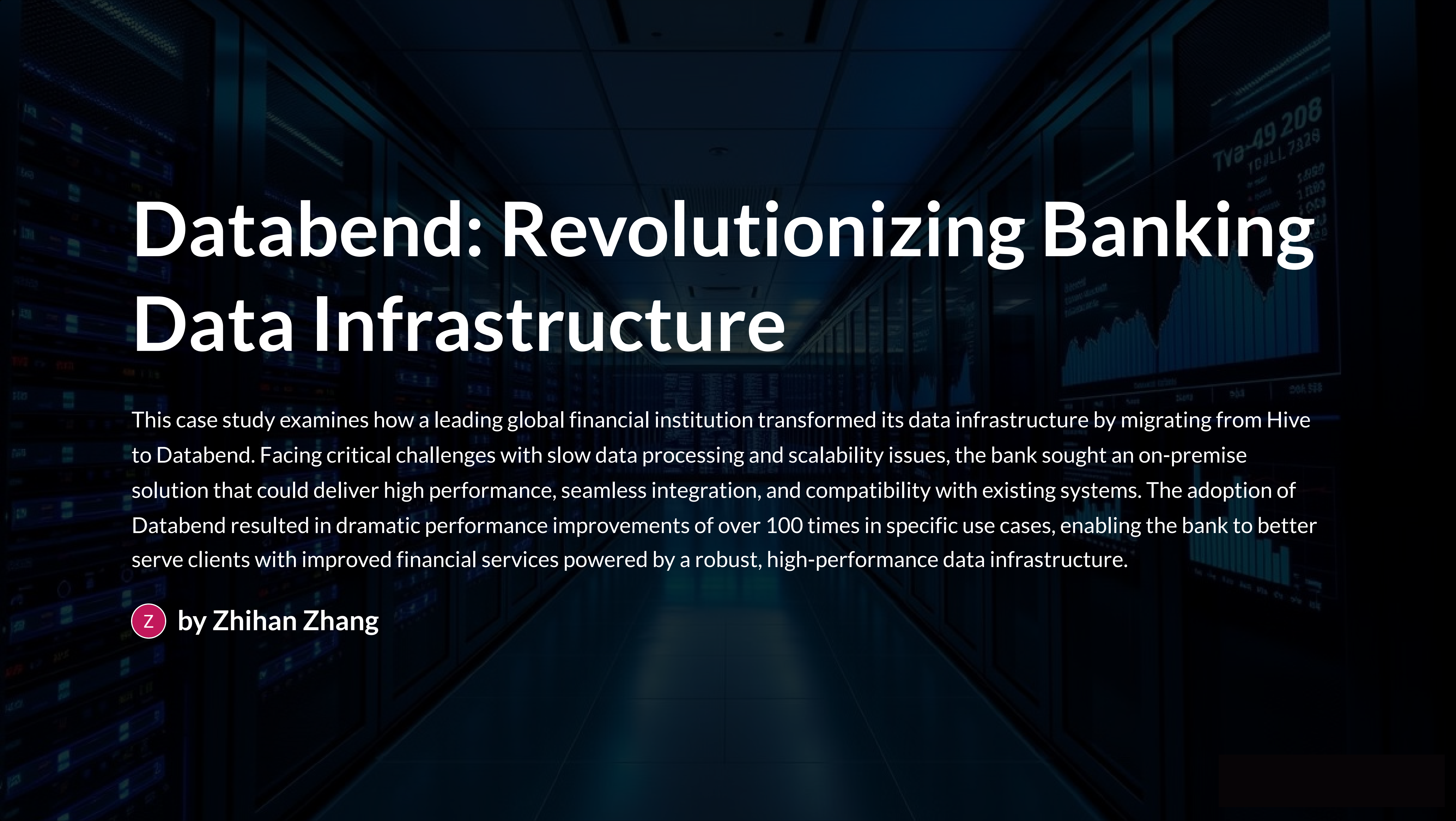


Databend: Revolutionizing Banking Data Infrastructure

This case study examines how a leading global financial institution transformed its data infrastructure by migrating from Hive to Databend. Facing critical challenges with slow data processing and scalability issues, the bank sought an on-premise solution that could deliver high performance, seamless integration, and compatibility with existing systems. The adoption of Databend resulted in dramatic performance improvements of over 100 times in specific use cases, enabling the bank to better serve clients with improved financial services powered by a robust, high-performance data infrastructure.

by Zhihan Zhang

Key Challenges in Banking Data Infrastructure

The global financial services company faced significant hurdles with its Hive-based system, impacting its ability to deliver efficient financial services. Poor performance and scalability were major concerns, with simple count queries taking hours to complete, especially in data auditing scenarios. The system struggled to handle tables with 4-10 billion records and complex multi-join queries, failing to meet required execution time targets.

Limited concurrency and inefficient data transformation across over 130,000 tables further hindered critical processes like financial reporting and risk analysis. Integration issues with the bank's custom Ceph storage led to suboptimal data operations, while high operational costs and time-consuming maintenance added to the overall inefficiency.

Performance Issues

- Slow query execution - Limited scalability

- Inefficient handling of large datasets

Operational Challenges

- Poor concurrency support

- Inefficient data transformation

- High maintenance costs

Integration Problems

- Suboptimal integration with Ceph storage

- Difficulty in expanding services

- Limited analytical capabilities

Databend Architecture Overview

The bank implemented a sophisticated multi-warehouse Databend architecture to efficiently manage diverse data processing needs. This architecture comprises four main warehouse types: Data Ingestion, Data Governance, Service, and Data Audit. Each warehouse is optimized for specific tasks, from high-volume real-time data intake to complex transformations and fast client-facing operations.

Databend Query nodes in the warehouses are designed with powerful hardware configurations to handle intensive data processing tasks. These nodes typically feature:

- CPU: 32 to 96 cores, providing substantial processing power
- Memory: Following a CPU to memory ratio of 1:4 to 1:8.

For example:

- A node with 32 cores might have 128GB to 256GB of RAM

The storage infrastructure is built on a custom, optimized version of Ceph, featuring NVMe caching for enhanced performance. This proprietary solution achieves throughput at bandwidth limits, with put/get operations taking just 10-30ms. Currently storing about 300-400TB of Databend data, the system has a total usable space of around 700TB, providing a scalable foundation for the bank's data operations.

1

Data Ingestion Warehouse

Handles high-volume, real-time data intake from various sources

2

Data Governance Warehouse

Manages complex data cleaning and transformation tasks

3

Service Warehouse

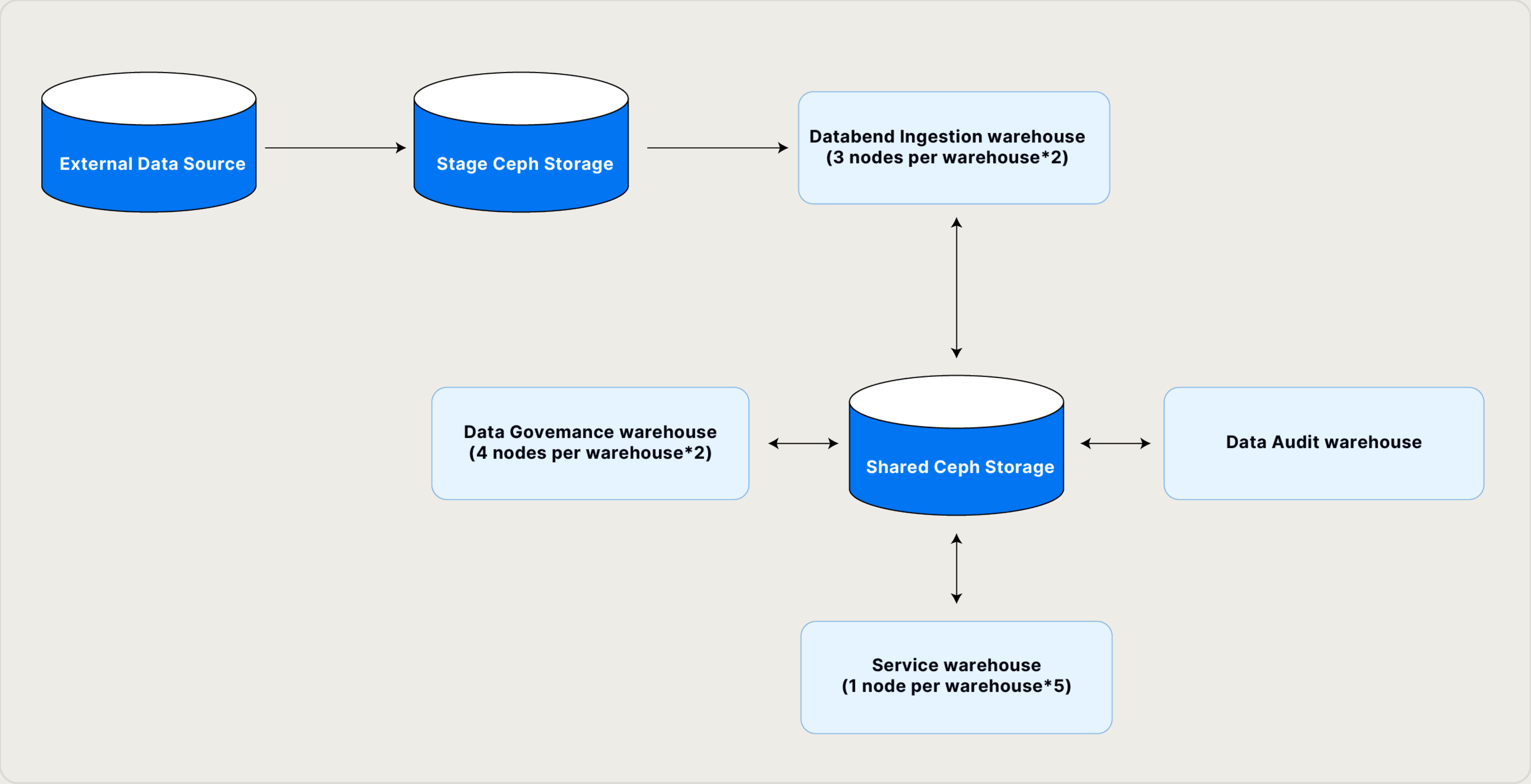
Serves client-facing operations with fast query execution

4

Data Audit Warehouse

Verifies data consistency across business systems

Databend Architecture



Data Ingestion Warehouse: Real-Time Financial Data Processing

The Data Ingestion Warehouse serves as the bank's high-speed conduit for diverse data streams, operating in multi-warehouse mode with 3 nodes per warehouse * 2. This warehouse efficiently processes millions of daily transactions from traditional RDBs while simultaneously capturing real-time social media sentiment data, enabling near-instant strategy adjustments in response to major economic events.

Demonstrating remarkable efficiency, the warehouse can ingest and process 500-record batches of social media reactions within 10 seconds. It also excels in large-scale data processing, handling 1 billion records in approximately 4 hours, including deduplication. This ensures that by the time traders arrive at work, they have access to clean, deduplicated data from overnight batch processes, ready for immediate analysis and decision-making.

1

Data Intake

Millions of transactions ingested daily from various sources

2

Real-Time Processing

500-record batches processed within 10 seconds

3

Large-Scale Handling

1 billion records processed in ~4 hours, including deduplication

4

Ready for Analysis

Clean, deduplicated data available for traders by morning

Data Governance Warehouse: Powering Complex Financial Data Transformations

At the heart of the bank's data infrastructure lies the unsung hero - the Data Governance Warehouse. This specialized warehouse, exclusively managed by skilled data engineers, is the backbone of the bank's data transformation and quality assurance processes.

With the ability to handle over **130,000 tables**, some containing **up to 1 billion rows each**, the Data Governance Warehouse showcases remarkable scale and efficiency. It's not uncommon for this warehouse to process complex **joins involving up to 38 tables** simultaneously - a task that now completes within a remarkable **one-hour timeframe**.

1

Massive Scale

Manages a staggering 130,000+ tables, some with up to 1 billion rows

2

Complex Operations

Effortlessly joins data from up to 38 tables in a single process

3

Dramatic Improvement

Slashes business data transformation from days to just one hour

4

High-Volume Processing

Handles tables with over 4 billion rows of historical data in around 10 hours

By efficiently handling **8-10 concurrent jobs** on complex data transformation tasks, the Data Governance Warehouse demonstrates its unparalleled strength in concurrency. This performance enables data engineers to maintain high data quality standards, ensure timely updates to critical datasets, and effectively support the bank's analytical and operational needs.

Serving as a crucial foundation for the bank's data infrastructure, the Data Governance Warehouse indirectly empowers various departments that rely on clean, well-managed data for their day-to-day operations and strategic decision-making. This unsung hero quietly powers the bank's data-driven transformation, ensuring data integrity and driving operational excellence.

Service Warehouse: Enabling Responsive Financial Services

The Service Warehouse stands at the forefront of the bank's customer-facing operations, leveraging Databend's capabilities to deliver responsive and efficient financial services. This warehouse's architecture is finely tuned for high concurrency and rapid response times, comprising 5 instances, each allocated 100GB memory and 500GB local disk cache. This robust configuration enables the warehouse to handle an impressive **1500 concurrent operations**, with each instance managing 200-300 concurrent queries.

Performance is a standout feature, with the warehouse consistently delivering results in about **1 second**, well within the bank's 3-second target for query execution. This is achieved through careful optimization, including the use of warehouse keys for most queries and controlled complexity for operations requiring joins. The max_threads setting is optimized at 8 (or 4 for Intel CPUs), allowing each query to utilize up to 8 threads for task execution, fully harnessing the power of concurrency in each node.

Instances	Memory per Instance	Local Disk Cache per Instance	Total Concurrent Operations	Concurrent Queries per Instance	Average Response Time
5	100GB	500GB	1500	200-300	~1 second

Data Audit Warehouse: Ensuring Financial Data Integrity

The Data Audit Warehouse serves as the bank's financial watchdog, continuously ensuring data consistency across all systems. This warehouse performs the critical task of reconciling data across various services, conducting meticulous row-by-row and field-by-field comparisons. For example, it compares customer account balances across the main ledger, online banking system, and ATM network, ensuring perfect alignment down to the last cent.

The warehouse efficiency has revolutionized the bank's auditing processes, reducing tasks that once took days to hours or even minutes. A full audit of the daily transaction log, previously a weekend-long process, now completes overnight. This speed not only allows for more frequent and thorough audits but also enables near-real-time error detection. In one notable instance, the warehouse detected a discrepancy in foreign exchange rates between the trading system and the customer-facing application within minutes of its occurrence, preventing potential financial losses and maintaining the bank's reputation for accuracy.

Rapid Audits

Full daily transaction log audit reduced from days to overnight

Detailed Reconciliation

Row-by-row and field-by-field data comparisons across systems

Enhanced Security

Near-real-time error detection prevents financial losses

Performance Boost

Over 100 times faster than previous Hive-based system

Conclusions: Databend's Transformative Impact on Banking Operations

1 **Revolutionizing Data Processing**

Databend's multi-warehouse architecture has revolutionized data processing and management for this leading financial institution, addressing critical challenges in speed, scalability, and data integrity.

2 **Enabling Data-Driven Decisions**

The implementation has yielded dramatic performance improvements across all warehouse types, with the Data Audit Warhouse demonstrating a 100-fold speed increase over the previous Hive-based system.

3 **Improving Customer Experience**

This boost in performance has transformed time-consuming processes into near-real-time operations, enabling more frequent audits, faster decision-making, and enhanced regulatory compliance.

4 **Freeing Up Resources for Innovation**

The scalability and flexibility of Databend's solution have allowed the bank to efficiently manage diverse data processing needs, from high-volume ingestion to complex transformations and real-time customer service queries.