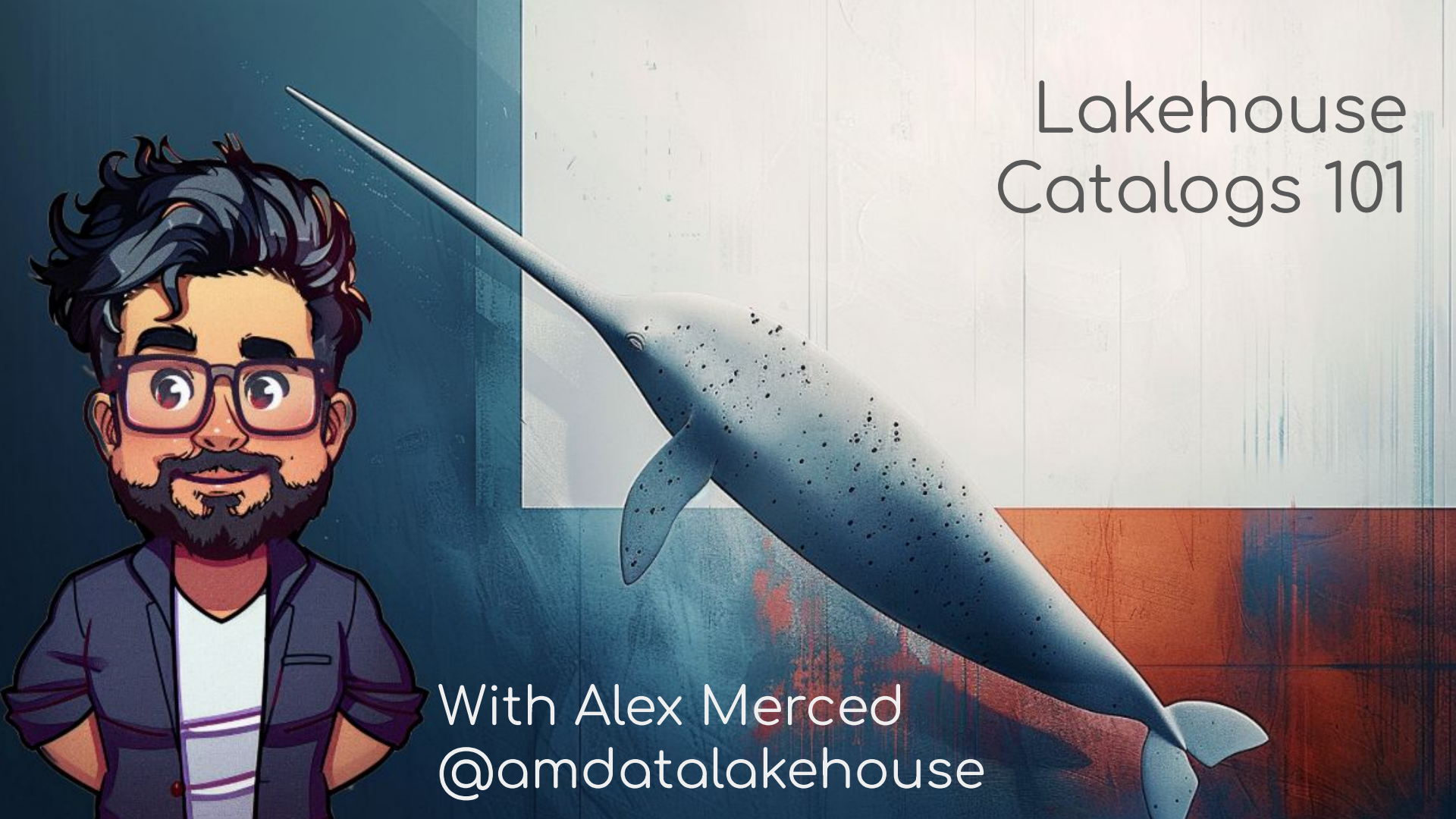
Lakehouse Catalogs 101



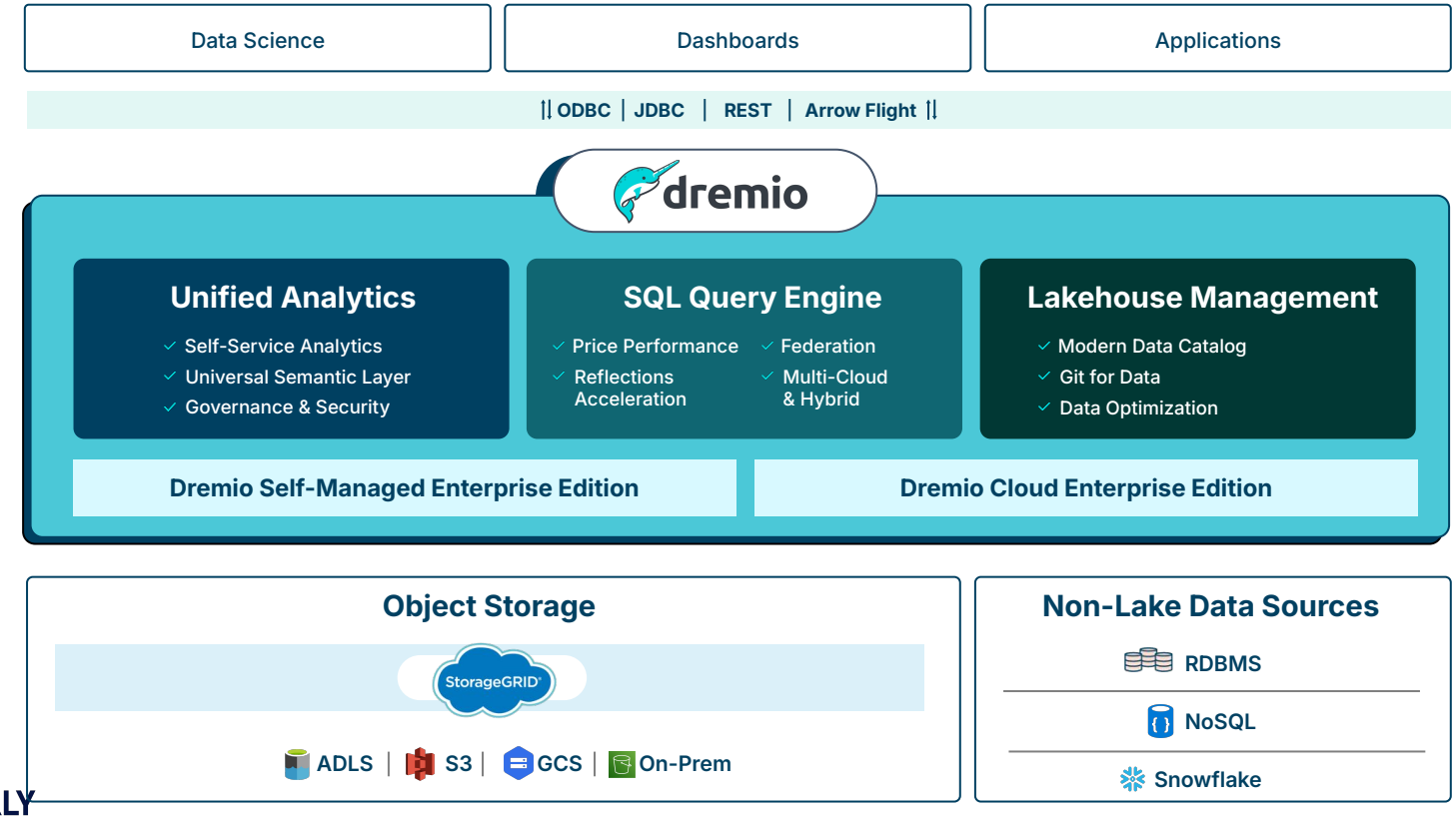
With Alex Merced
@amdatalakehouse

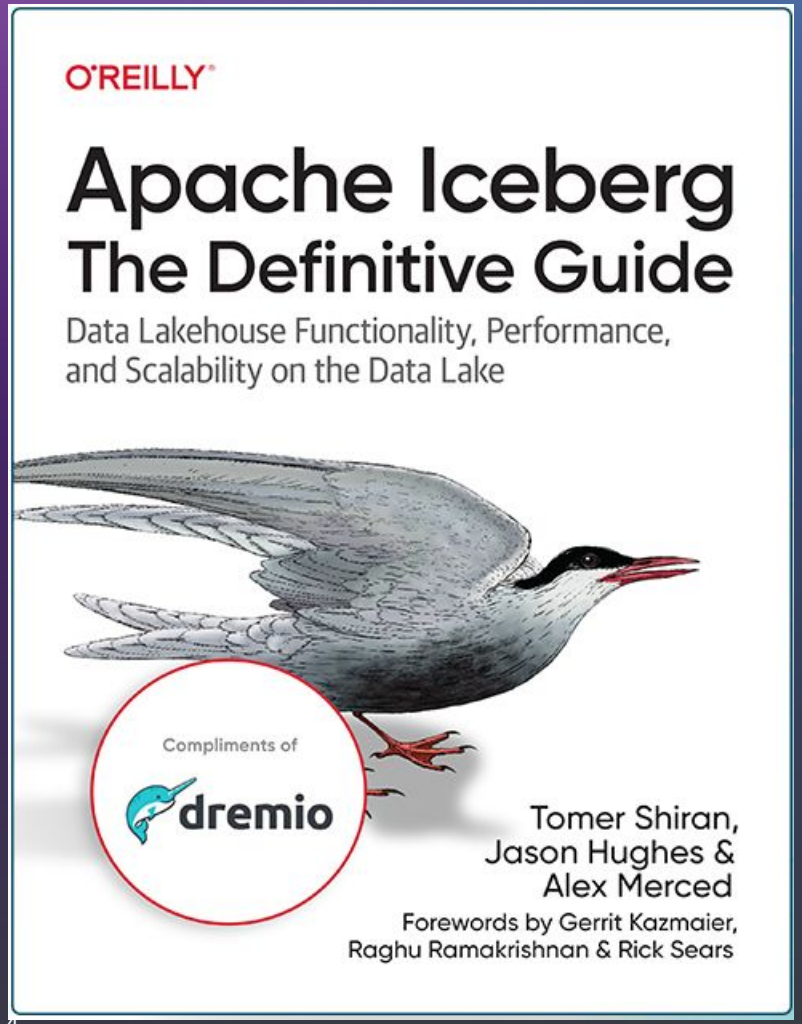
Alex Merced

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.

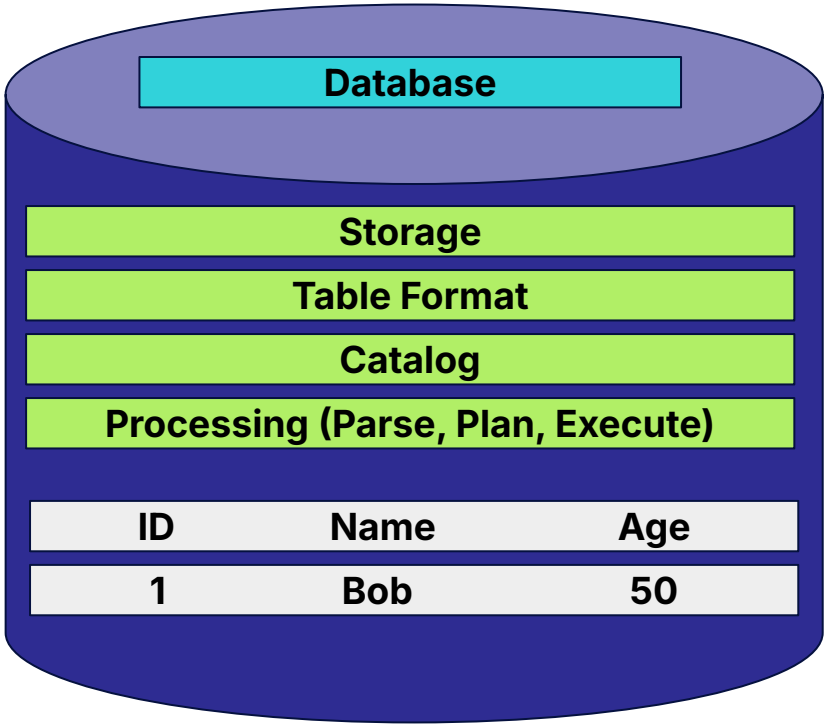
Dremio: The Unified Lakehouse Platform for Self-Service Analytics & AI

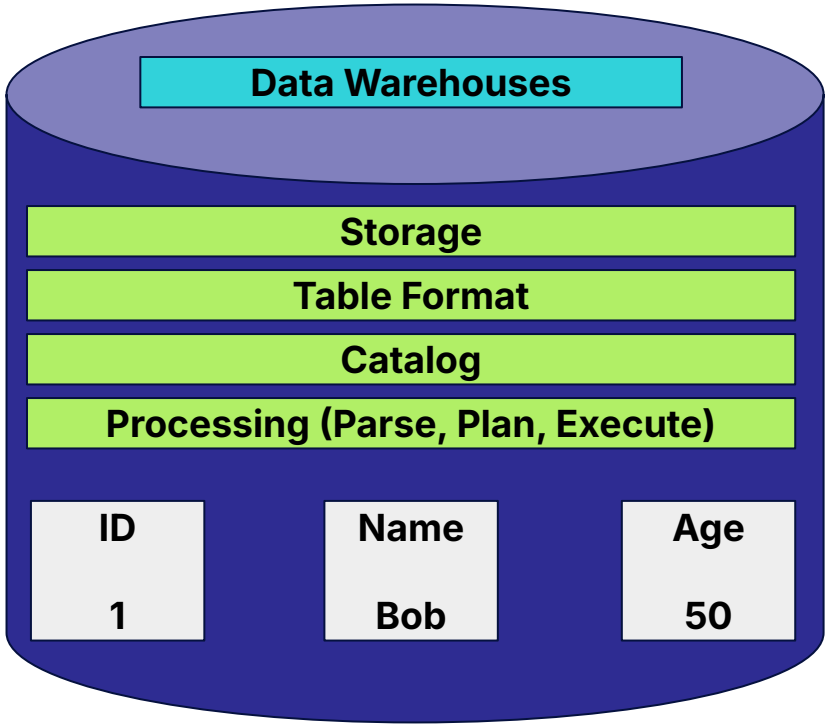



O'REILLY
Apache Iceberg
The Definitive Guide
Data Lakehouse Functionality, Performance, and Scalability on the Data Lake
Compliments of
dremio
Tomer Shiran, Jason Hughes & Alex Merced
Forewords by Gerrit Kazmaier, Raghu Ramakrishnan & Rick Sears

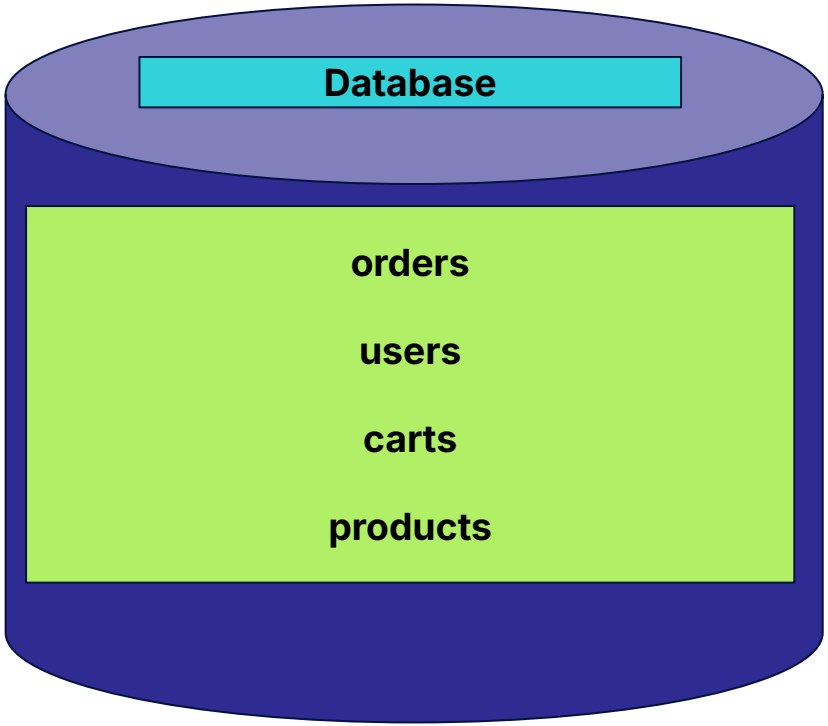
What is a catalog?

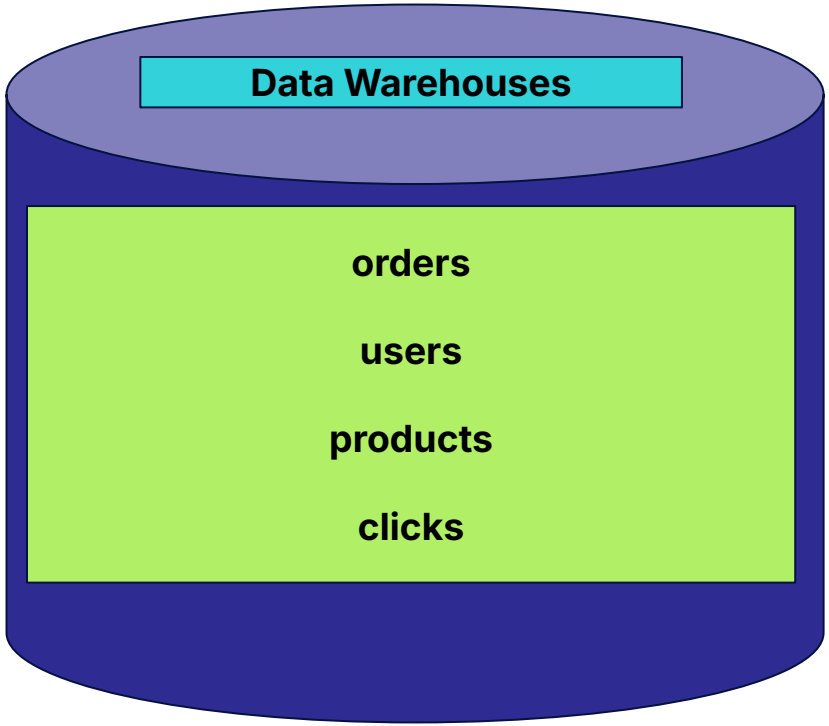
Traditional Data Systems



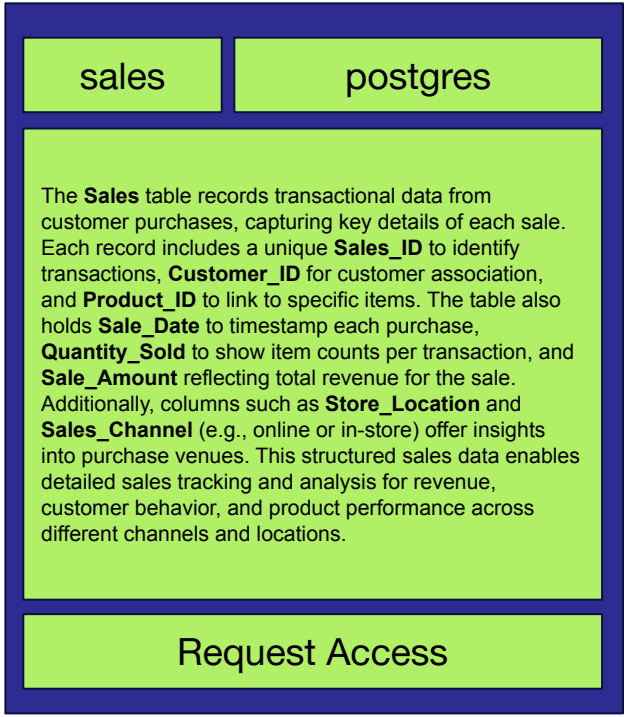


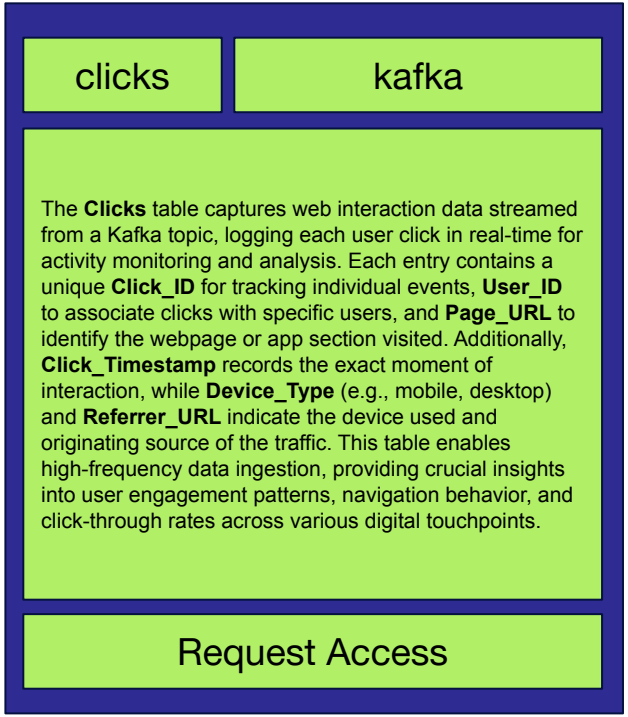
Traditional Data Systems





Enterprise Data Catalogs

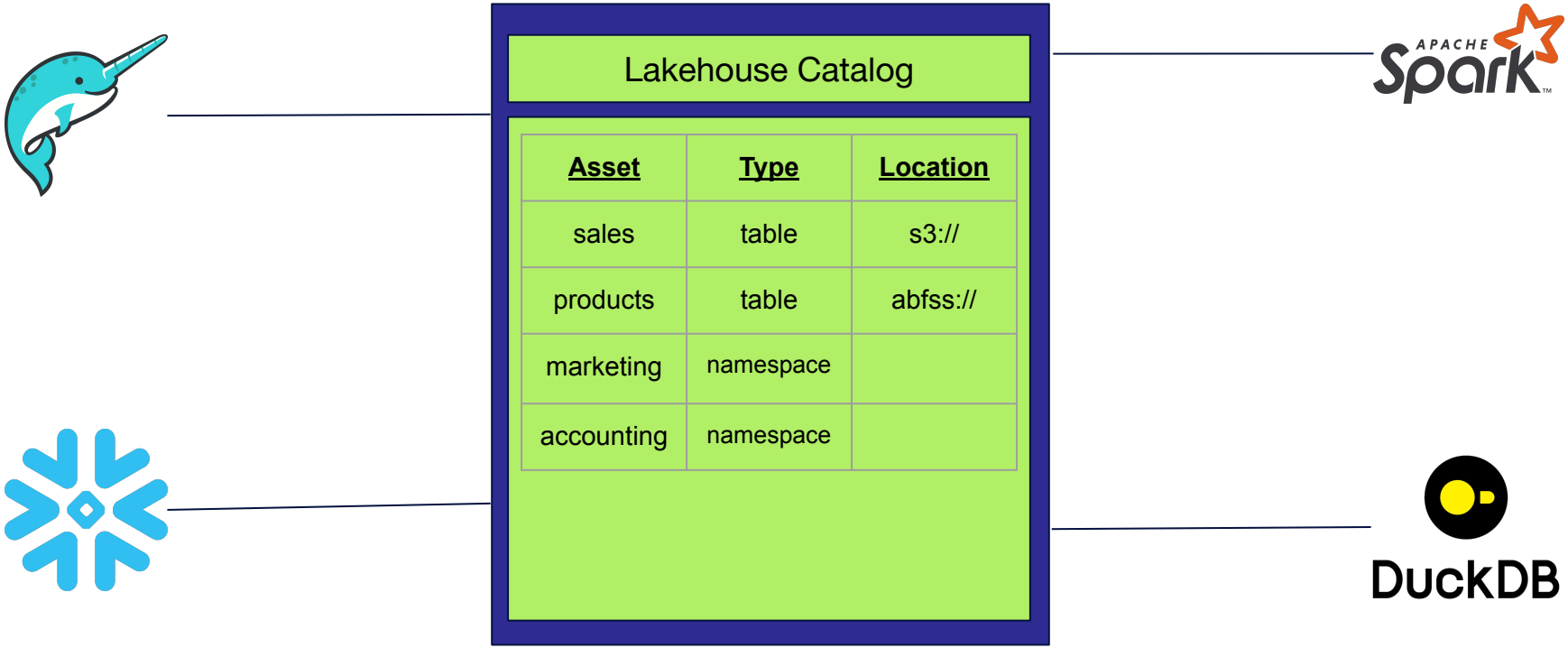




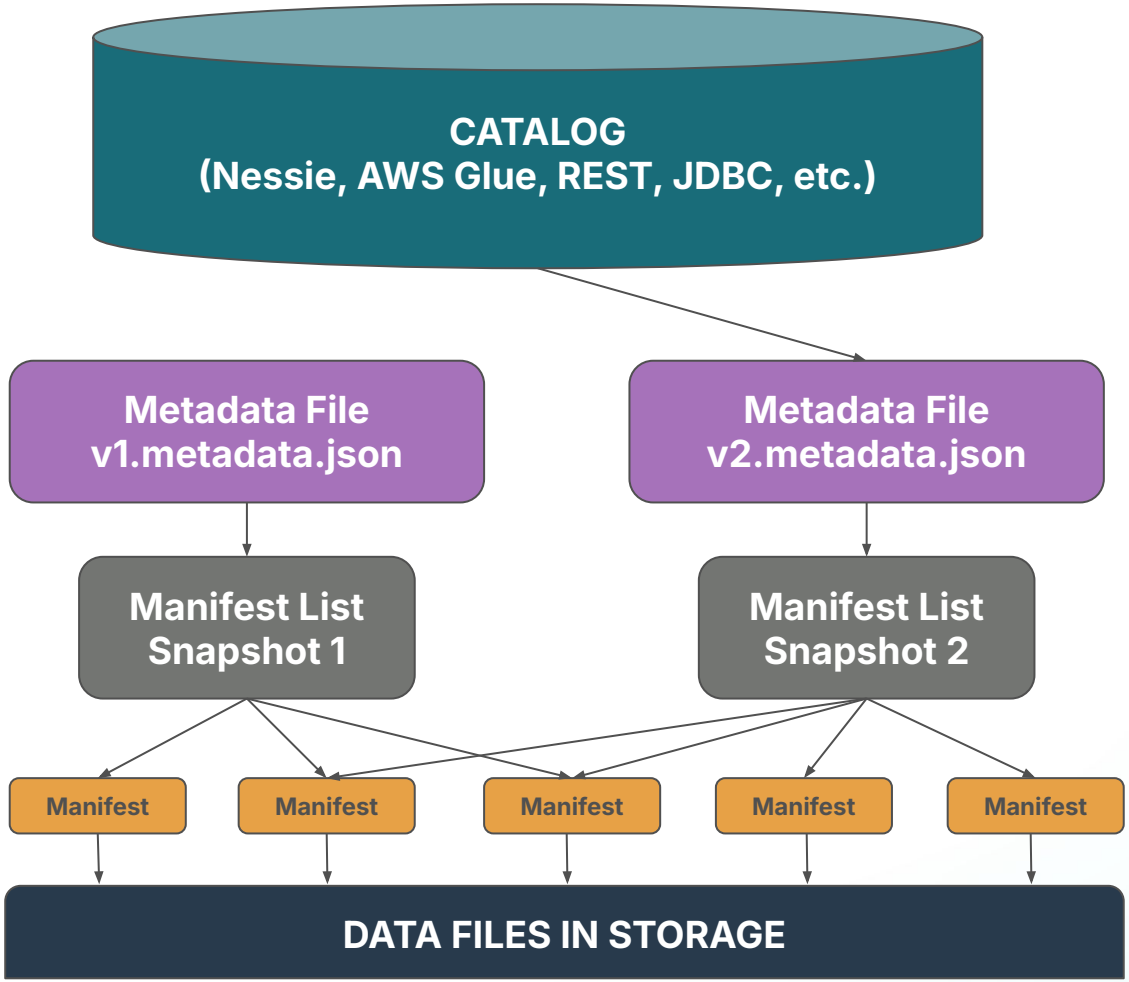
What is Data Lakehouse?

What is a Data Lakehouse?

Lakehouse Catalogs



What is Apache Iceberg?



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

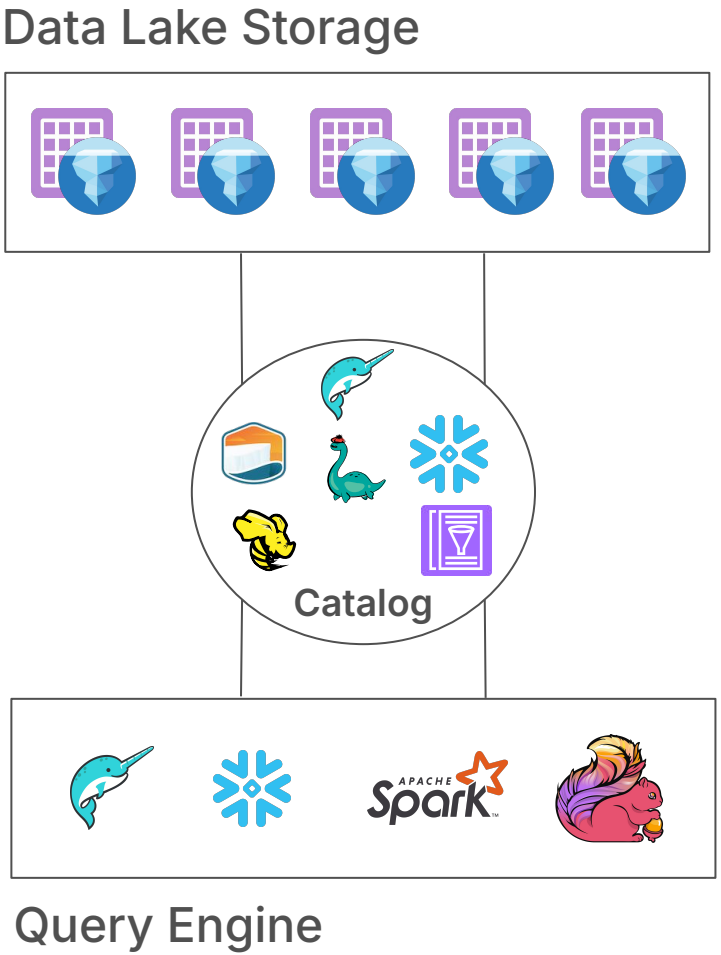
2. Metadata file is read for partitioning, schema and for location of manifest list of desired snapshot

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

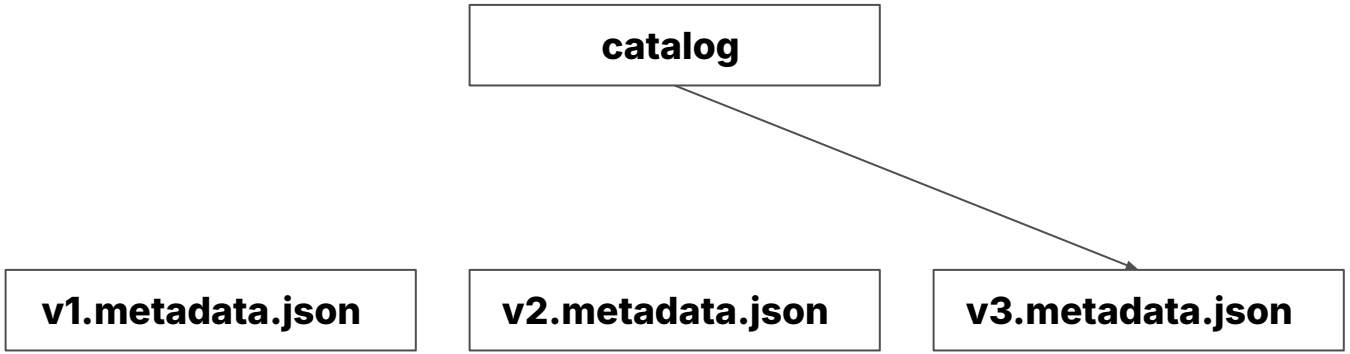
4. Manifests are using to do min/max filter to prune individual files, remaining files are then scanned for query.

Apache Iceberg Catalogs

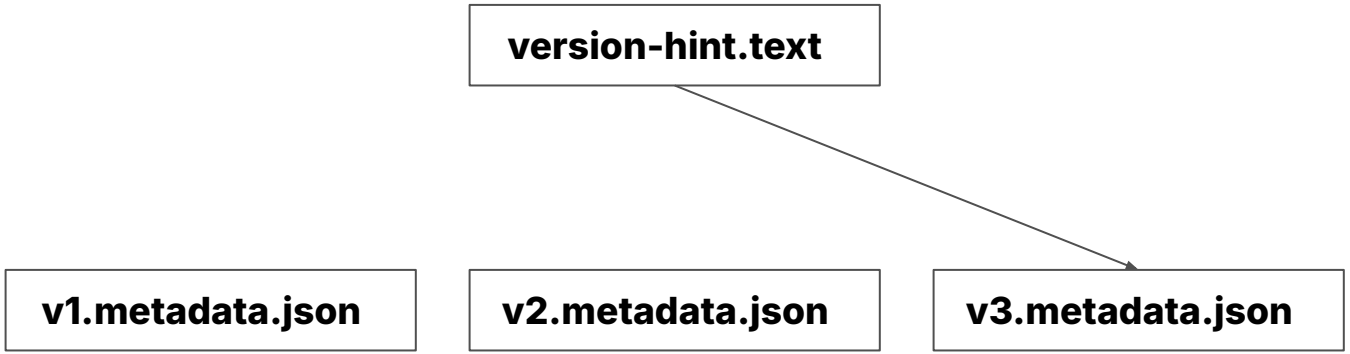
Role of the Iceberg Catalog



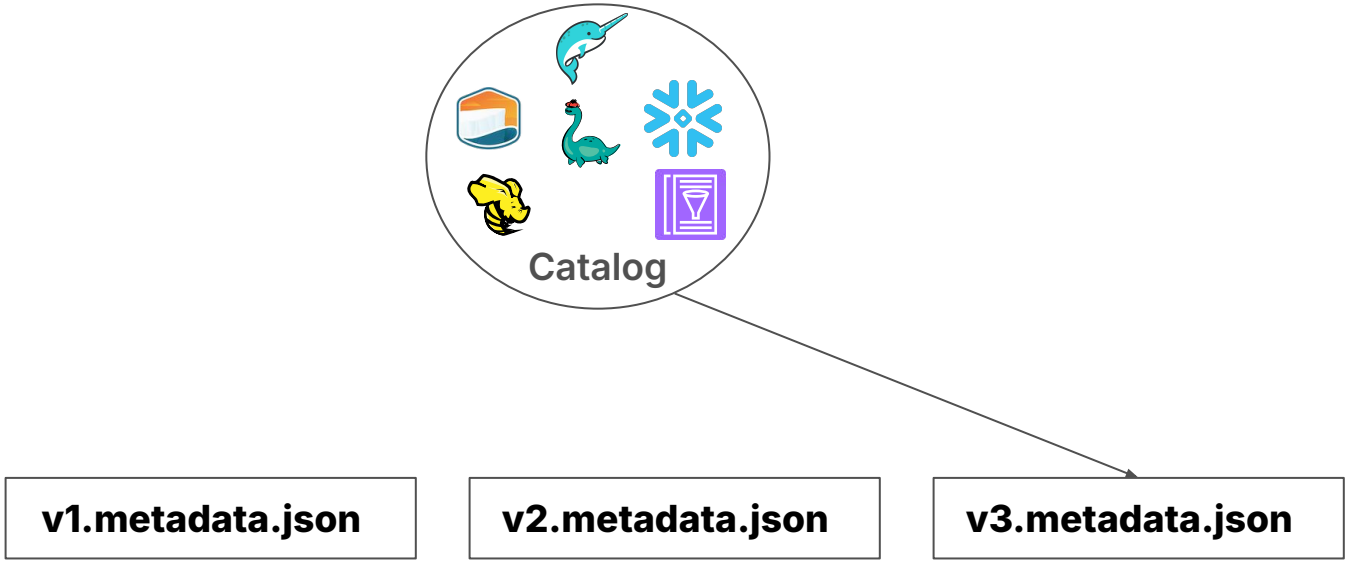
The Catalog Table Reference



The Catalog Table Reference - File System Catalog (Hadoop)



The Catalog Table Reference - Service Catalog

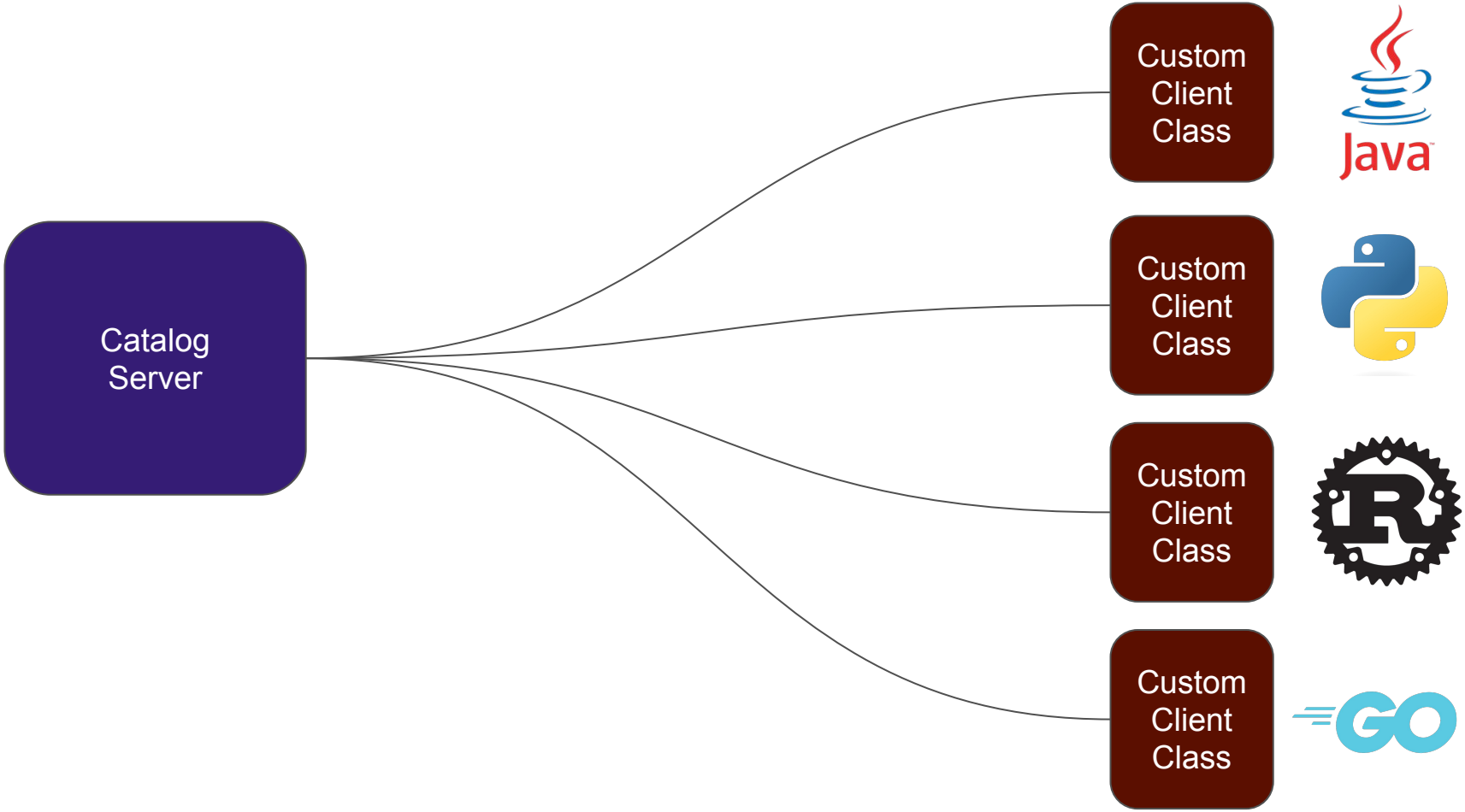


The Apache Iceberg REST Catalog

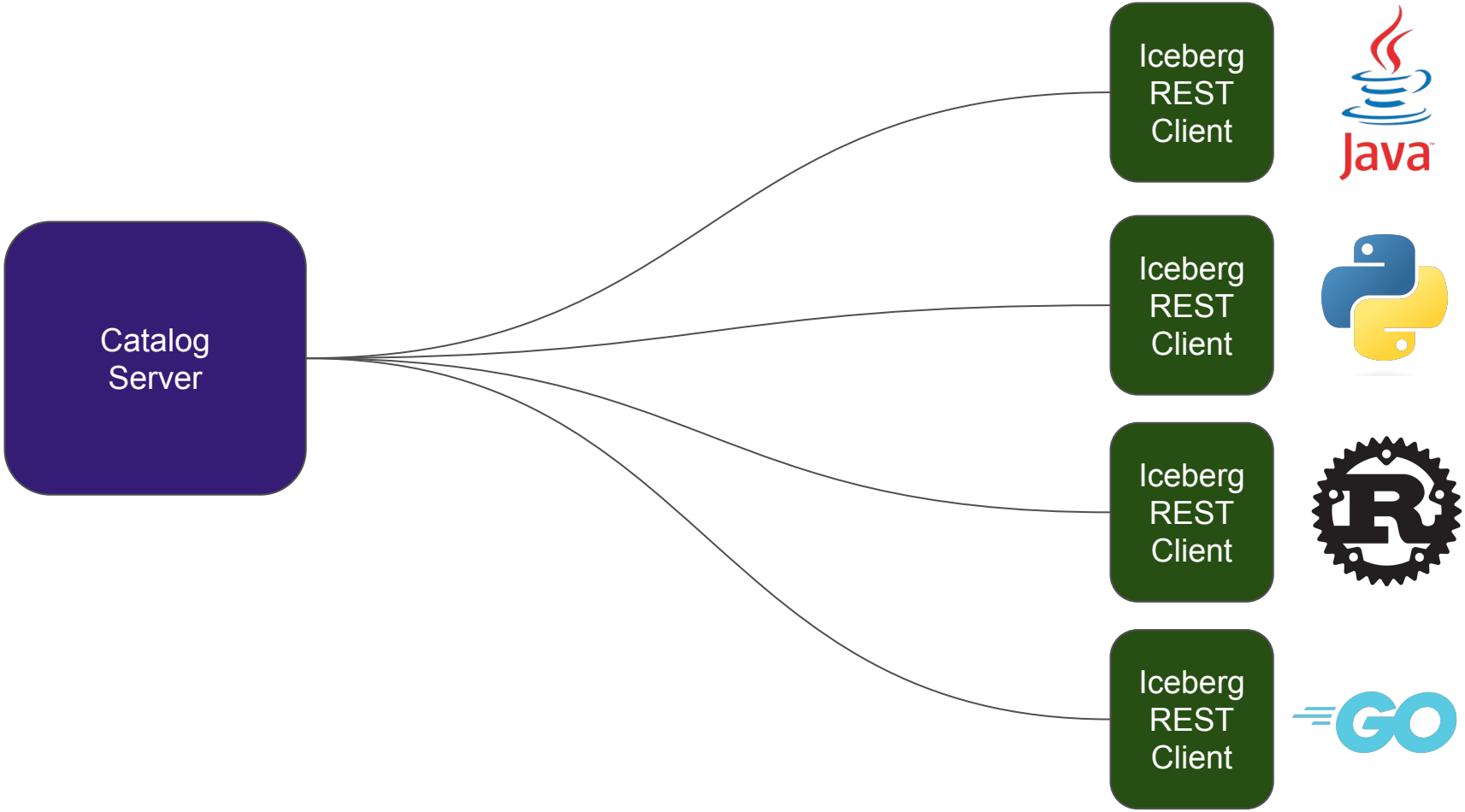
Problems

- 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

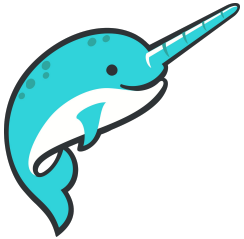
The Catalog Status Quo



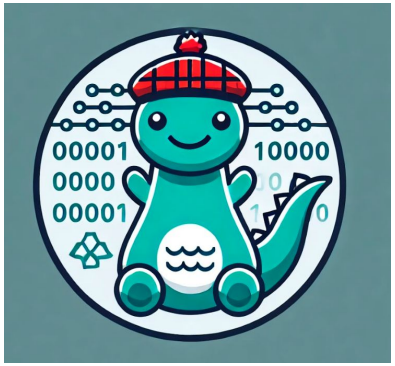
The Apache Iceberg REST Specification



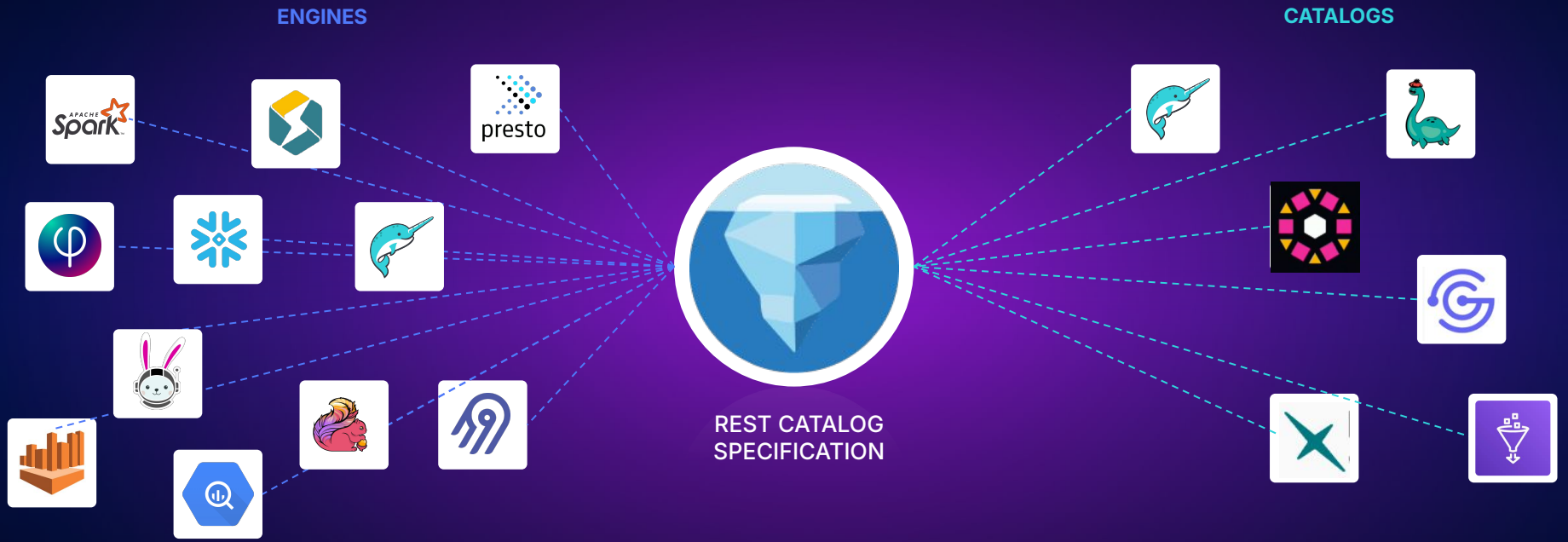
REST Catalog Client Support



Open Source Catalogs using REST CATALOG Spec



Announcing Support for the Iceberg REST Catalog Specification

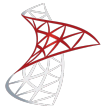


Any Engine That Supports REST Catalog Specification Can Connect to Any Catalog that Supports REST Catalog

Hands-On

Intro to Iceberg

Postgres -> Dashboard



SQLServer -> Dashboard

MongoDB -> Dashboard

dremio.com/blog