

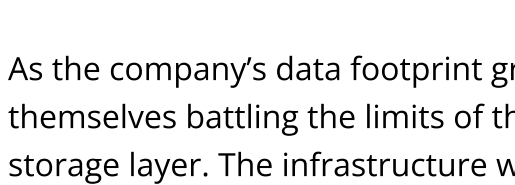
How DiDi Scaled To Hundreds Of Petabytes With Apache Ozone

Published on January 20, 2026

By The ASF

[◀ Back to Blog](#)

Building a cost-effective, high-performance data foundation for global mobility



When you're operating one of the world's largest ride-hailing and mobility platforms, every millisecond and megabyte counts. For DiDi Global, which generates over one petabyte of new data every day, scaling storage isn't just a technical challenge—it's a business imperative.

As the company's data footprint grew to more than 500PB annually, DiDi's engineers found themselves battling the limits of their legacy Apache Hadoop® Distributed File System (HDFS) storage layer. The infrastructure was struggling to keep pace with the company's explosive data growth, slowing downstream analytics and machine learning (ML) workloads that power everything from route optimization to dynamic pricing.

The Challenge: Scaling Without Compromise

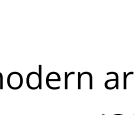
DiDi's HDFS-based infrastructure had served the company well, but it was beginning to show its age under the weight of petabyte-scale workloads. The team faced several interconnected problems:

- **Metadata bottlenecks:** File count limits in HDFS created stress on metadata services, driving up latency and throttling performance.
- **Read-heavy workloads:** RPC congestion and HDD I/O bottlenecks introduced lag for analytics and AI pipelines.
- **Escalating costs:** Triple replication inflated storage use and operational expenses.
- **Operational risk:** Even routine maintenance, such as decommissioning, carried stability concerns.

These issues had tangible business impacts. Slow metadata operations increased latency for end users, inflated costs, and created risks during peak demand periods.

"Metadata latency wasn't just a technical problem—it slowed down business units that rely on real-time analytics and AI insights," said JiangHua Zhu, Software Engineer, DiDi's Storage Team.

The Solution: Apache Ozone



After a rigorous evaluation, DiDi selected Apache Ozone™, a next-generation distributed storage system designed for scalability and performance in large, unstructured data environments.

Ozone's modern architecture—featuring RocksDB-based metadata management, separation of Object Manager (OM) and Storage Container Manager (SCM) services, and containerized data storage—provide the foundation DiDi needed to scale with confidence.

Key Benefits

- **Massive scalability:** Ozone comfortably supports tens of billions of files, removing HDFS metadata constraints.
- **Performance optimizations:** Features like OM Follower Read, multi-cluster routing, and NVMe caching help minimize latency and balance system load.
- **Cost efficiency through Erasure Coding:** Transitioning from 3x replication to EC 6-3 reduce storage overhead from 3.0x to roughly 1.5x—saving hundreds of petabytes.
- **Enhanced resilience:** Container-based data granularity improves fault tolerance and streamlined operations.

"Ozone gave us the flexibility to scale elastically across hundreds of petabytes without sacrificing performance," said Wei Ming, DiDi engineer.

The Results: Faster, Leaner, and More Reliable

The move to Apache Ozone delivered measurable, cross-functional benefits across DiDi's data ecosystem:

- **Latency:** P90 GetMetaLatency improved from 90ms to 17ms.
- **Throughput:** Production read throughput increased by more than 20% with OM follower reads.
- **Cost savings:** Erasure Coding cut the storage footprint nearly in half, saving both capital and operational expenses.
- **Stability under load:** The platform now operates smoothly even during cluster maintenance and peak traffic periods.
- **Developer productivity:** Application teams no longer need to manage small-file compaction, reducing complexity and accelerating data delivery.

Smooth Adoption Through Planning and Community Collaboration

DiDi's migration to Ozone was meticulous and deliberate. Engineers ensured data consistency with DistCp COMPOSITE_CRC checksums, implemented dual-write for rollback safety, and validated end-to-end compatibility with Hadoop, Apache Spark™, and S3 APIs.

The company also leaned heavily on the Apache Ozone open source community—which contribute bug fixes, performance enhancements, and feedback that benefit all users.

"The open source community was instrumental in our success—we gained support, shared knowledge, and received bug fixes that help everyone," said Shilun Fan, DiDi's storage leadership.

DiDi engineers even became active contributors, helping resolve issues such as metadata inconsistencies and Erasure Coding container handling. The collaboration ultimately strengthened both DiDi's deployment and the broader Ozone ecosystem.

Technical Highlights

- **Storage savings:** Hundreds of petabytes saved through Erasure Coding (6-3).
- **Read efficiency:** 20%+ improvement from OM follower reads and NVMe caching.
- **Unified access:** Hadoop API and S3 compatibility for batch, interactive, and ML workloads.
- **Scalability:** A single Ozone cluster can handle ~5 billion files, with the potential to scale to tens of billions.

Looking Ahead

DiDi's storage team continues to push the boundaries of performance and efficiency. Upcoming initiatives include:

- Integrating IO_URING and SPDK to enhance I/O performance.
- Developing AI-driven operational insights for anomaly detection and auto-remediation.
- Piloting tiered storage strategies for hot, warm, and cold data layers to optimize cost and performance.

"Ozone is more than a storage layer—it's the backbone of DiDi's data ecosystem and future AI innovation," said Hongbing Wang, DiDi technical lead.

The Takeaway

By embracing Apache Ozone, DiDi transformed its data storage infrastructure from a limitation into a competitive advantage. The move delivered lower costs, higher reliability, and faster access to the insights that power intelligent mobility.

At petabyte scale, even incremental improvements deliver outsized impact—and with Apache Ozone, DiDi has built a storage foundation ready for the next decade of data-driven innovation.

To learn more about Apache Ozone:

- Apache Ozone GitHub: <https://github.com/apache/ozone>
- Apache Ozone Getting Started: <https://ozone.apache.org/docs/edge/start/startfromdockerhub.html>
- Apache Ozone LinkedIn page: <https://www.linkedin.com/company/apache-ozone/>
- Apache Ozone X.com handle: <https://x.com/ApacheOzone>
- Apache Ozone Best Practices at Didi: <https://ozone.apache.org/assets/ApacheOzoneBestPracticesAtDidi.pdf>

Related Articles



The Apache Software Foundation Announces Apache Geode® 2.0

The ASF January 12, 2026

Wilmington, DE – January 12, 2026 – The Apache Software Foundation (ASF), the global home of open source software the world relies on, today...

[Read Post](#)



Celebrating 10,000 Committers: The People Who Power Apache Software Projects

The ASF December 5, 2025

By: Brian Proffitt, Vice President, Marketing & Publicity The Apache Software Foundation has come a long way since its founding in 1999, shepherding over...

[Read Post](#)



ASF Project Spotlight: Apache Geode

The ASF November 18, 2025

Can you tell us a bit about the project?Apache Geode is an in-memory data grid that provides real-time, consistent access to data-intensive applications at...

[Read Post](#)

Subscribe to ASF Plus One, Our Monthly Newsletter

SUBSCRIBE

DONATE



Contributors

- Contributor Getting Started
- Becoming a Committer
- Code of Conduct
- Community Resources
- Community Over Code
- Events

Project

- Projects
- Incubator Projects
- Projects Directory
- Mailing Lists
- Report a Vulnerability

Distributi

- Distributions
- Releases
- Infrastructure Status
- Infrastructure Statistics

Blog

- How the ASF Works
- The Apache Way
- Legal & Trademark
- Licenses
- Glossary
- FAQ

Developpe

- Wiki
- Issues
- Slack
- Self Serve Portal
- Infrastructure
- Whimsy
- Brand Guidelines
- Project Logos

About

- Our Sponsors
- Corporate Sponsorship
- Individual Supporters
- Leadership
- Members
- Diversity & Inclusion
- Newsroom
- Contact
- Privacy Policy