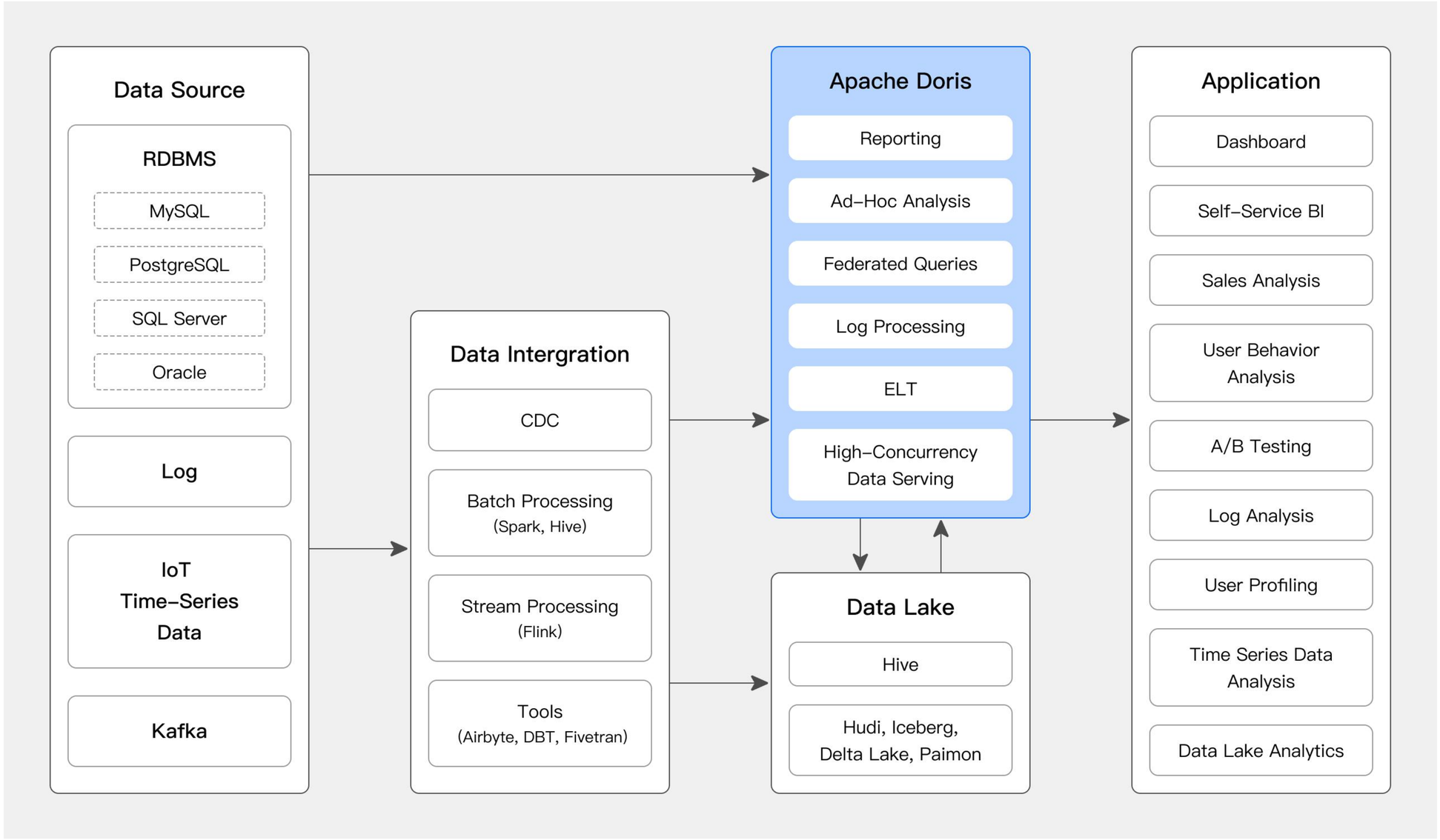
Apache Doris 2025 Roadmap Overview

What is Apache Doris?



DORIS

Contents

Doris 2024 Overview
Community Achievement

One of the world's most active open source communities in big data

Contributors	Releases	Commits	Change Lines
290	22	7K+	150K+

DORIS

Doris 2024 Overview
Things We've Done

Query Optimization

- Smart Optimizer
- Statistics
- Adaptive Scan Parallelism
- Local Shuffle
- Async Materialized View
- Optimization on ARM
- …

Storage Optimization

- Auto Incremental Column
- High Concurrency Import
- Multi-Statement Transaction
- CCR - cross cluster replication

Semi-Structured Data Analysis

- VARIANT Data Type
- IP Data Type
- More Analytic Functions
- …

Lakehouse

- Optimization on Iceberg/Paimon
- More Connectors
- SQL Dialect Support
- Arrow Flight SQL Protocol
- …

Cloud-Native Support

- Stateless Backends
- Multi Compute Group
- Smart Data Cache
- …

- https://doris.apache.org/docs/releasenotes/v2.0/release-2.0.0
- https://doris.apache.org/docs/releasenotes/v2.1/release-2.1.0
- https://doris.apache.org/docs/releasenotes/v3.0/release-3.0.0

DORIS

Contents

Doris 2025 Roadmap
Overview

Typical Use Cases

- Real-Time Analysis
- Lakehouse
- Semi-Structured Data Analysis

Stability

- Release Management
- Code Review and Approval Rules
- More Tests

Community

- Community Collaboration
- Community Support

Innovation

- GenAI & ML
- Batch Processing
- Incremental Processing

DORIS

Doris 2025 Roadmap Overview

Typical Use Cases

Real-Time Analysis

Becoming the fastest and most cost-effective analytical database

- Improving performance under x86 and ARM architectures
- Improving optimizer capabilities (CBO/RBO/HBO/AIBO)
- Optimization for Wide Tables with 10K+ Columns

Data Lakehouse

Solving unified data management, data sharing and high-performance data processing

- Query acceleration on open lake format
- Unified SQL gateway for multiple data sources
- Full-featured open lake format management

Semi-Structured Data Analysis

From Log to Observability

- Inverted index in production of PB scale
- Advanced features for VARIANT
- Ecosystem integration beyond Grafana, OpenTelemetry, Logstash and Filebeat

DORIS

Doris 2025 Roadmap Overview
Stability

Release Management

How to release
stable and latest version

- **2.1 & 3.0:** Stable version.
- **3.1:** Stable version with necessary new features and optimization.
- **4.0:** Data for AI

Code Review Rules

Make code review
easier, rigorous, and enforceable

- Pull request description
- Unit test coverage
- Code owner

More Test

More test scenarios

- Regression Tests
- Unit Tests
- Chaos Tests
- Stress Tests

DORIS

Doris 2025 Roadmap Overview
Community

Community Collaboration

Making community collaboration more open and efficient

- Doris Improvement Proposal
- Special Interest Group
- More deep dive articles
- More webinars

Community Support

Making community support smarter and more sustainable

- High-quality documentation
- Forum Construction
- Doris Expert AI Model

DORIS

Doris 2025 Roadmap Overview

Innovation

GenAI & ML

Data Infrastructure in the GenAI Era
DB for AI & AI for DB

- High-Throughput Data API Based on Arrow Flight (done)
- Vector semantic search
- Data Preparation & Feature Store
- Lakehouse Integration
- ChatBI & Agent

Batch Processing

How to run large query
with limited resource

- Spill to Disk (Done)
- Stage by Stage scheduler
- Mixed load management between real-time and batch process tasks

Incremental Processing

Making data refresher

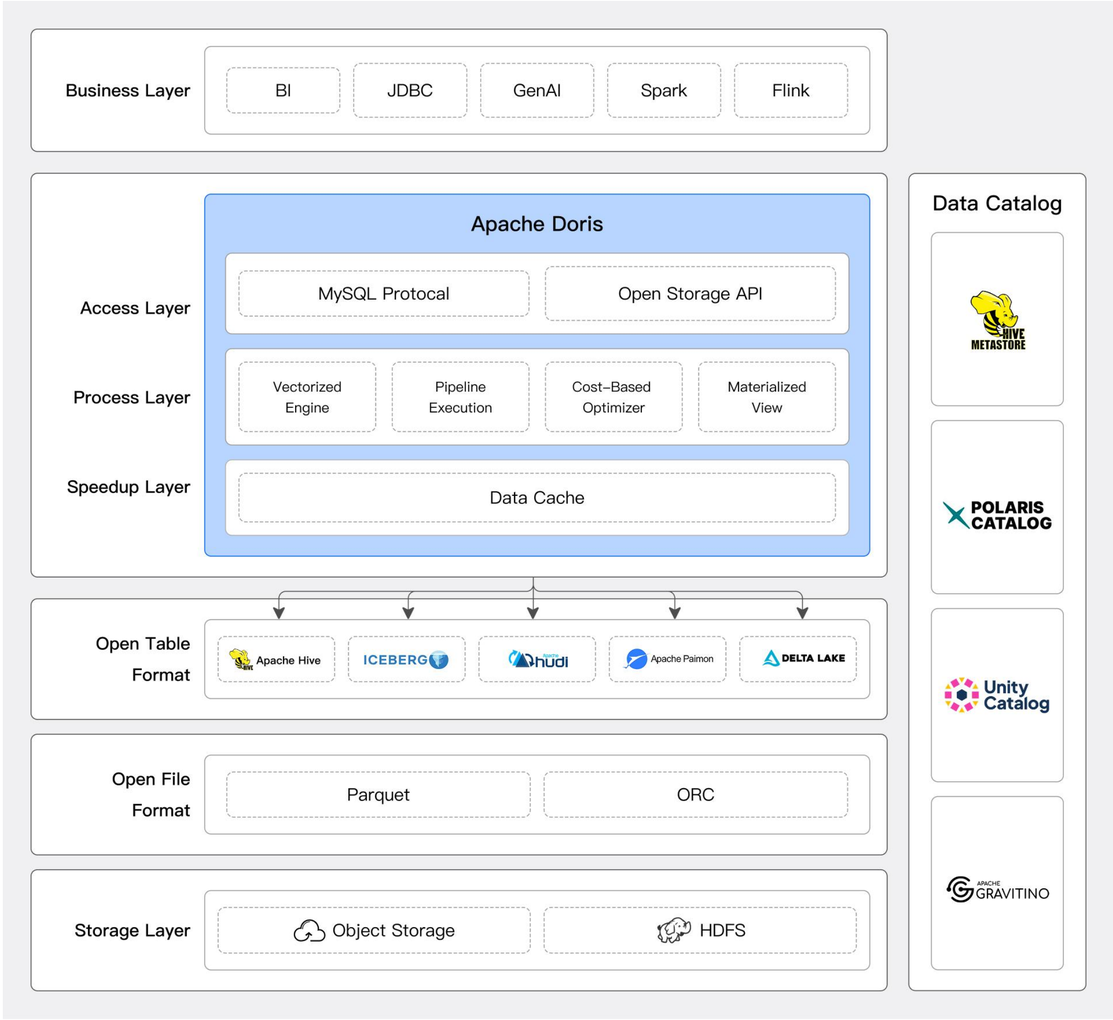
- Binlog Publishing and Subscription
- Realtime Materialized View

DORIS

Doris 2025 Roadmap

Data Lakehouse

Query Acceleration, Federation, Data Lake Processing



Enhancing Query Acceleration on Data Lake for Greater Stability and Efficiency

- Enhance the predictability of remote I/O latency and reduce long-tail query delays. This includes I/O scheduling, more intelligent I/O merging strategies, and data caching strategies.
- Address the efficiency and resource utilization challenges in scenarios with large volumes of metadata. This includes distributed query planning and more intelligent metadata file caching.
- Exploring the indexing capabilities of open lake formats and leveraging external indexes to accelerate data access.
- Enhance the multi-SQL dialect compatibility of the SQL convertor to help migration from Trino, Presto, Hive, PostgreSQL, and ClickHouse etc.

DORIS

Doris 2025 Roadmap

Data Lakehouse

Query Acceleration, Federation, Data Lake Processing

Fully Embracing Iceberg and Paimon for a More Comprehensive User Experience

- A more comprehensive lakehouse operation experience: support for DDL and DML.
- Write-Back to LakeHouse
- Support for new features: new column types (geo, variant), new data formats (Iceberg v3)
- Smoother cross-data source integration: support for Iceberg Rest Catalog, Snowflake, Databricks, and S3 Tables.

Enhanced Asynchronous Materialized Views for Seamless Data Integration

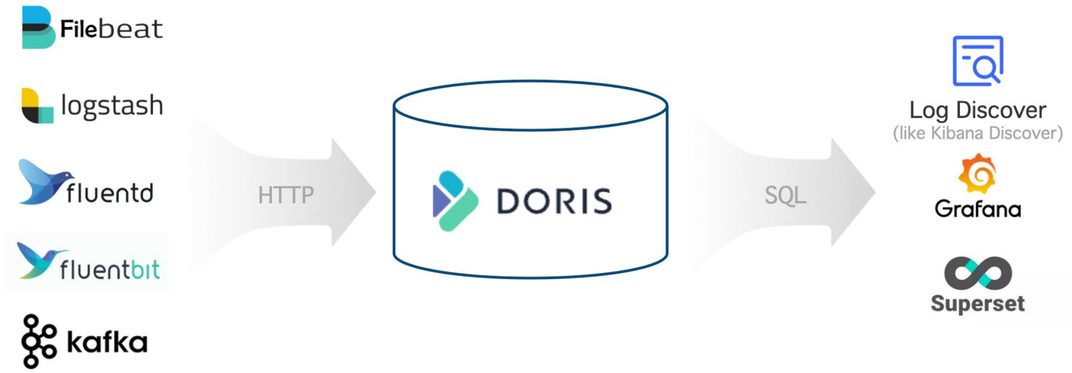
- Provide partition-level incremental build capabilities for materialized views in Iceberg, Paimon and Hudi.
- Support for converting between logical views and materialized views, offering greater flexibility for data modeling.
- Added capability to expose data lineage information for materialized views.
- Intelligent operations and maintenance, including features such as intelligent recommendations and automatic creation, automatic merging, automatic deletion, and automatic adjustment of build cycles.

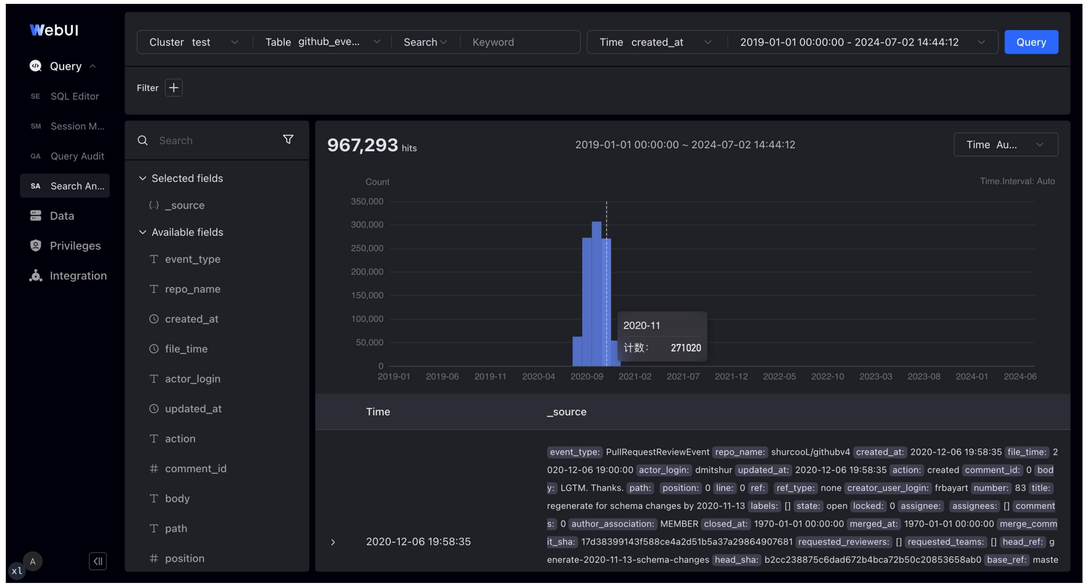
DORIS

Doris 2025 Roadmap

Semi-Structured Data Analysis

Inverted Index, VARIANT, Observability





SOTA inverted index in production of PB scale

- Support more tokenizers: Chinese ik, Unicode icu tokenizer, High-performance simple tokenizer for log scenarios, and custom dictionary management for tokenizers.
- Support custom dictionary and management for tokenizers.
- Support incremental index building in compute-storage decoupled mode.
- Further optimize inverted index space usage.
- Enhanced index observability, including write and query performance metrics.

DORIS

Doris 2025 Roadmap

Semi-Structured Data Analysis

Inverted Index, VARIANT, Observability

Advanced Semi-structured Data Type VARIANT

- Supports 10,000 sub-columns in compute-storage decoupled architecture.
- Sparse columns support more sparse sub-columns
- Supports complex structure expansion of JSON array nested objects
- Supports specifying sub-column types
- Supports building indexes for specified fields

Log and Observability Ecosystem Improvement

- Observability ecosystem integration: Opentelemetry, Jeager
- Support more log collector plugins: ilogtail, vector
- Output plugin supports writing to multiple tables: filebeat, logstash

DORIS

Doris 2025 Roadmap

Query Execution

Execution Adaptivity, Universality, Resource Management

Performance Optimization in Complex Scenarios

- Automatically detect and adapt to data skew scenarios: upon detecting data skew, utilize the new data skew handling capabilities of the execution engine to automatically rewrite plans and improve execution efficiency.
- ARM architecture tuning and optimization: Support more hardware architectures, improve operational efficiency.
- More general top-n and global lazy-materialization ability.
- Global dict.

Enhanced Resource Management for Stability and Observability

- Unified resource management framework for resource auditing and observability for query, load, compaction, schema change.
- Provide real-time resource monitor system tables and metrics for all tasks.
- Unify resource control logics such as Workload Group Policy, Spill-to-Disk, Query Breaker.
- More smarter scheduling algorithm to allocate resource between multi queries in a single workload group to reduce affect between big queries and small queries.
- Enhance mix-load memory management.

DORIS

Doris 2025 Roadmap
Query Optimizer

Plan Performance, Quality & Operability

Enhancing the quality and stability of query planning in complex scenarios.

- Optimize simple queries to better handle high-concurrency query scenarios.
- Further improve the efficiency of Join planning to address more complex join query scenarios.
- Introduce HBO to enhance query planning accuracy and stability based on historical statistics.
- Enhance plan management capabilities by providing plan fixation and evolution to address stability issues caused by query plan fluctuations.
- Introduce partition-level statistics and histograms to better handle data skew scenarios. Optimize sampling algorithms to improve accuracy and execution efficiency.

Improving query planning observability

- Develop real-time diagnostics, execution path tracing, and plan capture tools to facilitate troubleshooting.
- Expose more internal real-time operational states to enhance maintenance and monitoring convenience.

DORIS

Doris 2025 Roadmap
Storage & Cloud-Native Support

Data Security, Easy ETL, Stability

Fault Tolerance & High Availability

- Cross-Cluster Replication (CCR) product ready in cloud mode
- CCR support master-standby switching

Security

- Support storage encryption
- IAM Role

Enhancing Compute-Storage Decoupled

- Improving cold data query performance
- Enhanced data caching strategy
- Enhanced read-write isolation

Enhancing ETL capability

- Support temporary table
- Support write-write confliction detection in multi-statement transaction

DORIS

Welcome to Doris Community

Subscribe

- Mailing list: dev@doris.apache.org

Get technical support

- Slack: apachedoriscommunity.slack.com

Follow us on:

- X: https://x.com/doris_apache
- Linkedin: https://www.linkedin.com/company/doris-apache/
- Youtube: https://www.youtube.com/@apachedoris

DORIS