy data lakes into environments that are easy to use & describe themselves.

Scalability is perhaps the most important benefit. In places where petabytes of data are created every day, traditional manual methods for tagging & categorizing information don't work. ML models simplify this process by figuring out the schema, lineage, and linkages, even from forms that are only partially organized or that are the latest to them. This capability allows data teams to easily add new sources and stay up to date on how data ecosystems are developing.

Another good thing is that it is accurate. As time goes on, ML models become better at figuring out what kinds of data there are, how they are utilised, and what they mean in context. They aren't flawless, but they're a lot better than solely depending on their human categorisation, particularly when it comes to discovering duplicates, outliers, or hidden linkages in big datasets.

Making things easier for users is a significant bonus. Because of smart recommendations, semantic search capabilities, and easy-to-use lineage tracking, non-technical users may be able to identify and comprehend data assets without much support from data engineers. The modification allows analysts, product managers, and other people who are interested to make choices based on data on their own.

6.2. Problems and Limitations Need More Thought

ML-driven data catalogues have a lot of problems, even if they seem like they will work.

The cold-start problem happens a lot. ML algorithms can't identify the historical patterns they require to draw correct conclusions when there isn't much or any additional information, such in freshly produced data lakes. This usually means that teams have to employ human data input until a baseline is established, which is the opposite of what automation is supposed do.

Another limit is how easy it is to understand. As these algorithms find relationships, classifications, and methods to assess data quality, it may not be evident why they do what they do. This lack of transparency might be a big concern in areas where trust and the ability to be audited are particularly essential. Companies need a mechanism to verify where metadata originates from, and it would be great if this came with a lot of explicit explanations or metrics that show how confident they are. A third, more essential barrier is security and privacy. Automatically collecting metadata from sensitive datasets may unintentionally reveal private or controlled information. If you don't label personal identifiers correctly or at all, you might face many serious legal problems under laws like GDPR or HIPAA. The protection of ML pipelines' security and privacy is the latest area of study.

6.3. Trends on the Way

Future advancements are set to enhance the effectiveness & the reliability of machine learning-based libraries.

Generative AI is changing the way metadata is put together. Instead of only finding existing material, large language models (LLMs) may now use contextual signals and domain-specific corpora to create documentation that people can read, suggest rules for governance & fill in blank fields automatically. This adds a whole latest level of usefulness to data catalogues.

Connecting with knowledge graphs is a big step forward. When ML-derived information is founded on structured domain ontologies, the result is much more contextually enriched and interoperable. This relationship makes things more accurate & makes it easier to do better cross-system searches, semantic searches, and produce the latest insights.

Active learning & methods that include people are becoming more popular. Allowing human curators to check or fix ML predictions makes the system even smarter over time. This hybrid technique combines the effectiveness of automation with the contextual knowledge of subject experts, which leads to more reliable & responsible metadata ecosystems.

7. Conclusion

As modern businesses rely more and more on huge, varied data ecosystems, intelligent, automated data management has become a must. This study investigated the transformative role of ML-driven data catalogues in enabling self-describing massive data environments, namely via the enhancement of metadata discovery, classification & governance.

Because they rely on their human processes and set standards, traditional metadata management methods typically don't work. These solutions don't work well with the pace & diversity of the information that modern data lakes, cloud platforms & distributed these pipelines create. On the other hand, data catalogues that employ ML may learn on their own from data patterns, query histories, and user behavior to figure out, improve & verify content. This improvement makes data descriptions more accurate, speeds up the time it takes to get insights & boosts trust in an organization's data assets.

We showed how ML makes metadata operations deeper & more scalable by looking at several ways to infer information, such as schema prediction, semantic tagging & lineage tracing. We also gave an architectural framework for building a ML-driven data catalogue, which showed its main parts and how they fit together. A case study from the actual world showed that there are many genuine benefits, such as better data quality, easier data discovery, and more efficient operations.

The journey is full of problems, even if the advantages are more tremendous. Model interpretability, metadata versioning & user trust are all problems that demand constant innovation and careful design. Combining human-in-the-loop approaches with generative AI opens up intriguing the latest ways to build more complete and user-friendly metadata ecosystems.

In short, ML-powered data catalogues help make data more accessible to everyone and are the building blocks of data infrastructures that are ready for the future. As businesses work to become more data-driven, they will need to invest in smart, flexible metadata systems to get the most out of their data.

References

[1] Chard, K., D'Arcy, M., Heavner, B., Foster, I., Kesselman, C., Madduri, R., ... & Toga, A. (2016, December). I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets. In *2016 Ieee international conference on big data (big data)* (pp. 319-328). IEEE.

[2] Nagorny, K., Scholze, S., Ruhl, M., & Colombo, A. W. (2018, May). Semantical support for a CPS data marketplace to prepare Big Data analytics in smart manufacturing environments. In *2018 IEEE Industrial Cyber-Physical Systems (ICPS)* (pp. 206-211). IEEE.

[3] Suryadevara, S. S. K., & Shaik, K. (2023). Real-Time Anomaly Detection and Attack Mitigation for Cloud-Based Content Delivery Paths Using AI. International Journal of Emerging Research in Engineering and Technology, 4(1), 175-185. https://doi.org/10.63282/3050-922X.IJERET-V4I1P119

[4] Kumar Doodala, A. N. (2024). Validating UX consistency Across Omnichannel Platform. American International Journal of Computer Science and Technology, 6(6), 87-97. https://doi.org/10.63282/3117-5481/AIJCST-V6I6P109

[5] Lokers, R., Van Randen, Y., Knapen, R., Gaubitzer, S., Zudin, S., & Janssen, S. (2015, September). Improving access to big data in agriculture and forestry using semantic technologies. In *Research Conference on Metadata and Semantics Research* (pp. 369-380). Cham: Springer International Publishing.

[6] Vppalapati, M. (2024). Power-Bound Storage Design: Architecting Systems for Electrical Scarcity. International Journal of AI, BigData, Computational and Management Studies, 5(1), 208-217. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I1P121

[7] Khalifa, S., Elshater, Y., Sundaravarathan, K., Bhat, A., Martin, P., Imam, F., ... & Statchuk, C. (2016). The six pillars for building big data analytics ecosystems. *ACM Computing Surveys (CSUR)*, *49*(2), 1-36.

[8] Parakala, A. (2023). Citizen-Facing Automation: Chatbots and Self-Service in Public Services. International Journal of AI, BigData, Computational and Management Studies, 4(4), 108-118. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P112

[9] Gaddam, R. R., & Krishna, K. (2023). KFP v2 Artifact-Centric ML Pipeline Governance. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(2), 142-153. https://doi.org/10.63282/3050-9262.IJAIDSML-V4I2P116

[10] Eynard-Bontemps, G., Abernathey, R., Hamman, J., Ponte, A., & Rath, W. (2019, January). The PANGEO Big Data Ecosystem and its use at CNES. In *Big Data from Space (BiDS'19).... Turning Data into insights... 19-21 february 2019, Munich, Germany*.

[11] Muppaneni, R. K. (2022). Data Privacy in the Age of AI: How Dynamics 365 Handles Regulatory Challenges. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(4), 159-170. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P117

[12] Kumar Doodala, A. N. (2024). Service Virtualization for API-First development: A shift-Left Testing Strategy. American International Journal of Computer Science and Technology, 6(4), 50-58. https://doi.org/10.63282/3117-5481/AIJCST-V6I4P105

[13] Ghiringhelli, L. M., Carbogno, C., Levchenko, S., Mohamed, F., Huhs, G., Lüders, M., ... & Scheffler, M. (2017). Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *npj computational materials*, *3*(1), 46.

[14] Suryadevara, S. S. K., & Nakirikanti, S. (2023). Privacy-Preserving Personalization Using Federated Learning in AEM . International Journal of AI, BigData, Computational and Management Studies, 4(4), 190-199. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I4P119

[15] Srigadde, B. R., & Devaraju, J. M. (2024). Building a Reusable AI Connection Utility Class. International Journal of Emerging Research in Engineering and Technology, 5(2), 188-200. https://doi.org/10.63282/3050-922X.IJERET-V5I2P119

[16] Berman, J. J. (2013). *Principles of big data: preparing, sharing, and analyzing complex information*. Newnes.

[17] Shiramalla, R. (2024). Secure Multi-Cloud API Orchestration between Salesforce, Oracle CPQ, and Azure. American International Journal of Computer Science and Technology, 6(3), 102-

113. https://doi.org/10.63282/3117-5481/AIJCST-V6I3P108

[18] Vppalapati, M. (2024). Cooling Domains as First-Class Failure Boundaries in Storage Architecture. American International Journal of Computer Science and Technology, 6(2), 96-106. https://doi.org/10.63282/3117-5481/AIJCST-V6I2P110

[19] Curry, E. (2020). *Real-time linked dataspaces: Enabling data ecosystems for intelligent systems*. Springer Nature.

[20] Takkalapally, D. (2024). ShiftLeft-AI: Machine Learning Framework for Proactive Performance Assurance in CI/CD Pipelines. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(4), 285-296. https://doi.org/10.63282/3050-9262.IJAIDSML-V5I4P126

[21] Verma, D., Overton, J., Wright, B., Purcell, J., Santhar, S., Das, A., ... & Prasad, S. (2022, August). Self-describing digital assets and their applications in an integrated science and engineering ecosystem. In *Smoky Mountains Computational Sciences and Engineering Conference* (pp. 274-287). Cham: Springer Nature Switzerland.

[22] Allenki, S. S. (2023). Reducing Security Vulnerabilities with Encryption, IAM, and Regular Audits. International Journal of Emerging Trends in Computer Science and Information Technology, 4(1), 265-275. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P127

[23] Srigadde, B. R. (2024). Agents, LLMs, and Salesforce with Multi-Cloud Provider (MCP). International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(3), 277-288. https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P127

[24] Amirian, P., van Loggerenberg, F., & Lang, T. (2017). Big data and big data technologies. In *Big Data in Healthcare: Extracting Knowledge from Point-of-Care Machines* (pp. 39-58). Cham: Springer International Publishing.

[25] Shiramalla, R. (2023). Optimizing Cross-Platform Enterprise Integrations Using Workato: A Case Study of Salesforce and Oracle SaaS Applications. International Journal of Emerging Trends in Computer Science and Information Technology, 4(1), 232-243. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P124

[26] Muppaneni, K. (2024). Progressive Web Apps: Offline UX Benchmarking. International Journal of Emerging Trends in Computer Science and Information Technology, 5(2), 174-183. https://doi.org/10.63282/3050-9246.IJETCSIT-V5I2P119

[27] Nagarjuna, D. N., & Yogesh, N. (2015). A Survey on Hadoop Architecture & its Ecosystem to Process Big Data-Real World Hadoop Use Cases.

[28] Takkalapally, D., & Takkellapally, M. R. (2024). AI-SynPerf: Synthetic Data Intelligence Framework for 5G Mobile Performance Simulation. International Journal of Emerging Trends in Computer Science and Information Technology, 5(1), 182-194. https://doi.org/10.63282/3050-9246.IJETCSIT-V5I1P118

[29] Gaddam, R. R. (2023). Progressive Delivery for Models with Quality KPIs. American International Journal of Computer Science and Technology, 5(4), 33-47. https://doi.org/10.63282/3117-5481/AIJCST-V5I4P104

[30] Niu, C., Zhang, W., Byna, S., & Chen, Y. (2023, December). PSQS: Parallel Semantic Querying Service for Self-describing File Formats. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 536-541). IEEE.

[31] Muppaneni, R. K. (2022). From Legacy ERP to Cloud-First: A Transformation Story with Dynamics 365. International Journal of Emerging Research in Engineering and Technology, 3(4), 153-164. https://doi.org/10.63282/3050-922X.IJERET-V3I4P117

[32] Parakala, A. (2023). Vendor Highlights – IoT, AI, and Process Mining. International Journal of Emerging Trends in Computer Science and Information Technology, 4(4), 135-146. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I4P115

[33] Mehta, S., & Mehta, V. (2016). Hadoop ecosystem: An introduction. *International Journal of Science and Research (IJSR)*, *5*(6), 557-562.

[34] Muppaneni, K., & Palem, V. (2024). Micro-Frontend Design Patterns for Multi-Framework Applications. International Journal of Emerging Research in Engineering and Technology, 5(3), 181-190. https://doi.org/10.63282/3050-922X.IJERET-V5I3P120

[35] Allenki, S. S. (2023). Applying Cloud Security Best Practices in Regulated Environments. American International Journal of Computer Science and Technology, 5(3), 48-60. https://doi.org/10.63282/3117-5481/AIJCST-V5I3P105

[36] Zender, C. S. (2008). Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). *Environmental Modelling & Software*, *23*(10-11), 1338-1342.

[37] Lukowski, M., Prokhorenkov, A., & Grossman, R. L. (2023). Towards self-describing and FAIR bulk formats for biomedical data. *PLOS Computational Biology*, *19*(3), e1010944.

[38] Taluri, R. (2024). A Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management, and Automated Data Quality Assessment on AWS. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(2), 230-240. https://doi.org/10.63282/3050-9262.IJAIDSML-V5I2P126