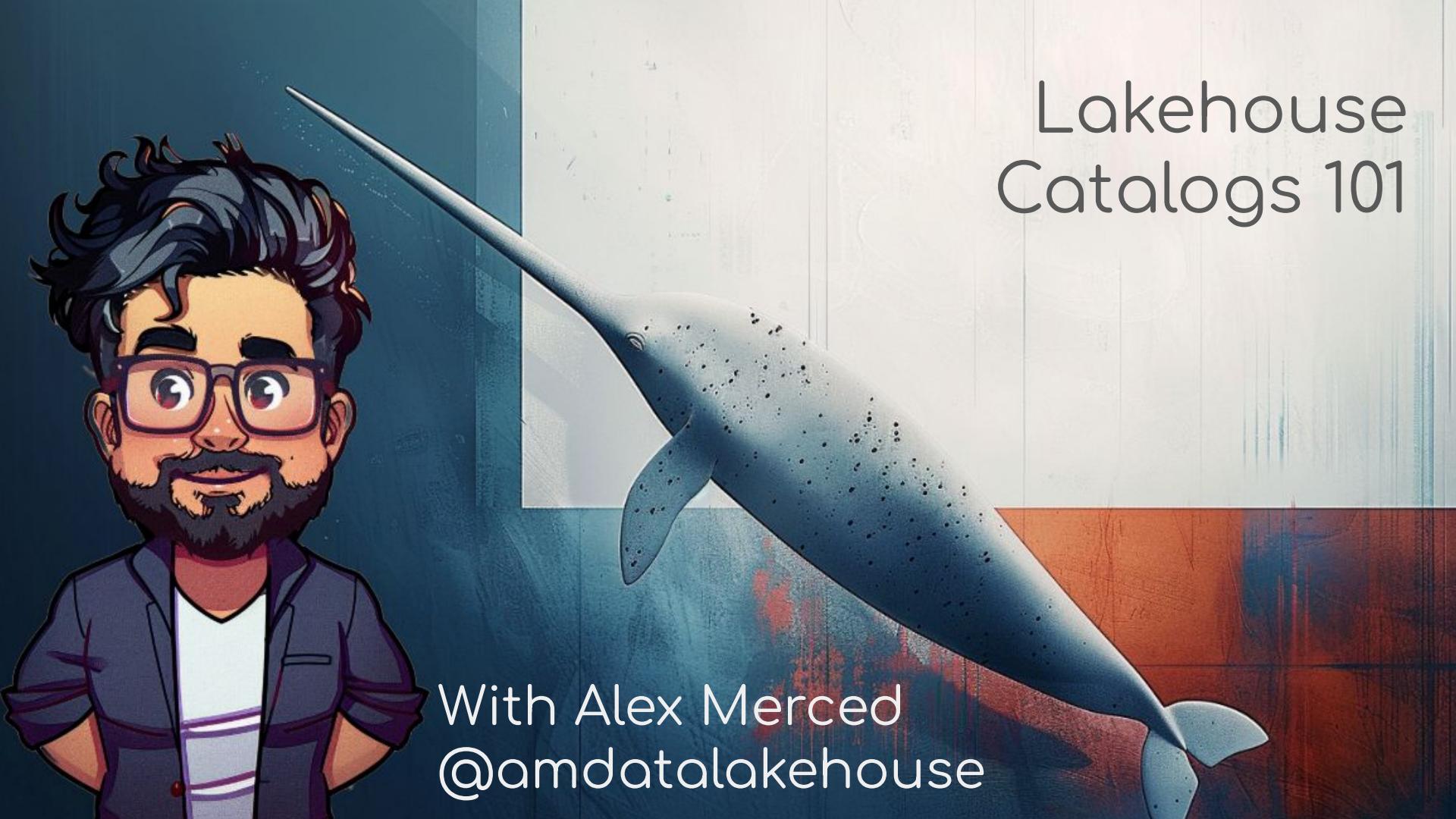


Lakehouse Catalogs 101

With Alex Merced
[@amdatalakehouse](https://twitter.com/amdatalakehouse)



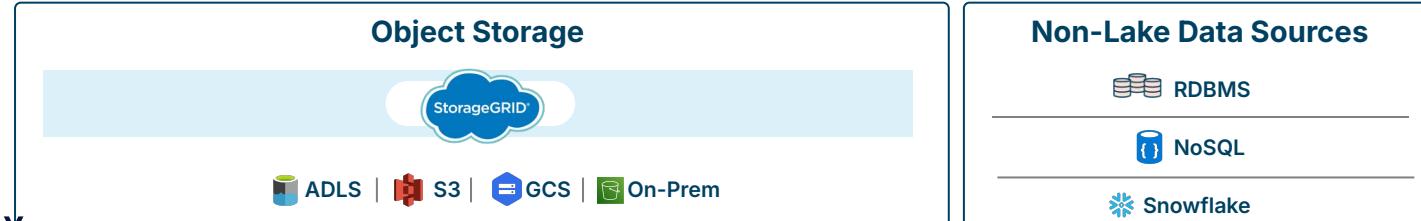
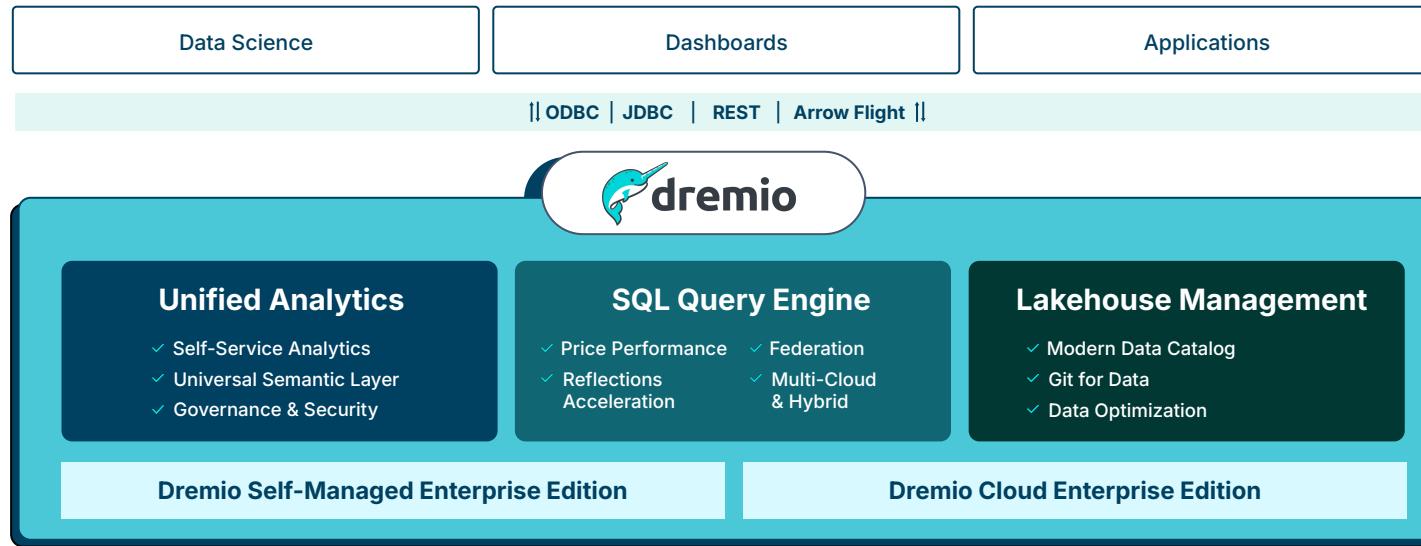


Alex Merced

Senior Technical Evangelist, Dremio

Alex Merced is a senior technical evangelist at Dremio with experience as a developer and instructor. His professional journey includes roles at GenEd Systems, Crossfield Digital, CampusGuard, and General Assembly. He co-authored "**Apache Iceberg: The Definitive Guide**" published by O'Reilly and has spoken at notable events such as Data Day Texas and Data Council. Alex is passionate about technology, sharing his expertise through blogs, videos, podcasts like Datanation and Web Dev 101, and contributions to the JavaScript and Python communities with libraries like SencilloDB and CoquitoJS.

Dremio: The Unified Lakehouse Platform for Self-Service Analytics & AI

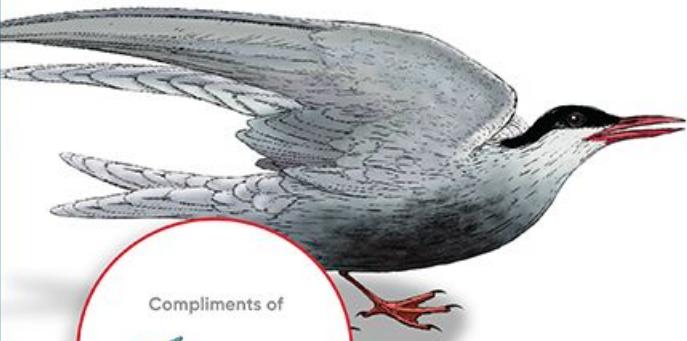


O'REILLY®

Apache Iceberg

The Definitive Guide

Data Lakehouse Functionality, Performance,
and Scalability on the Data Lake



Compliments of



Tomer Shiran,
Jason Hughes &
Alex Merced

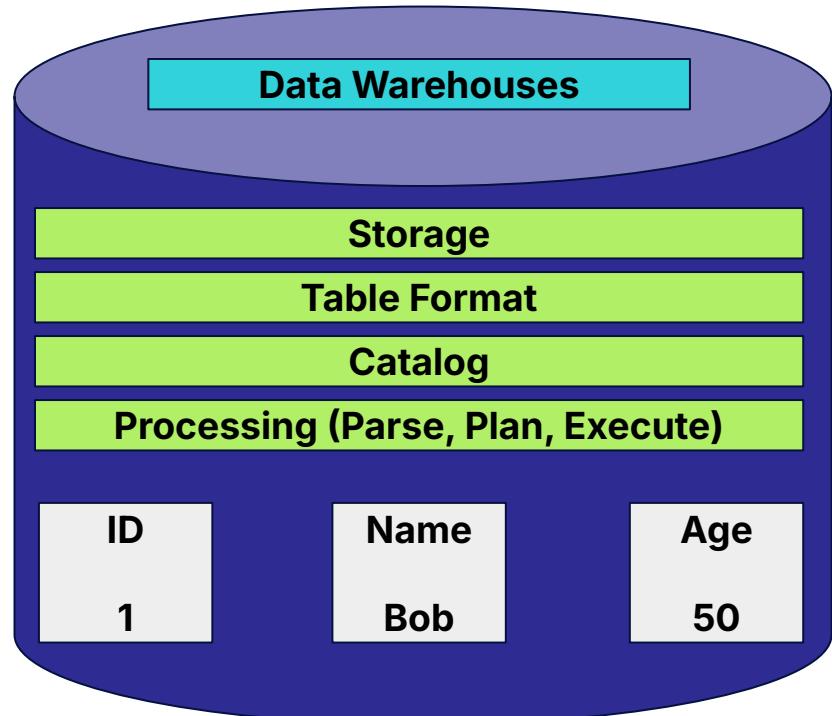
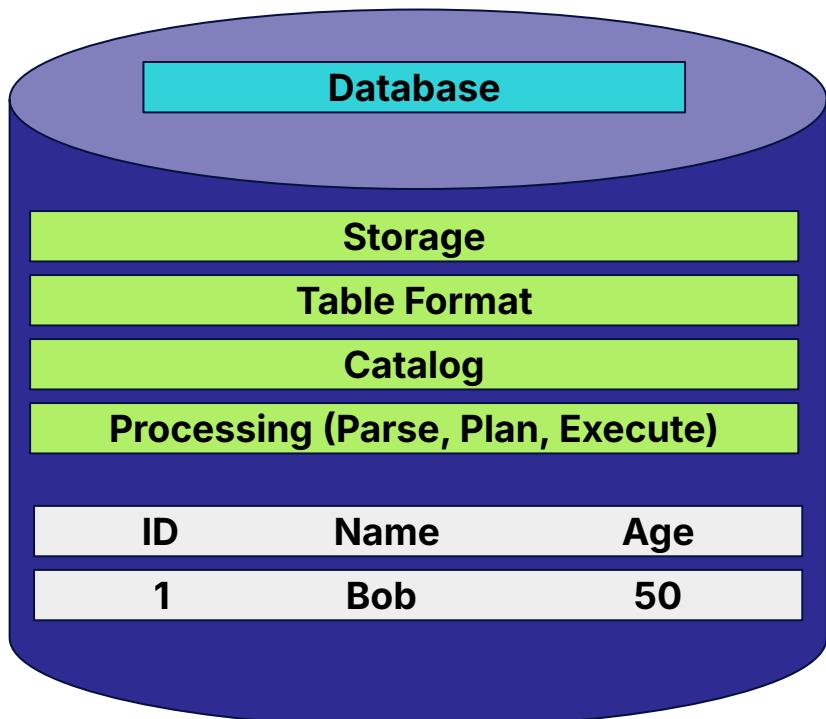
Forewords by Gerrit Kazmaier,
Raghu Ramakrishnan & Rick Sears



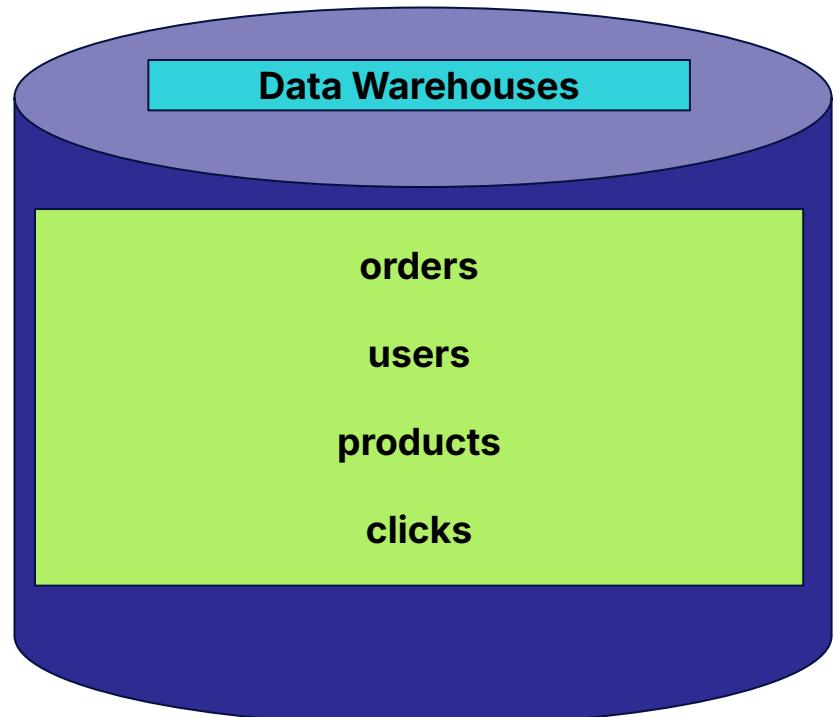
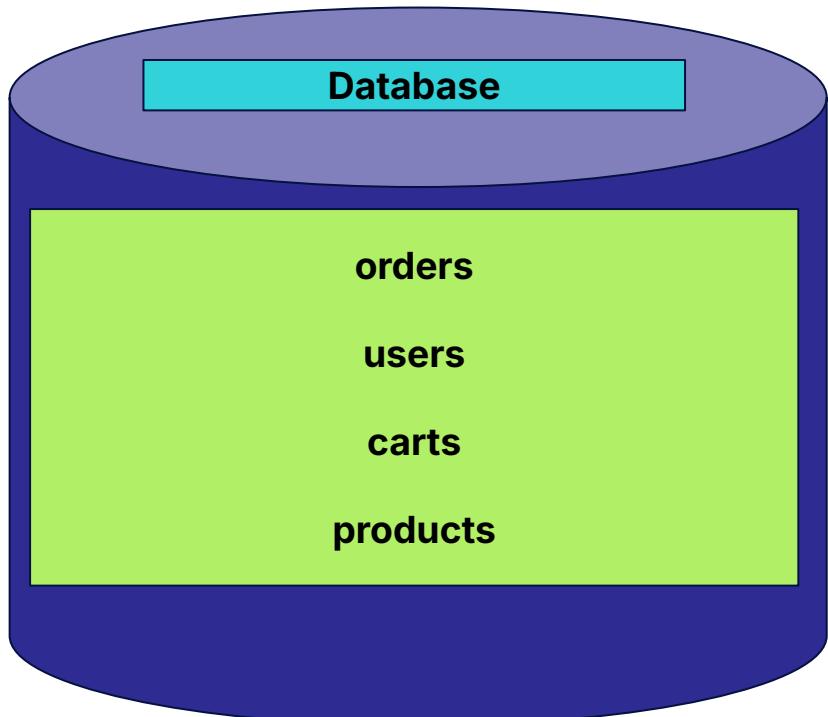
dremio

What is a catalog?

Traditional Data Systems



Traditional Data Systems



Enterprise Data Catalogs

sales	postgres
<p>The Sales table records transactional data from customer purchases, capturing key details of each sale. Each record includes a unique Sales_ID to identify transactions, Customer_ID for customer association, and Product_ID to link to specific items. The table also holds Sale_Date to timestamp each purchase, Quantity_Sold to show item counts per transaction, and Sale_Amount reflecting total revenue for the sale. Additionally, columns such as Store_Location and Sales_Channel (e.g., online or in-store) offer insights into purchase venues. This structured sales data enables detailed sales tracking and analysis for revenue, customer behavior, and product performance across different channels and locations.</p>	
Request Access	

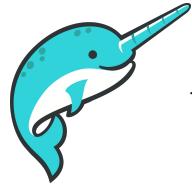
clicks	kafka
<p>The Clicks table captures web interaction data streamed from a Kafka topic, logging each user click in real-time for activity monitoring and analysis. Each entry contains a unique Click_ID for tracking individual events, User_ID to associate clicks with specific users, and Page_URL to identify the webpage or app section visited. Additionally, Click_Timestamp records the exact moment of interaction, while Device_Type (e.g., mobile, desktop) and Referrer_URL indicate the device used and originating source of the traffic. This table enables high-frequency data ingestion, providing crucial insights into user engagement patterns, navigation behavior, and click-through rates across various digital touchpoints.</p>	
Request Access	

What is Data Lakehouse?

What is a Data Lakehouse?



Lakehouse Catalogs

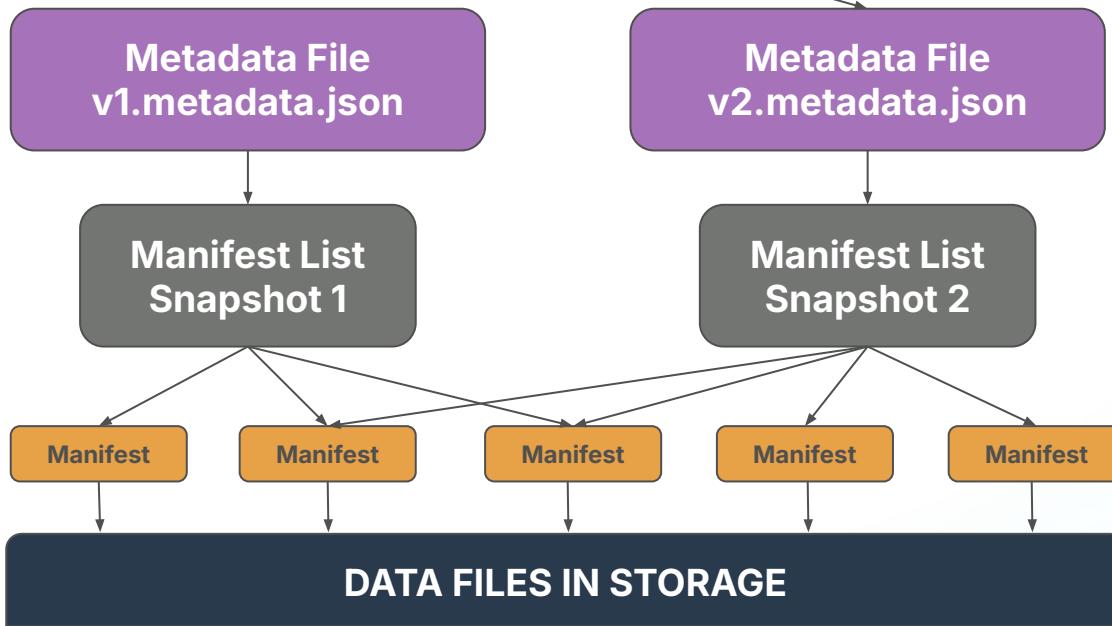


Lakehouse Catalog		
<u>Asset</u>	<u>Type</u>	<u>Location</u>
sales	table	s3://
products	table	abfss://
marketing	namespace	
accounting	namespace	



DuckDB

What is Apache Iceberg?



1. Query Engine consults catalog to discover location of latest metadata file.

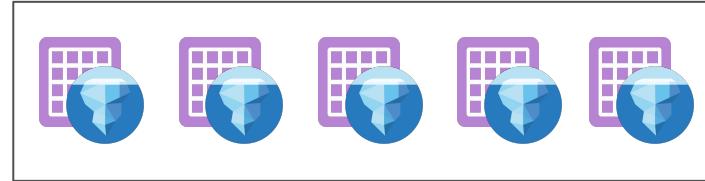
2. Metadata file is read for partitioning, schema and for location of manifest list of desired snapshot

3. Manifest list is used to do partition pruning and identify manifest with relevant files.

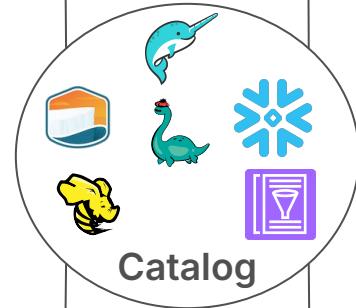
4. Manifests are used to do min/max filter to prune individual files, remaining files are then scanned for query.

Apache Iceberg Catalogs

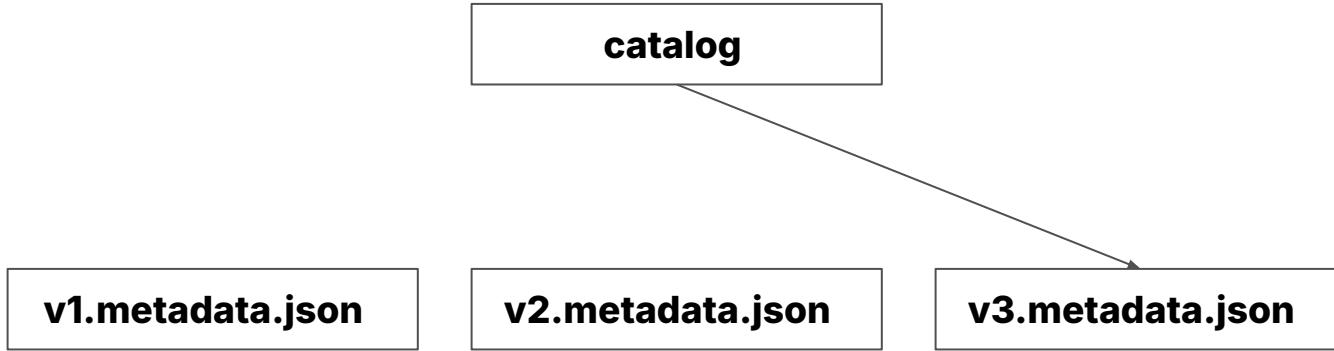
Data Lake Storage



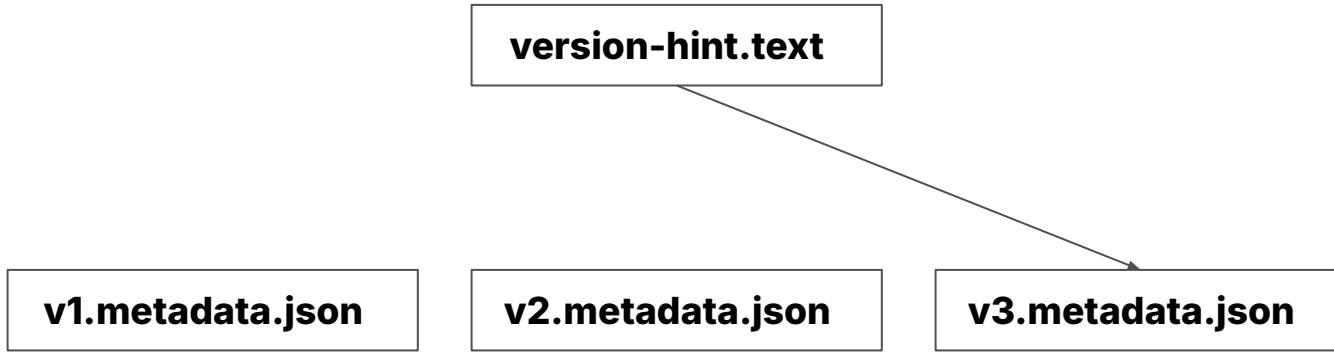
Role of the Iceberg Catalog



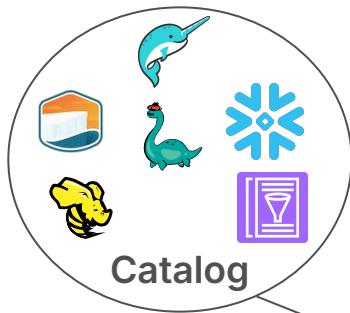
The Catalog Table Reference



The Catalog Table Reference - File System Catalog (Hadoop)



The Catalog Table Reference - Service Catalog



v1.metadata.json

v2.metadata.json

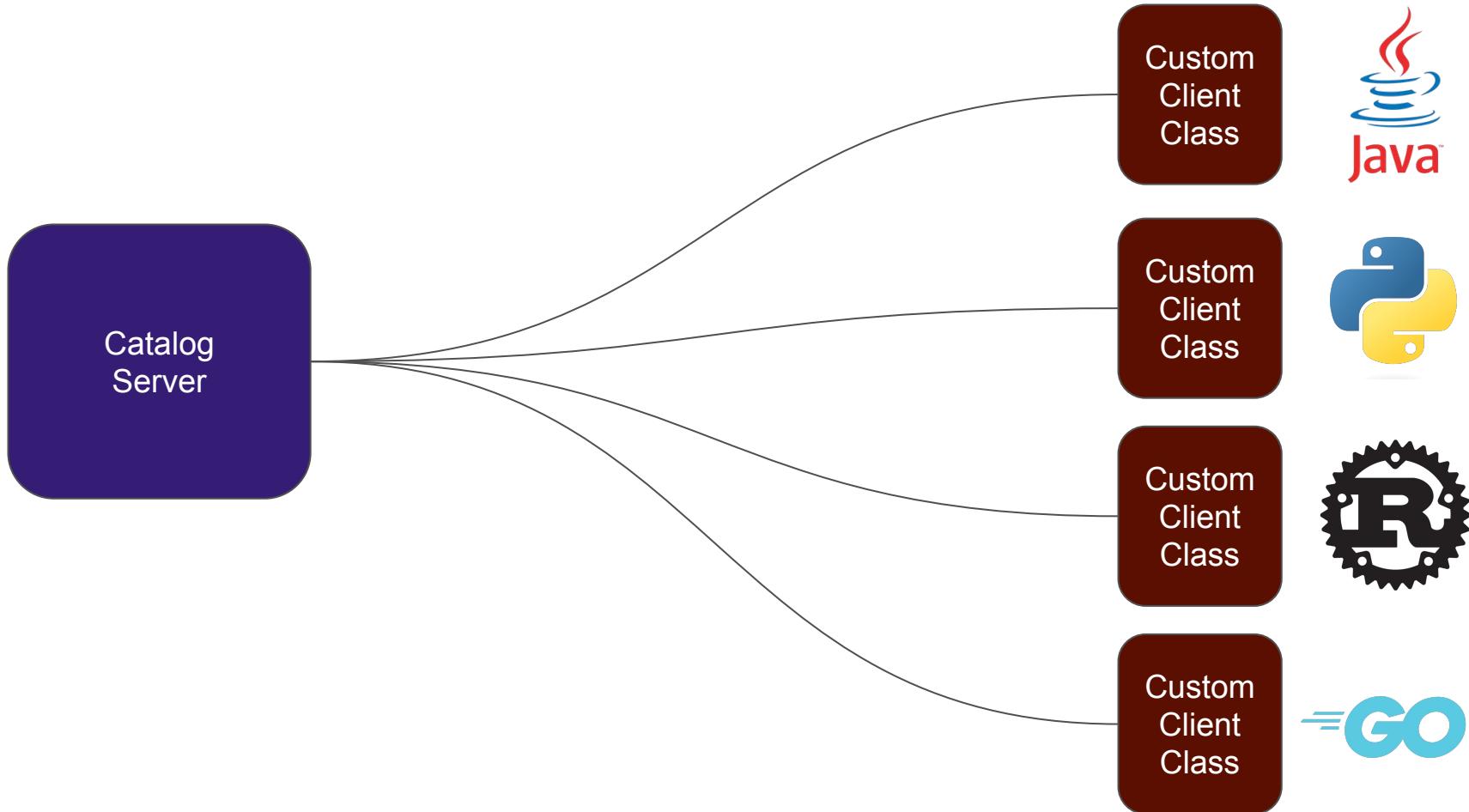
v3.metadata.json

The Apache Iceberg REST Catalog

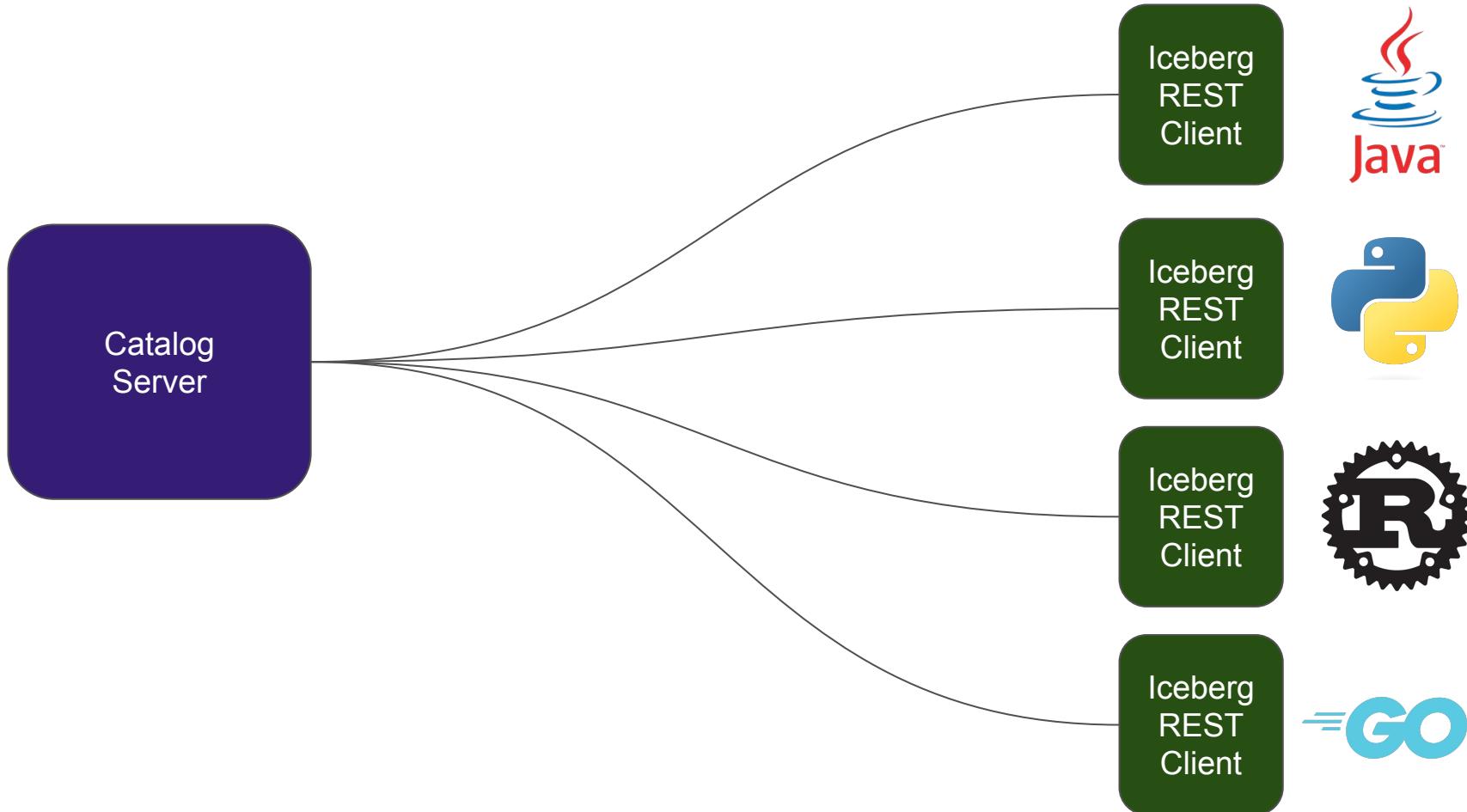
Problems

- Have to be re-implemented per language
- Engines would have to implement support for each catalog
- No control whether end-users are using most up to date version

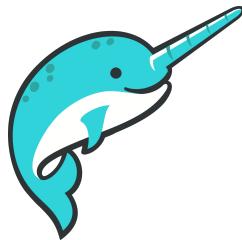
The Catalog Status Quo



The Apache Iceberg REST Specification



REST Catalog Client Support



DuckDB



Open Source Catalogs using REST CATALOG Spec



Announcing Support for the Iceberg REST Catalog Specification



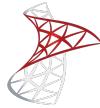
Any Engine That Supports REST Catalog Specification Can Connect to Any Catalog that Supports REST Catalog



Intro to Iceberg



Postgres -> Dashboard



SQLServer -> Dashboard



MongoDB -> Dashboard



dremio.com/blog