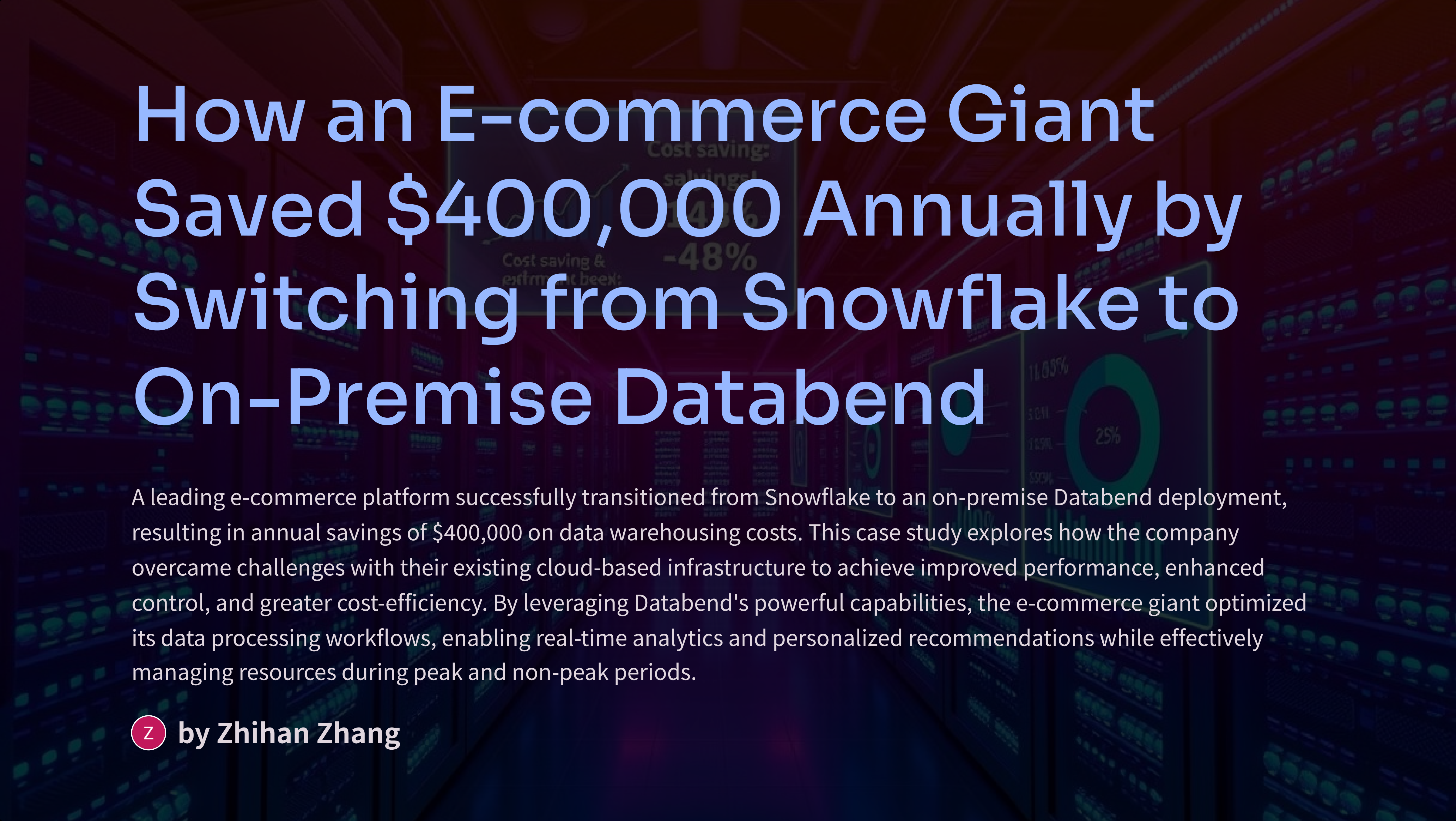


How an E-commerce Giant Saved $400,000 Annually by Switching from Snowflake to On-Premise Databend

A leading e-commerce platform successfully transitioned from Snowflake to an on-premise Databend deployment, resulting in annual savings of $400,000 on data warehousing costs. This case study explores how the company overcame challenges with their existing cloud-based infrastructure to achieve improved performance, enhanced control, and greater cost-efficiency. By leveraging Databend's powerful capabilities, the e-commerce giant optimized its data processing workflows, enabling real-time analytics and personalized recommendations while effectively managing resources during peak and non-peak periods.

by Zhihan Zhang

Goals and Benefits of Migration

Key Objectives

- Enhance real-time data ingestion and cleansing for personalized recommendations
- Improve complex offline analysis capabilities for user retention and market trend forecasting
- Ensure system flexibility to handle increased data processing demands during peak periods

Realized Benefits

1

Substantial Cost Savings

The e-commerce platform achieved approximately $400,000 in annual cost reductions on data warehousing expenses by migrating to an on-premise Databend deployment.

2

Enhanced Performance

Databend's efficient large-scale data processing capabilities enabled faster complex offline analysis and real-time data ingestion, meeting the platform's performance requirements.

3

Resource Optimization

The ability to dynamically adjust resources according to business needs during low and peak periods significantly reduced unnecessary compute resource consumption.

4

Improved Control and Security

On-premise deployment allowed full control over data infrastructure, ensuring compliance with data privacy regulations and implementing tailored security measures.

Challenges with Cloud-Based Data Warehousing

The e-commerce platform faced significant challenges with their Snowflake-based data infrastructure, primarily revolving around cost escalation and resource management inefficiencies. During peak periods, the Multi-Cluster Warehouse mode led to substantial cost increases of 60% to 80%, forcing the platform to limit the frequency of critical offline tasks to meet cost goals.

Resource utilization was another major concern. Small Warehouses scaled from 2 to 6 instances for data ingestion, while Medium Warehouses scaled similarly for transformation tasks. This scaling pattern often resulted in over-provisioning and increased costs, especially during peak periods when both Small and Medium Warehouses scaled to 4-6 instances. Plus, they have used a X-Large size warehouse, dedicated to offline analysis, during peak periods (every 30 minutes instead of hourly), dramatically increasing operational costs.

1

Escalating Costs

The platform experienced a 60-80% increase in monthly costs during peak periods, significantly impacting the overall operational budget.

2

Resource Inefficiencies

Snowflake's elastic scaling led to underutilization during non-peak times and excessive resource consumption during peak periods.

3

Scalability Concerns

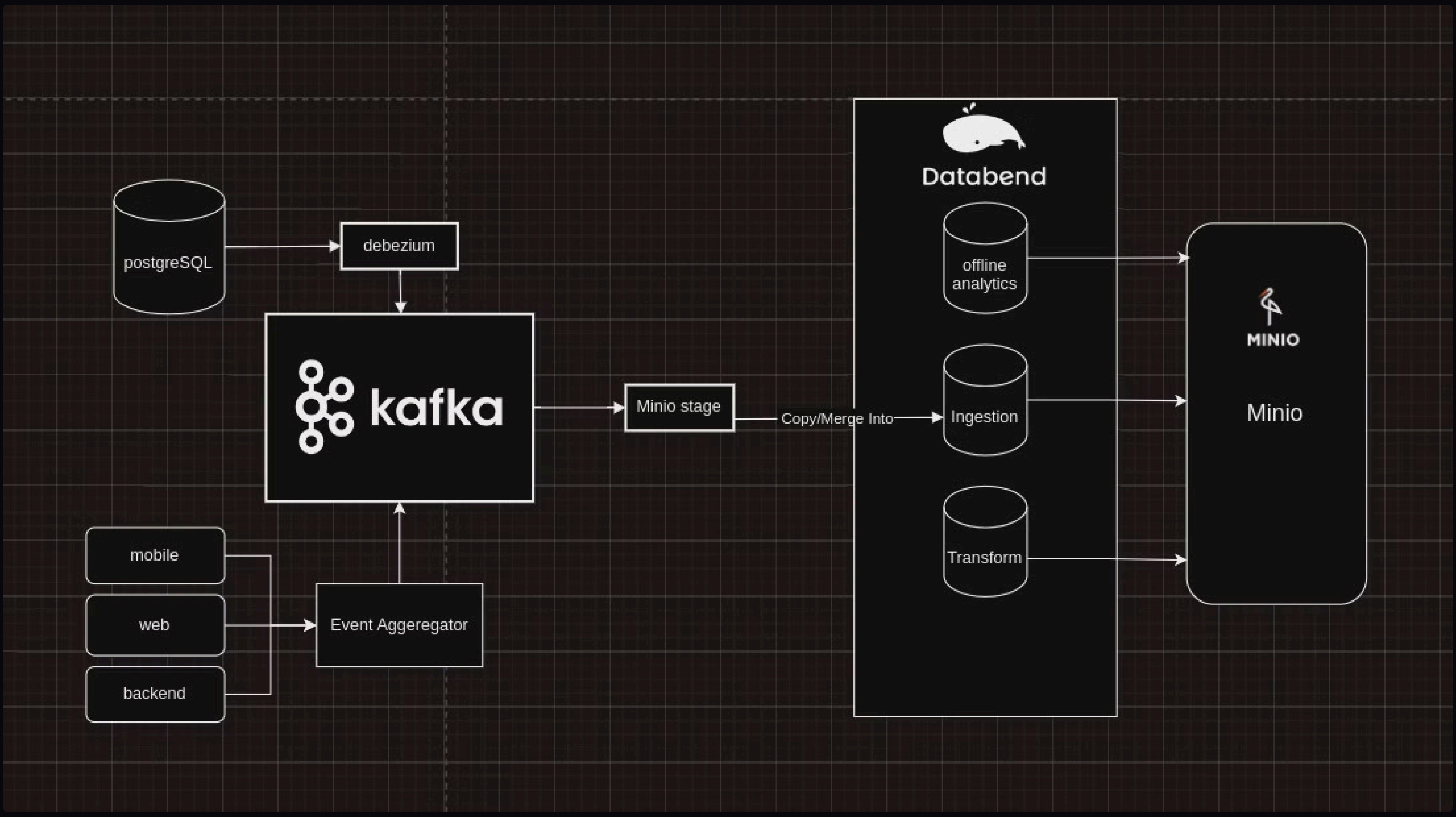
As data volume continued to grow, there were increasing worries about the long-term scalability and cost-effectiveness of the Snowflake solution.

4

On-Premise Limitations

The platform was unable to leverage existing on-premise infrastructure and idle machines due to Snowflake's cloud-based nature.

Architecture Overview



The e-commerce platform's data infrastructure is designed to handle massive amounts of user behavior data and provide real-time analytics capabilities. The system ingests data from two primary sources into a unified Apache Kafka pipeline: transactional data from a PostgreSQL database and event data from user interactions across web and mobile applications.

For transactional data, the platform utilizes Change Data Capture (CDC) technology with Debezium's PostgreSQL connector within the Kafka Connect framework. This captures real-time changes from PostgreSQL, which stores critical business information. Simultaneously, a custom-built event ingestion system processes user interaction events, performing schema validation before publishing to Kafka.

From Kafka, a Sink connector streams all ingested data into a MinIO object storage stage, acting as an optimized intermediate storage layer for Databend's ingestion process. Databend then efficiently ingests the data from MinIO, enabling high-performance analytics and processing.

Databend as the Data Warehouse Solution

The e-commerce platform leverages Databend warehouses for various operations, including ingestion, data transformation, and offline processing. The scale of data handling is impressive, with the system processing trillions of user behavior records at an ingestion rate of 500,000 records per second. The uncompressed data volume reaches approximately 2PB, which is efficiently compressed to 200TB for storage in MinIO object storage.

Databend's powerful processing capabilities enable the platform to handle this massive data volume while providing real-time analytics and insights. The combination of Databend's performance and MinIO's scalable object storage creates a cost-effective and highly efficient data warehousing solution that meets the platform's growing needs.

Data Volume

Trillions of user behavior records

2PB uncompressed, 200TB compressed

Ingestion Rate

500,000 records per second

Storage Solution

MinIO object storage for long-term retention and analysis. Up to 500TB.

Ingestion Warehouse Deep Dive

The data ingestion layer of the e-commerce platform's architecture is designed to efficiently consume stage data from MinIO, with built-in idempotency guarantees to prevent duplicate data ingestion. This critical component of the data pipeline ensures that all incoming data is accurately and reliably processed, forming the foundation for downstream analytics and business intelligence operations.

Ingestion Warehouse Configuration

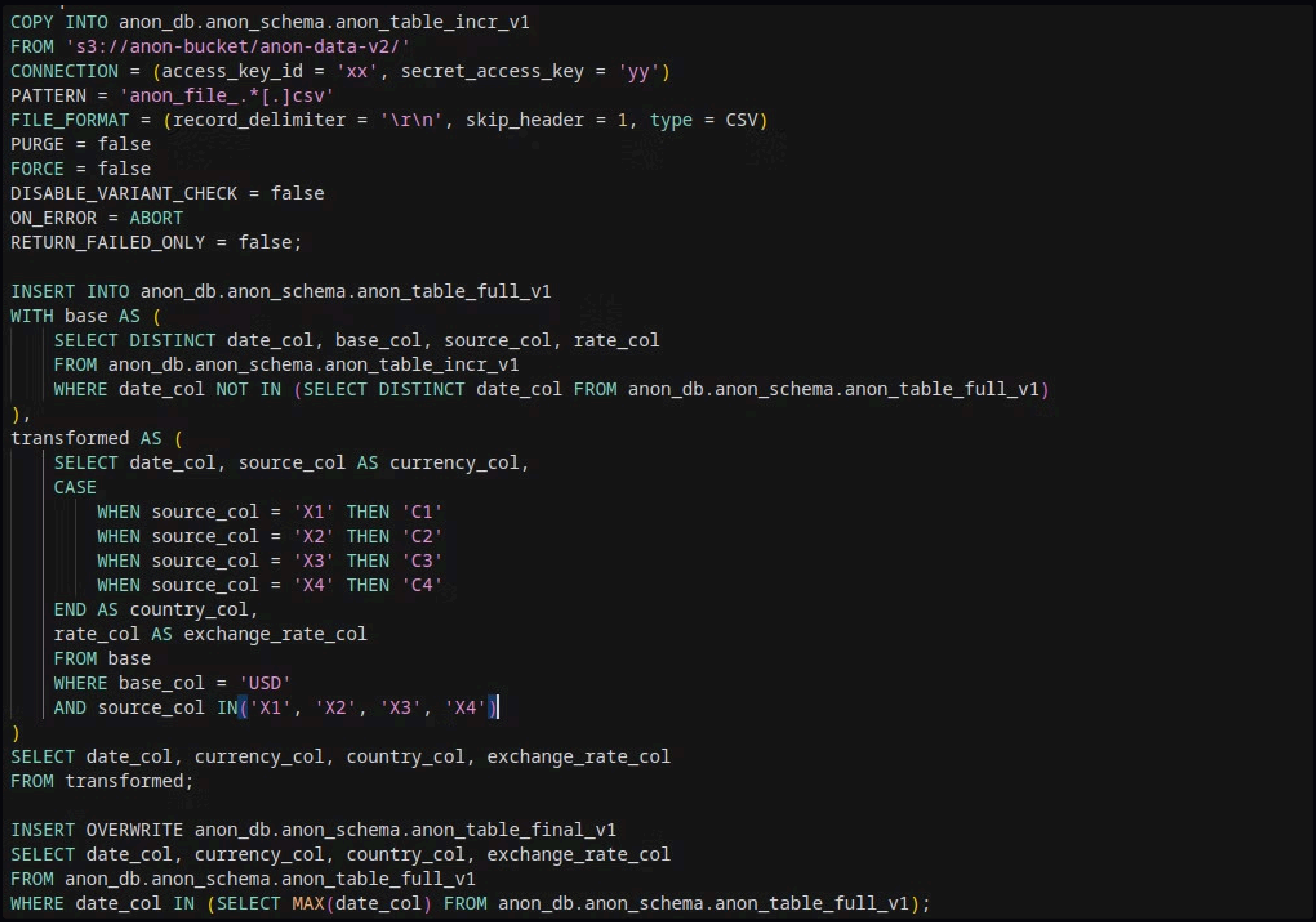
- 1 Databend Warehouse (64 cores, 256GB memory)
- Handles data ingestion rate of 500,000 records per second
- Single warehouse meets needs during both low and peak periods

Workload Metrics

Scenario	Traffic	Processing Time
Normal	100,000-300,000 records/second	0.5-1 second
Peak	800,000-1,000,000 records/second	1-1.5 seconds

The ingestion warehouse's ability to maintain consistent performance across varying traffic levels demonstrates Databend's scalability and efficiency in handling real-time data processing demands.

Sample Ingestion SQL(NDA with **Anonymization**)



```sql
COPY INTO anon_db.anon_schema.anon_table_incr_v1
FROM 's3://anon-bucket/anon-data-v2/'
CONNECTION = (access_key_id = 'xx', secret_access_key = 'yy')
PATTERN = 'anon_file_.*[.]csv'
FILE_FORMAT = (record_delimiter = '\r\n', skip_header = 1, type = CSV)
PURGE = false
FORCE = false
DISABLE_VARIANT_CHECK = false
ON_ERROR = ABORT
RETURN_FAILED_ONLY = false;

INSERT INTO anon_db.anon_schema.anon_table_full_v1
WITH base AS (
    SELECT DISTINCT date_col, base_col, source_col, rate_col
    FROM anon_db.anon_schema.anon_table_incr_v1
    WHERE date_col NOT IN (SELECT DISTINCT date_col FROM anon_db.anon_schema.anon_table_full_v1)
),
transformed AS (
    SELECT date_col, source_col AS currency_col,
    CASE
        WHEN source_col = 'X1' THEN 'C1'
        WHEN source_col = 'X2' THEN 'C2'
        WHEN source_col = 'X3' THEN 'C3'
        WHEN source_col = 'X4' THEN 'C4'
    END AS country_col,
    rate_col AS exchange_rate_col
    FROM base
    WHERE base_col = 'USD'
    AND source_col IN('X1', 'X2', 'X3', 'X4')
)
SELECT date_col, currency_col, country_col, exchange_rate_col
FROM transformed;

INSERT OVERWRITE anon_db.anon_schema.anon_table_final_v1
SELECT date_col, currency_col, country_col, exchange_rate_col
FROM anon_db.anon_schema.anon_table_full_v1
WHERE date_col IN (SELECT MAX(date_col) FROM anon_db.anon_schema.anon_table_full_v1);
```

Transformation Warehouse Deep Dive

The data transformation warehouses play a crucial role in enriching and cleaning up the ingested data. These warehouses perform tasks such as normalizing data formats, resolving inconsistencies, and applying business rules to ensure data quality. For instance, they standardize product categories across different suppliers and merge customer profiles from various touchpoints.

Transformation Warehouse Configuration

- 1-2 Databend Warehouses (64 cores, 256GB memory each)
- Scales up to 2 warehouses during peak periods
- Ensures efficient data cleansing even during high-load times

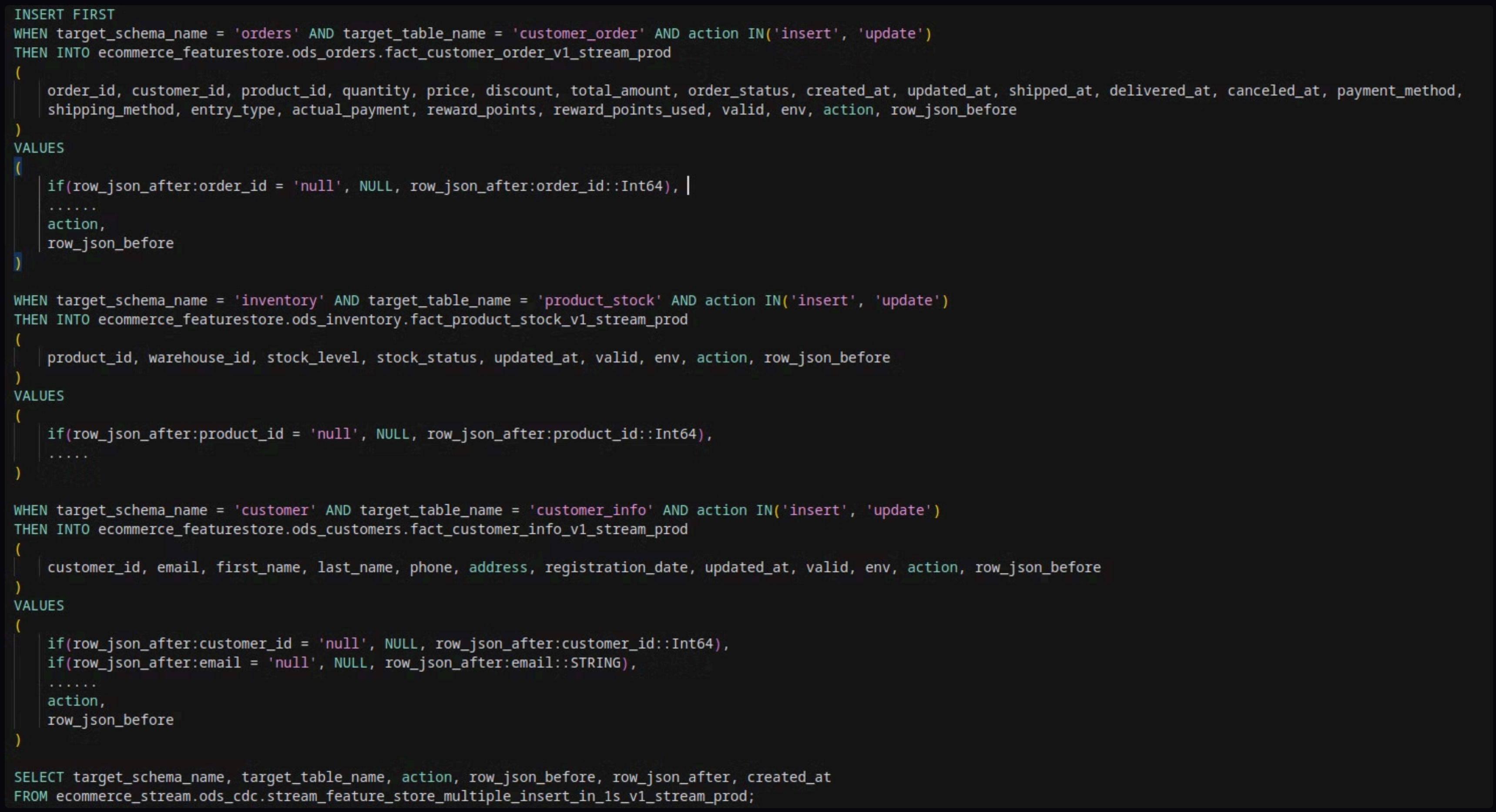
Normal Workload

Data volume: 500,000-1,000,000 records
Processing time: 10 seconds per transformation

Peak Workload

Data volume: 1,000,000-2,000,000 records
Processing time: 10 seconds per transformation

Sample Transformation SQL (NDA with **Anonymization**)



```sql
INSERT FIRST
WHEN target_schema_name = 'orders' AND target_table_name = 'customer_order' AND action IN('insert', 'update')
THEN INTO ecommerce_featurestore.ods_orders.fact_customer_order_v1_stream_prod
(
    order_id, customer_id, product_id, quantity, price, discount, total_amount, order_status, created_at, updated_at, shipped_at, delivered_at, canceled_at, payment_method,
    shipping_method, entry_type, actual_payment, reward_points, reward_points_used, valid, env, action, row_json_before
)
VALUES
(
    if(row_json_after:order_id = 'null', NULL, row_json_after:order_id::Int64),
    ......
    action,
    row_json_before
)

WHEN target_schema_name = 'inventory' AND target_table_name = 'product_stock' AND action IN('insert', 'update')
THEN INTO ecommerce_featurestore.ods_inventory.fact_product_stock_v1_stream_prod
(
    product_id, warehouse_id, stock_level, stock_status, updated_at, valid, env, action, row_json_before
)
VALUES
(
    if(row_json_after:product_id = 'null', NULL, row_json_after:product_id::Int64),
    .....
)

WHEN target_schema_name = 'customer' AND target_table_name = 'customer_info' AND action IN('insert', 'update')
THEN INTO ecommerce_featurestore.ods_customers.fact_customer_info_v1_stream_prod
(
    customer_id, email, first_name, last_name, phone, address, registration_date, updated_at, valid, env, action, row_json_before
)
VALUES
(
    if(row_json_after:customer_id = 'null', NULL, row_json_after:customer_id::Int64),
    if(row_json_after:email = 'null', NULL, row_json_after:email::STRING),
    ......
    action,
    row_json_before
)

SELECT target_schema_name, target_table_name, action, row_json_before, row_json_after, created_at
FROM ecommerce_stream.ods_cdc.stream_feature_store_multiple_insert_in_1s_v1_stream_prod;
```

Offline Analytics Warehouse Deep Dive

The offline analytics warehouses are dedicated to processing complex queries and generating insights that drive business decisions. These warehouses run sophisticated e-commerce analytics, including funnel analysis to track user journeys, cohort analysis for understanding customer retention and lifetime value, and inventory turnover analysis to optimize stock levels.

Offline Analysis Warehouse Configuration

Normal Workload

Scanned data volume: Tens of billions of records.
Processing time: 5 minutes per analysis
Analysis frequency: Once every 25 minutes

Peak Worload

Scanned data volume: Tens of billions of records

Processing time: 5 minutes per analysis
Analysis frequency: Once every 5 minutes

Ensures large-scale personalized recommendations and other critical analytical tasks are performed with increased frequency during peak periods, supporting real-time business decision-making.

- 1 Databend Warehouse (64 cores, 256GB memory)
- Executes analysis every 25 minutes during non-peak periods
- Increases frequency to every 5 minutes during peak period

Conclusion: Benefits of On-Premise Data Warehousing

The migration from Snowflake to on-premise Databend has proven to be a game-changing decision for the e-commerce platform. It has delivered substantial benefits in terms of performance, cost-efficiency, and operational flexibility. Databend's robust warehouses have significantly enhanced the platform's data processing capabilities, with a single warehouse handling an impressive ingestion rate of 500,000 records per second and enabling near real-time personalized recommendations during peak periods.

Cost efficiency has been a major win, with the platform achieving approximately $400,000 in annual savings on data warehousing expenses. Databend's flexible resource management allows for dynamic adjustment based on business needs, significantly reducing unnecessary compute resource consumption during non-peak periods.

The on-premise deployment has provided enhanced control over data infrastructure, allowing for customized security measures and ensuring compliance with data privacy regulations. It has also enabled the platform to leverage existing on-premise resources, optimizing overall IT resource utilization.

Enhanced Performance

Improved data processing capabilities for real-time analytics and personalized recommendations

Cost Savings

$400,000 annual reduction in data warehousing expenses

Operational Flexibility

Dynamic resource adjustment and utilization of existing infrastructure

Enhanced Control

Improved data security and compliance with privacy regulations