Building Lakehouse with Apache Doris & Iceberg

Apache Doris Webinar 20250410

Content

Table Format and Iceberg

Build Lakehouse with Doris and Iceberg

Live Demo

User Cases and Roadmap

Apache Doris Webinar 20250410

Table Format & Apache Iceberg
Why We Need Open Table Format?

Performance

Tabular projects are essentially the management of file collections.

- By recording the metadata of these files, the files can be classified, such as which files belong to the same table or partition. This avoids the cost of tools searching and classifying these files, making file queries more efficient and cost-effective.

Interoperability

Operations on data are essentially operations on files

- It can manage conflicts, record file change history, support different schema versions, etc. This ensures that multiple tools can safely operate and access the same data.

Openess

Table management does not make any requirements or assumptions about the storage system and computing engine.

- This allows users to easily use different storage and computing systems to meet their needs, or to upgrade and migrate. The openness of the table format provides better long-term scalability and adaptability for the file format.

Apache Doris Webinar 20250410

DORIS

Table Format & Apache Iceberg
Why Choose Iceberg?

Expressive SQL

- Iceberg supports flexible SQL commands to merge new data, update existing rows, and perform targeted deletes. Iceberg can eagerly rewrite data files for read performance, or it can use delete deltas for faster updates.

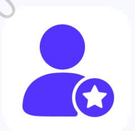
Full Schema Evolution

- Schema evolution just works. Adding a column won't bring back "zombie" data. Columns can be renamed and reordered. Best of all, schema changes never require rewriting your table.

Hidden Partitioning

- Iceberg handles the tedious and error-prone task of producing partition values for rows in a table and skips unnecessary partitions and files automatically. No extra filters are needed for fast queries, and table layout can be updated as data or queries change.

Time Travel and Rollback



- Time-travel enables reproducible queries that use exactly the same table snapshot, or lets users easily examine changes. Version rollback allows users to quickly correct problems by resetting tables to a good state.

Data Compaction

- Data compaction is supported out-of-the-box and you can choose from different rewrite strategies such as bin-packing or sorting to optimize file layout and size.

Apache Doris Webinar 20250410

DORIS

Table Format & Apache Iceberg
What is Iceberg Missing?

Efficient Data Maintainance

File layout directly affects query efficiency and storage cost

- Iceberg itself does not provide data management services. You need to customize data maintenance operations based on Iceberg's specifications, including small file merging, index management, snapshot management, etc.

Suitable Compute Engine

How to use SQL to access and manage iceberg tables efficiently

- Iceberg itself is not bound to a computing engine. Therefore, only a computing engine that is highly compatible with Iceberg can truly bring out the full potential of Iceberg.

Apache Doris Webinar 20250410

DORIS

Content

Table Format and Iceberg

Build Lakehouse with Doris and Iceberg

Live Demo

User Cases and Roadmap

Apache Doris Webinar 20250410

Build Lakehouse with Doris and Iceberg

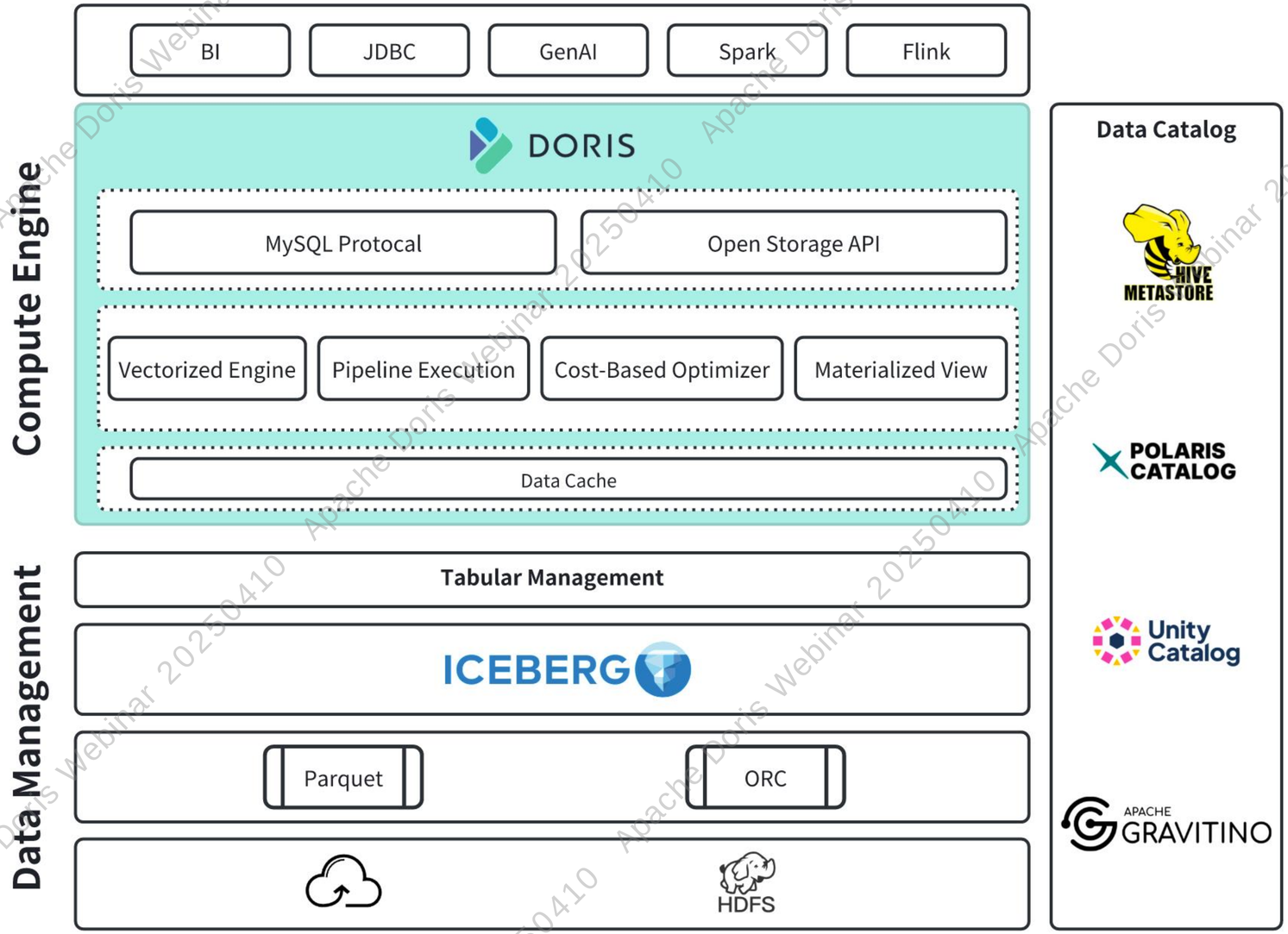
Apache Doris + Managed Iceberg Table

Compute Engine

- Apache Doris

Data Management

- Apache Amoro
- AWS S3 Tables



Apache Doris Webinar 20250410

Build Lakehouse with Doris and Iceberg

What is S3 Tables?

Store Iceberg data at scale in S3

Scalability

- Simplify data lakes at any scale, whether you're just getting started or managing thousands of tables in your Iceberg environment.

Fully managed

- Perform continual table maintenance tasks such as compaction, snapshot management, and unreferenced file removal to automatically optimize query efficiency and costs over time.

Enhanced performance

- Get up to 3x faster query performance through continual table optimization compared to unmanaged Iceberg tables, and up to 10x higher transactions per second compared to Iceberg tables stored in general purpose S3 buckets.

Simplified security

- Create tables as first-class AWS resources and apply permissions to easily govern access to them.

Apache Doris Webinar 20250410

DORIS

Build Lakehouse with Doris and Iceberg

Doris Iceberg Features

Fastest and Open-Source Lakehouse Compute & Query Engine on Iceberg

Query

High performance query engine

- C++ Implement table format reader
- Vectorized query engine
- Native position/equality delete support

Time Travel

Flexible version control

- View snapshot metadata
- Query with snapshot

Write Back

Data management with single engine

- Insert Iceberg table using Doris SQL
- Auto hanlde partition data skew

Apache Doris Webinar 20250410

DORIS

Content

Table Format and Iceberg

Build Lakehouse with Doris and Iceberg

Live Demo

User Cases and Roadmap

Apache Doris Webinar 20250410

Live Demo

Doris & S3 Tables Live Demo

Components

- Apache Doris: Query Engine for Iceberg
- AWS S3 Tables: Managed Iceberg on Cloud
- SelectDB Studio for Apache Doris: The Desktop GUI for Apache Doris

Demo

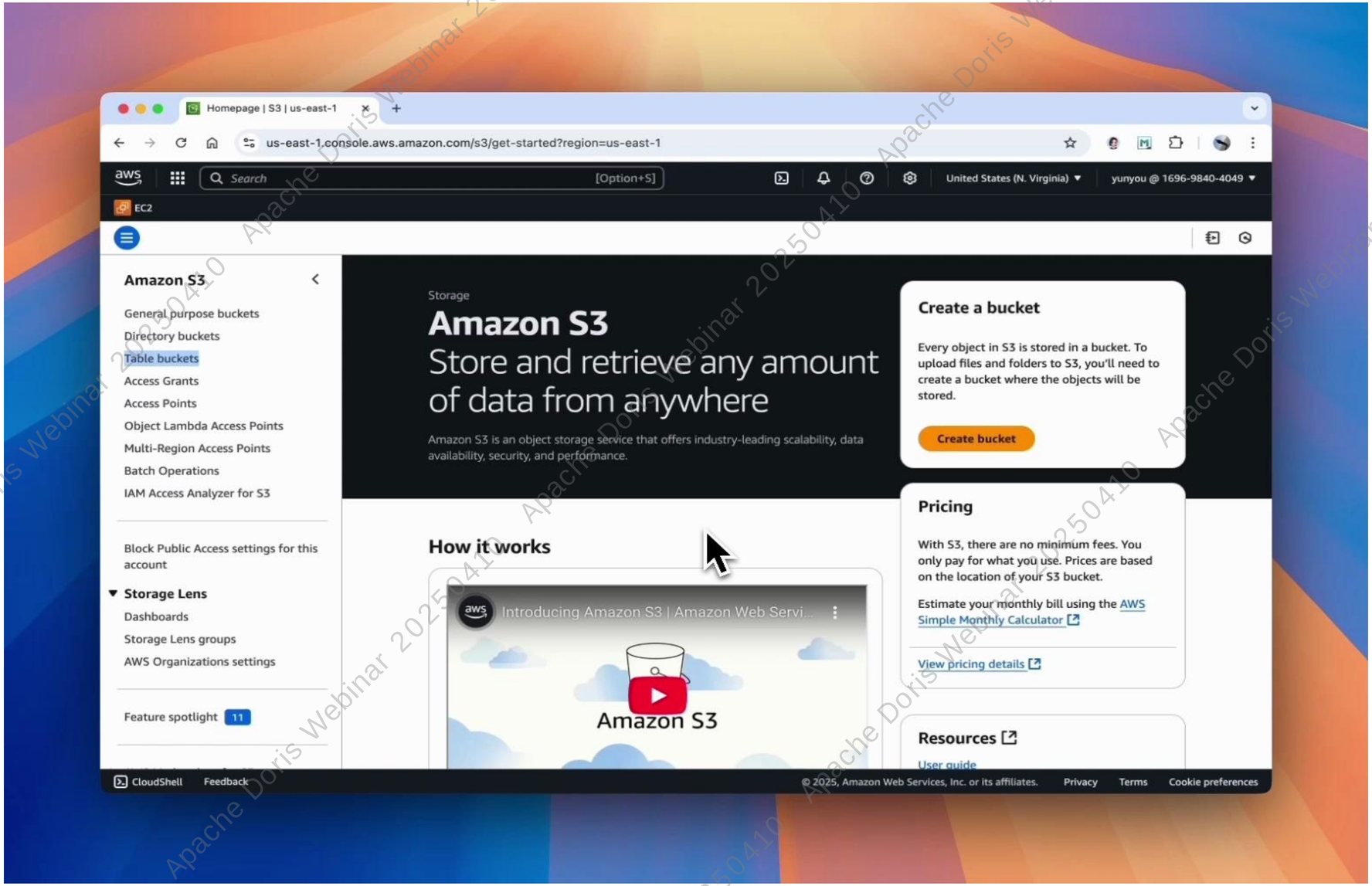
- Catalog Integration: Connecting to AWS S3 Tables
- Data Querying: Querying and time traveling on Iceberg using Doris.
- Table Management: Creating Iceberg table using Doris.
- Data Manipulation: Generating and Inserting TPCH benchmark data to Iceberg using Doris

Apache Doris Webinar 20250410

DORIS

Live Demo

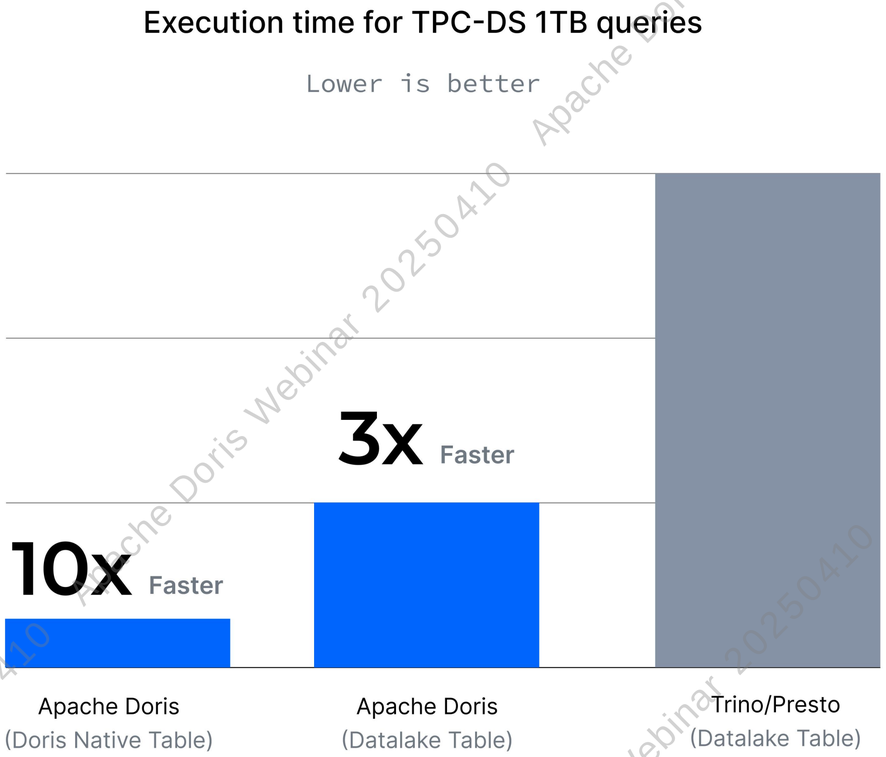
Doris & S3 Tables Live Demo

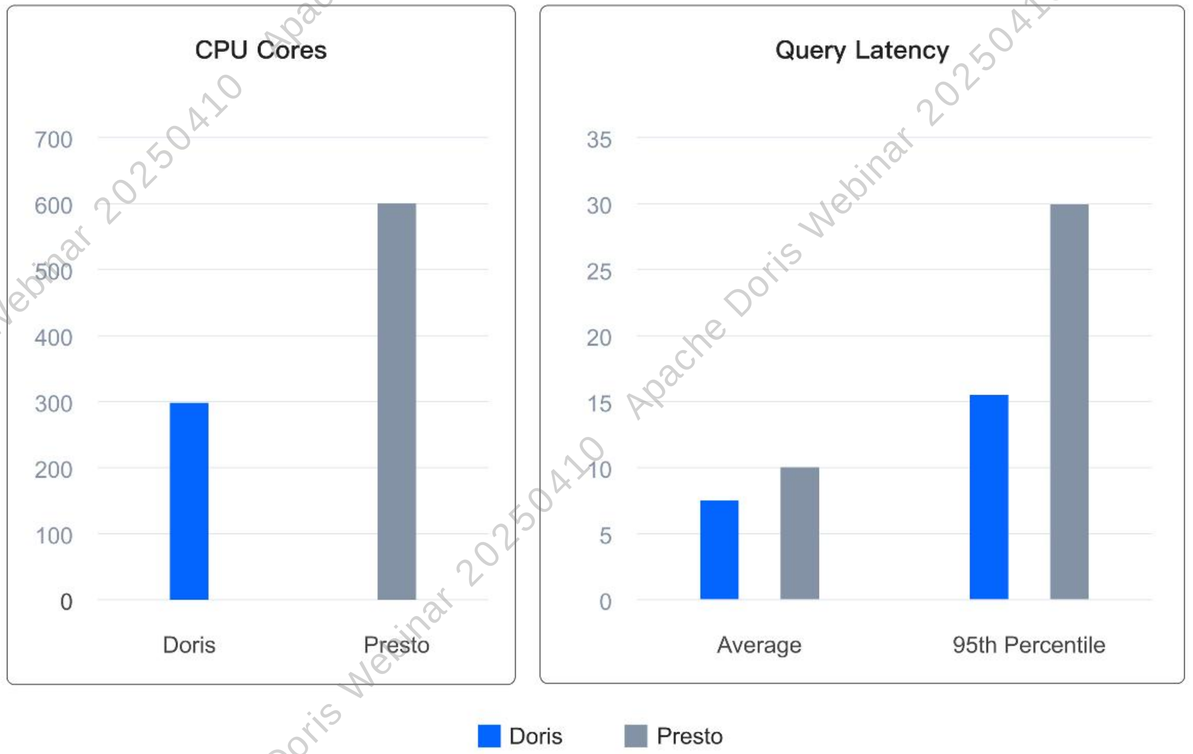


Apache Doris Webinar 20250410

DORIS

Live Demo
Performace





Apache Doris Webinar 20250410

Content

Table Format and Iceberg

Build Lakehouse with Doris and Iceberg

Live Demo

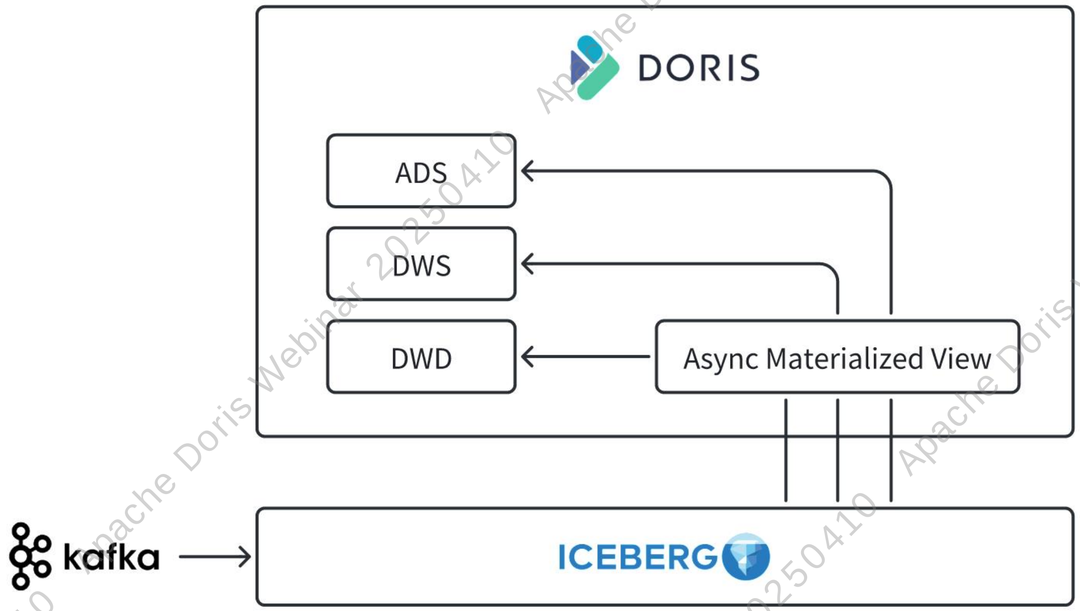
User Cases and Roadmap

Apache Doris Webinar 20250410

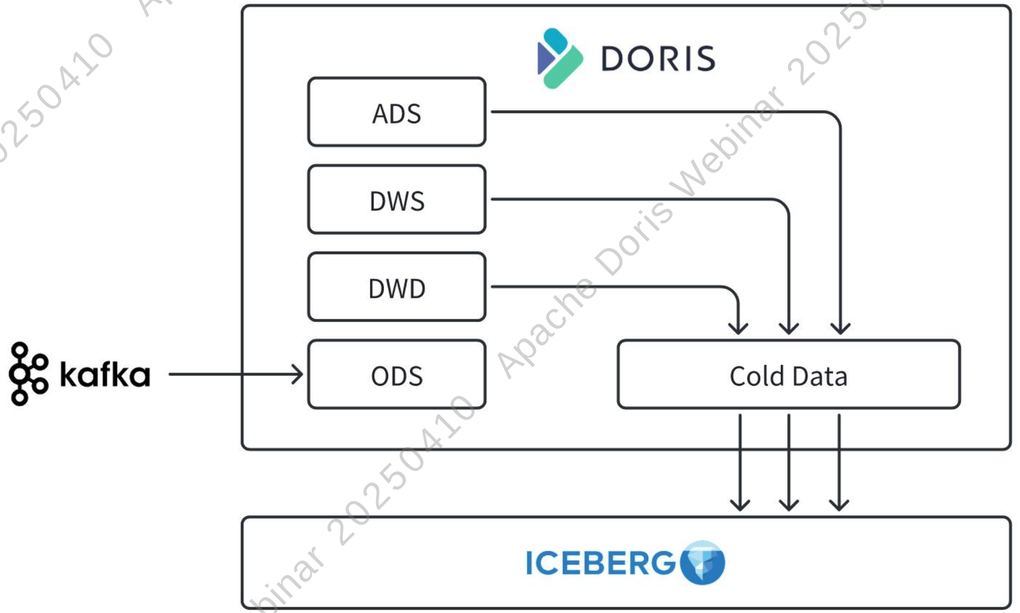
User Cases and Roadmap

Query Acceleration & Tiered Storage

網易 NETEASE



- Data is loaded into Iceberg first
- Hierarchical modeling in Doris using materialized View
- Query acceleration



- Data is loaded into Doris first
- Hierarchical modeling in Doris
- Cold data is offloaded to Iceberg

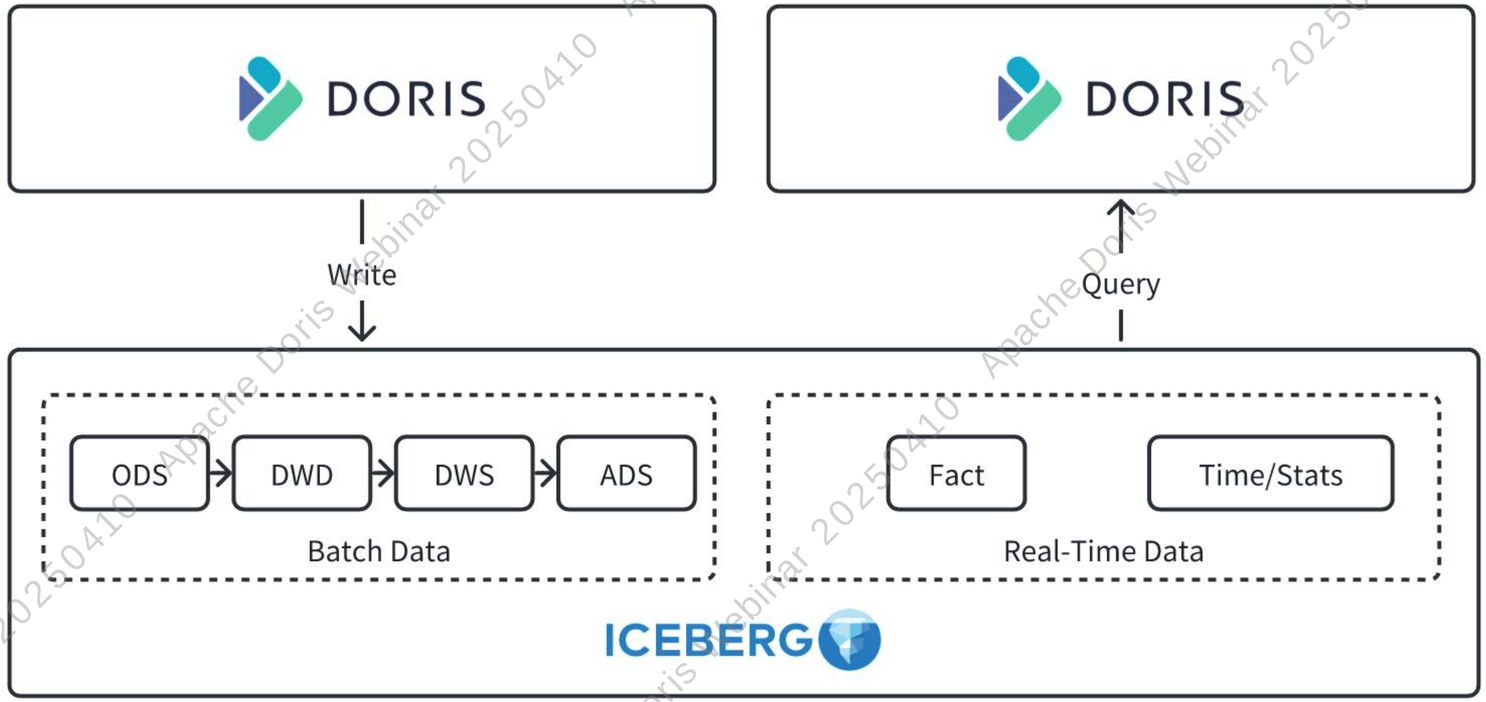
DORIS

Apache Doris Webinar 20250410

User Cases and Roadmap
Data Sharing

天翼云 Chinatelecom Cloud

- Iceberg data as unified storage
- Sharing data among diff Doris clusters



Apache Doris Webinar 20250410

DORIS

User Cases and Roadmap

Roadmap

More complete Iceberg operation capabilities

- Data compaction(Data rewrite)
- Branch/Tag support
- Various procedures support

More new data types

- Variant
- Geo Types

New Iceberg spec

- Iceberg format V3
- Binary deletion vection support

Iceberg REST Catalog API

- Unity Catalog
- Apache Polaris
- Apache Gravitino

Apache Doris Webinar 20250410

DORIS

Welcome to Doris Community

Subscribe

Subscribe to our mailing list and join our discussion:

dev@doris.apache.org

Get technical support

- **Slack**: apachedoriscommunity.slack.com
- **Wechat Group**: Scan the QR code on the right.

Apache Doris Webinar 20250410

Thanks !

Apache Doris Webinar 20250410