COMPUTER SOCIETY OF INDIA
ESTD. 1965
सर्वे भवन्तु सुखिनः

e6data

HACKATHON PROJECT OVERVIEW

Project Name: MetaView

Problem Statement & ID: Problem Statement 2 - Metastore Viewer for Parquet, Iceberg, Delta & Hudi Tables on S3

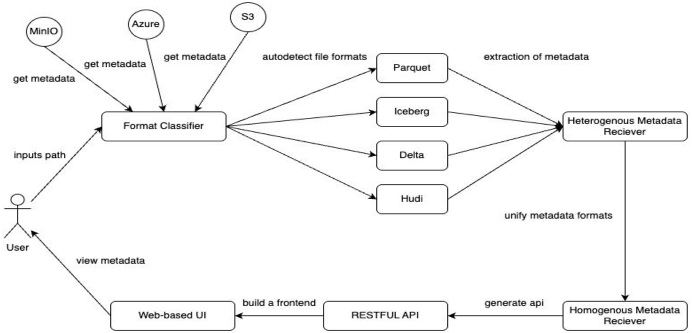
Team Name: Mjolnir

College: COEP Technological University

City: Pune

PROPOSED SOLUTION

Solution Overview & Key Components



Detailed Explanation on how the solution works

1. User inputs Object Store Path,
2. Classifier **Auto-Detects** Table format,
3. **Extract Heterogenous Metadata** which is format specific
4. **Parse it** and convert it to a **homogenous**, unified format,
5. Using a **REST API** send it to website frontend,
6. UI Displays schema, partitions, snapshots and Metrics

How It Addresses the Problem (Advantages)

1. Eliminates Dependency on Traditional Metastores like **Glue** and **Hive**,
2. No need to register tables manually, just provide the **S3** (or equivalent service) **path** and our system retrieves metadata instantly
3. Supports multiple formats – **Parquet**, **Iceberg**, **Delta** and **Hudi**,
4. Extracts rich Metadata to provide valuable insights,
5. Improves data governance by implementing a **uniform** metadata format.

Innovation & Uniqueness

1. No table registration required for Metadata exploration,
2. Has the ability to **visualise** schema evolution over time,
3. **Interactive** UI for **Metadata exploration** instead of users querying in SQL,
4. Built for a combination of Lakehouse formats.

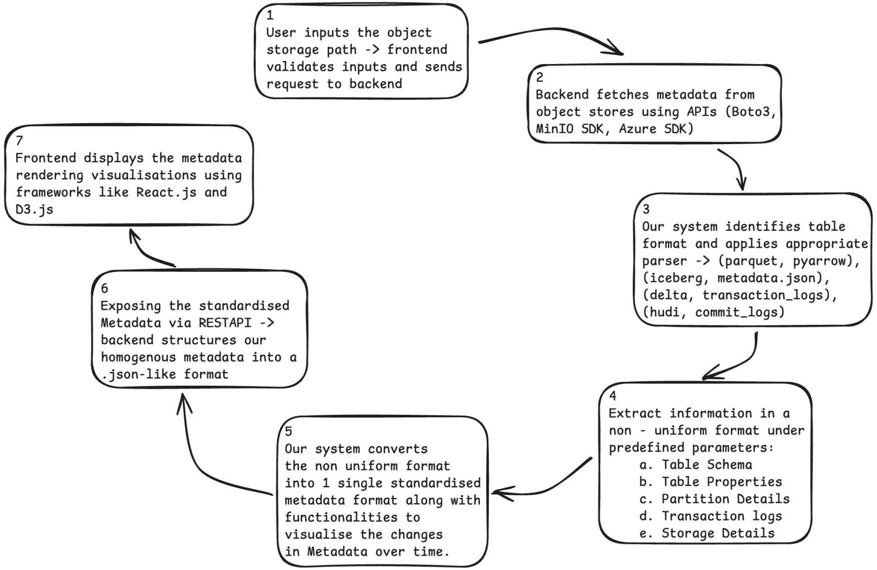
Cutting-edge technologies or novel approaches

1. **Format Classifier**: Automatically Detect the format of the table stored in S3 (or Azure, MinIO) using NLP or and equivalent fine tuned LLM,
2. **AWS SDK and Object Store APIs**: Interact directly with S3, MinIO, or Azure for real time metadata retrieval,
3. **React.js and D3.js**: Provide an appealing GUI for visualising transformations in metadata,
4. **Serverless & Cloud-Native Approach**: Can be deployed as a **Lambda function** or as a **Docker container**.

APPROACH

Methodology to Solve the Problem

Step-by-step explanation & Logical Breakdown



1 User inputs the object storage path -> frontend validates inputs and sends request to backend

2 Backend fetches metadata from object stores using APIs (Boto3, MinIO SDK, Azure SDK)

3 Our system identifies table format and applies appropriate parser -> (parquet, pyarrow), (iceberg, metadata.json), (delta, transaction_logs), (hudi, commit_logs)

4 Extract information in a non - uniform format under predefined parameters:
a. Table Schema
b. Table Properties
c. Partition Details
d. Transaction logs
e. Storage Details

5 Our system converts the non uniform format into 1 single standardised metadata format along with functionalities to visualise the changes in Metadata over time.

6 Exposing the standardised Metadata via RESTAPI -> backend structures our homogenous metadata into a .json-like format

7 Frontend displays the metadata rendering visualisations using frameworks like React.js and D3.js

Key Challenges & How they are addressed

1. Handling different table formats
   - Classification of different formats to apply the right parser
   (parquet -> **pyarrow**, delta -> **delta_api**, iceberg -> **metadata.json_parser**)
2. Employing the right parser
   - Scanning metadata files using **Natural Language Processing** to choose the right parser for extracting data
3. Efficiently Extracting Metadata from large datasets
   - Do not scan the whole dataset, instead use the **manifest files** & **transaction logs** to retrieve metadata
4. API Performance & Scalibility
   - Use **FastAPI** (async processing) and deploy in a serverless / cloud – native architecture (**AWS Lambda**, **Kubernetes**)
5. Ensuring a User- Friendly Interface
   - Interactive UI with metadata visualisation features that can **interpolate** the “state” of the data

USPs & Features

COMPUTER SOCIETY OF INDIA
ESTD. 1965
सर्वे भवन्तु सुखिनः

e6data

Unique Selling Points

- Unlike Hive/Glue, our tool requires **no table registration**—just provide the object storage path and explore instantly.
- Unlike specialized viewers that focus on one format, **MetaView seamlessly handles** Parquet, Iceberg, Delta and Hudi tables through a **single interface**.
- Automatically **classifies table formats** and processes metadata intelligently, reducing manual effort.
- Provides **quick metadata insights** for decision-making without the need for expensive data warehousing tools.

What makes the solution special?

- **No more Hive or Glue dependency** – Just plug in the storage path & explore metadata.
- **No SQL expertise needed** – Intuitive UI lets users browse metadata visually.
- **Multi-Cloud Support** – Works seamlessly with AWS S3, MinIO, and Azure Blob Storage.
- **Optimized for the Lakehouse** – Handles evolving data lakes **without manual intervention**.

Key Features

1. Auto-Detection and Classification of Table Formats
2. Direct Metadata Access and Instant Retrieval
3. Interactive UI for Metadata Exploration

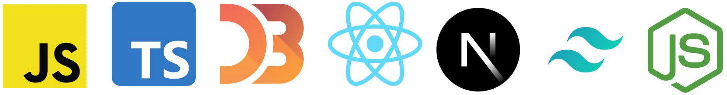
Technologies & Implementation

COMPUTER SOCIETY OF INDIA
ESTD. 1965
सर्वे भवन्तु सुखिनः

e6data

Tech Stack

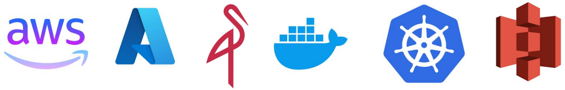
Frontend (UI and Visualization)



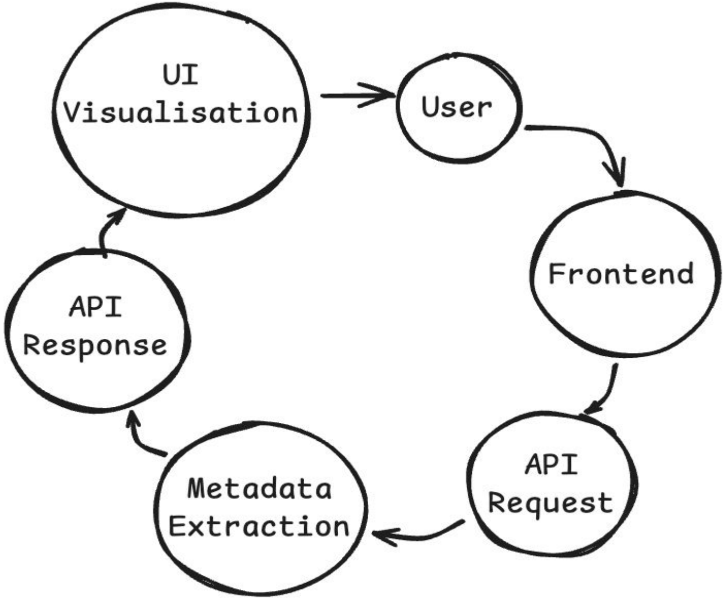
Backend (Metadata Extraction and API Layer)

DELTA LAKE
Flask

Frameworks and Additional Tools



Data Flow Diagram (Implementation Methodology)



UI Visualisation → User → Frontend → API Request → Metadata Extraction → API Response → UI Visualisation

Potential Impact

Impact on Target Audience

1. Gain instant metadata insights without needing a meta store (**data engineers** & **analysts**)
2. Quickly explore table schema and structure before running queries (**data scientists**)
3. Optimise storage and performance with metadata driven insights (**cloud architects**)
4. Manage tables without complex infrastructure (**organisations** & **data lakes**)

Social Impact

1. Encouraging **Open Source** & **Cloud Native Communities** by reducing dependencies on proprietary meta stores,
2. Making large-scale metadata easily accessible
3. Increasing adoption of **modern data lake technologies**

Economic Benefits

1. Cost savings as no need to use expensive meta store services like **AWS Glue**, **Databricks Unity**, or **Hive Metastore**,
2. Quick metadata access improves **data pipeline efficiency**,
3. Better resource utilisation leading to optimise storage and compute costs

Benefits of the Solution

1. No meta store required – directly read metadata from object store,
2. **Faster** metadata **retrieval** – no need to scan full tables,
3. **Multi – format support** – works with Parquet, Iceberg, Delta, and Hudi,
4. Scalability & Flexibility – **Cloud based**, **serverless** architecture,
5. User-Friendly Interface – Interactive UI for schema browsing, version tracking & **visualisation**
6. **Agentic Environment** – An agent-like system that redirects metadata retrieval based on the various formats and then converts into a single unified metadata format

Environmental Contribution

1. Enables **efficient data pruning** by leveraging metadata insights,
2. Reducing **unnecessary data scans** & processing, leading to **lower cloud energy consumption**,

References & Additional Links

Links to References & Research Papers

Lakehouse Architecture & Metadata Management

1. https://iceberg.apache.org/
2. https://delta.io/
3. https://hudi.apache.org/
4. https://www.databricks.com/glossary/what-is-parquet

Research Paper Links

1. "Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores" - https://dl.acm.org/doi/pdf/10.14778/3415478.3415560

Additional Documentation

Official API & SDK Docs for Storage & Table formats

1. https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/s3.html
2. https://min.io/docs/minio/macos/index.html
3. https://learn.microsoft.com/en-us/azure/storage/blobs/
4. https://devdocs.io/fastapi/

Code Repositories & Open Source Projects for Reference

1. https://github.com/apache/iceberg
2. https://github.com/delta-io/delta
3. https://github.com/apache/hudi
4. https://github.com/trinodb/trino

Team Details

Anshul Shelokar	TY	CSE	COEP Tech. Uni.	anshulshelokar777@gmail.com
Tejas Kolhe	TY	CSE	COEP Tech. Uni.	tejaskolhe0505@gmail.com
Paras Dhole	TY	CSE	COEP Tech. Uni.	parasdhole23@gmail.com
Chinmay Mulmule	TY	CSE	COEP Tech. Uni.	mulmulechinmay10@gmail.com