## Apache Doris for Al & 2025 RoadMap

Kang Xiao

Apache Doris PMC Member, SelectDB VP



Meetur



## Apache Doris: Open-Source Real-Time Analytics Database



🔊 mato	omo				Dashboard	All Websites	English 🗸
Q Search	APRIL 2024 🛱	2*					
	All Websites dash	board (Total: <b>5,037,193</b> visits, <b>24</b> ,	<b>577,438</b> pageviews, <b>2</b>	6,286,470 actions, 0 reve	nue)		
	WEBSITE	VISITS	▼ PAGEVIEWS	REVENUE	EVOLU	JTION Visits 🗸	
	Apache.org	920,826	9,096,548	\$0	<b>.</b> -4	.8%	
	Apache Doris	170,815	6,625,777	\$O	<b>⊕</b> 1	.8%	f
	Apache Maven	939,402	1,980,331	\$0	<ul> <li>-2</li> </ul>	6%	
	Apache Spark	883,449	1,688,773	\$0	• C	.9%	
	Apache Airflow	605,469	1,623,579	\$0	• 2	4%	Anna I
	Apache Flink	335,857	1,151,182	€0	• -0	.4%	· ·····
	Apache Kafka	369,545	630,851	\$0	• -C	.2%	





## Why Doris

## Why you need it

- Your dashboards refresh in real-time (not 15 minutes later)
- Complex queries finish in milliseconds, not minutes
- Handles 1000+ concurrent users without breaking
- Costs 60% less than cloud warehouses

## How it's different

- vs Snowflake: Real-time ingestion instead of batch loading
- vs BigQuery: Sub-second queries on live data
- vs ClickHouse: Better for mixed analytical workloads
- · vs Traditional databases: Built for analytics, not transactions

- E-commerce dashboards showing live sales
- Fraud detection needing instant alerts
- Customer analytics requiring fresh data
- Any use case where "wait 15 minutes" isn't acceptable



# vvny Doris is an Al-Ready Database

0612

Meetur



obla poris AMA

## Top 3 Use Cases of Apache Doris NPP.

## **Real-Time Analytics**

Ne

Data Warehousing

0610

## Observability



## **Real-Time Analytics**

## From User-Facing Analytics to Agent-Facing Analytics



## Agent-Facing Analytics

With the rise of AI technologies, especially AI Agents, more analytical decisions will be made automatically by AI. This will improve efficiency and accuracy in decision-making.

Fraud Detection | Ad Serving | Personalized Recommendation

-Real-Time Ingestion & Update: ~ 1 S minimum data latency

-Blazing-Fast Analytics: < 100ms average query latency

-High-Concurrent Queries: > 10,000QPS maximum query concurrency

- Agent Native: Doris MCP Server



## Data Warehousing



je Ai

## Better together: Open Data Lakehouse and Real-Time Analytics Database

## **Real-Time Analytics Engine**

Use **Doris** as the real-time analytics engine, primarily responsible for supporting interactive analytics and lightweight ETL computational workloads.

## **Batch Processing Engine**

Use Spark-like batch processing engines, primarily responsible for supporting longrunning ETL and machine learning computational workloads.

## Open Lakehouse Storage

Build an open lakehouse storage based on Data Lake using open table formats and open Catalog.

## **Open Data Lakehouse**

Lakehouse is designed for AI era

## Why Choose Doris

Build the Fastest Real-Time Data Warehouse in the Lakehouse, replacing Trino/Presto, SparkSQL, ...

- Fast and cost effective: 3x to Trino/Presto
- Open
- Unified



## Observability

## From Cloud Native + Microservice to LLM + Al Agent



2013: The concept of observability began to gain traction as companies like Twitter started to adopt it to manage

2020: Observability became a hot topic in DevOps, with more companies recognizing its importance in managing

C	Search or jun	np to	🖾 X+k	+~ ③ 🤎	🖵 Sign
				Edit Export	~ Share
c9e709ff409	95af96		② Last 24 hor	urs y Q Q I	Refresh ~
			🗐 Give fe	edback D Trace ID	🛱 Export
				12 chans ()	Novt
					7.92ms
		_	-		
1.98ms		3.96ms		5.94ms	7.92ms
1.24ms	_				
1.04ms					
4.32	ms 📿	_			
4.2	1ms				>
4.04ms		_			
3.8ms					
3.8ms	3.19ms				

					Log	JS						
う → Home → Dashboa	rds → Doris Den	no > 2 Logs	5				Q Searc	ch or jump to		🖾 🗱 + 🗸	•	n 🖵 Sig
											Edit Exp	port - Share
service_name fraud-de	tection ~ log	g level Ente	r value	match ~	Enter value	trace_id	Enter value	interval 1n	n ~ (	<ol> <li>Last 1 hour</li> </ol>	× Q :	Cancel
Histogram												
5 4 3 2 1 19:05	₽ 19:10	19:15	19:20	<b>- - - - -</b>	:25 1	9:30	19:35	19:40	19:45	19:50	19:55	5 20
Time	serviceName	level	traceID	body								
2025-04-18 20:00:02	ad	INFO	a166f2ff2	Targeted	ad request rece	eived for [bin	oculars]					
2025-04-18 19:59:46	ad	INFO	581d9e07	Targeted	ad request rece	eived for [ass	embly]					
2025-04-18 19:59:18	ad	INFO	e7d352a3	Non-targe	eted ad request	received, pr	eparing random	response.				
2025-04-18 19:59:10	ad	INFO	b5add4e1	Targeted	ad request rece	eived for [tele	scopes]					
2025-04-18 19:59:06	ad	INFO	57d4cef1	Targeted	ad request rece	eived for [bin	oculars]					
2025-04-18 19:58:54	ad	INFO	68f546a2	Non-targe	eted ad request	received, pr	eparing random	response.				
2025-04-18 19:58:50	ad	INFO	1734616a	Targeted	ad request rece	eived for [tele	scopes]					

-10x Cost Effective Compared to Elasticsearch

- Flexible Semi-Structured Data Variant Type

- Open Integration for ELK, OpenTelemetry, Grafana and more



## Al features being developed in Doris obten Andertun obten Andertun obten Andertun obten Andertun Andert

Dois

Meetur



oble Anne

KINA

0612

## Hybrid Search for RAG



## Why Vector-Capable General-Purpose Databases are Better for Enterprise GenAI ?

- Lack of Hybrid Query Capabilities
- Limited Integration with Structured Data
- Operational Complexity and Increased Costs
- Large scale data volume, eg PBs

:	SELECT
	id,
÷	name,
:	description
	FROM docs
	WHERE
:	update_time BETWEEN '2025-01-01' AND '2025-05-30'
	AND description MATCH 'GenAI'
0	AND version >= 2.0
	ORDER BY
	<pre>cosine_distance(embedding, [])</pre>
:	LIMIT 100
	·





## Al Operators in Doris

## Integrating GenAl for Enhanced Text Analysis in Database

```
WITH reviews(review) AS (
   VALUES
   ('The product has a great battery life.')
   ('Noise cancellation does not work as advertised. Avoid this
product.'),
    ('The product has a good design, but it is a bit heavy. Not recommended
for travel.'),
    ('Music quality is good but call quality could have been better.')
SELECT
    review,
    ai.extract(
        'Concisely extract features from the review, leave n/a or 0 when
unsure: ' || review,
       ARRAY [
           'battery: string - battery life of the product',
           'design: string - design features of the product',
            'sound: string - sound quality (e.g., music, call, noise
cancellation) of the product',
           'sentiment: number - sentiment score of the review; 1 (lowest)
to 5 (highest)
     AS data
    reviews;
                          NEAN
```

review	data		
The product has a great battery life	{"sound": "n/a", "design": "n/a", "battery": "great		
The product has a great battery me.	battery life", "sentiment": 5}		
Noise cancellation does not work as	{"sound": "Noise cancellation does not work as		
advertised Avoid this product	advertised.", "design": "n/a", "battery": "n/a",		
auventiseu. Avoiu tilis product.	"sentiment": 1}		
The product has a good design, but it is	{"sound": "n/a", "design": "good design, but a bit		
a bit heavy. Not recommended for travel.	heavy", "battery": "n/a", "sentiment": 3}		
Music quality is good but call quality	{"sound": "Music quality is good, call quality could		
could have been better	have been better.", "design": "n/a", "battery": "n/a",		
could have been bettel.	"sentiment": 3}		

ob A Dori



## Apache Doris 2025 RoadMap obt A Doris

Meetur

obta ports Al Meeturo



## Doris 2024 Overview Community Achievement

## One of the worlds' most active open source communities in big data





NEAN





## **Doris 2024 Overview** Things We've Done

ND.

Query Optimization	Storage Optimization	Semi-Structured Data Analysis	Lakehouse	<b>Cloud-Native Support</b>
<ul> <li>Smart Optimizer</li> <li>Statistics</li> <li>Adaptive Scan Parallelism</li> <li>Local Shuffle</li> <li>Async Materialized View</li> <li>Optimization on ARM</li> <li></li> </ul> https://doris.apache.org/ https://doris.apache.org/ https://doris.apache.org/	<ul> <li>Auto Incremental Col</li> <li>High Concurrency Im</li> <li>Multi-Statement Tran</li> <li>CCR - cross cluster re</li> </ul>	<ul> <li>VARIANT Data Type</li> <li>IP Data Type</li> <li>More Analytic Functions</li> <li></li> </ul>	<ul> <li>Optimization on Iceber</li> <li>More Connectors</li> <li>SQL Dialect Support</li> <li>Arrow Flight SQL Protoc</li> <li></li> </ul>	<ul> <li>Stateless Backends</li> <li>Multi Compute Group</li> <li>Smart Data Cache</li> <li></li> </ul>

- 0.



## Doris 2025 Roadmap Overview Innovation

## GenAl & ML

## Data Infrastructure in the GenAl Era DB for Al & Al for DB



- High-Throughput Data API Based on Arrow Flight (Done)
- MCP server (Done)
- Vector search (To release)
- Al operators

## **Incremental Processing**

Making data refresher



 Binlog Publishing and Subscription

• Realtime Materialized View



## **Doris 2025 Roadmap Overview Typical Use Cases**

## **Real-Time Analysis**

Becoming the fastest and most costeffective analytical database

## **Data Lakehouse**

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



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

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

## Log and Observability

## From Log to Observability





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





## **Doris 2025 Roadmap Overview Stability**

## **Release Management**

How to release stable and latest version

## **Code Review Rules**



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

- Pull request description
- Unit test coverage
- Code owner

Make code review easier, rigorous, and enforceable

6



## **More Test**

## More test scenarios



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



## Doris 2025 Roadmap Overview Community

## **Community Collaboration**

Making community collaboration more open and efficient



Doris Improvement Proposal

Ne

• Special Interest Group

00

• More deep dive articles & Webinars

## **Community Support**

Making community support smarter and more sustainable



- High-quality documentation
- Forum Construction

0612

• Doris Expert Al Model



## **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.

NEANE

## Welcome to Doris Community

0610-





# obt Anneeture

A Meetup

