JDT 京东科技

Development of Cloudnative Architecture for ClickHouse in JD.com

JD Technology – JD Cloud

Xuewei Wang (王学伟)

resume

JDT 京东科技

- Graduated from Zhejiang University in mid-2019 and joined JD.com
- Core Developer of JD Cloud ClickHouse
- core developer of cloud native real-time data warehouse

JDT 京东科技

contents

JDT 京东科技

01

The first stage of JD Cloud-native ClickHouse

The Background of the Birth of JD Clickhouse

JDT 京东科技

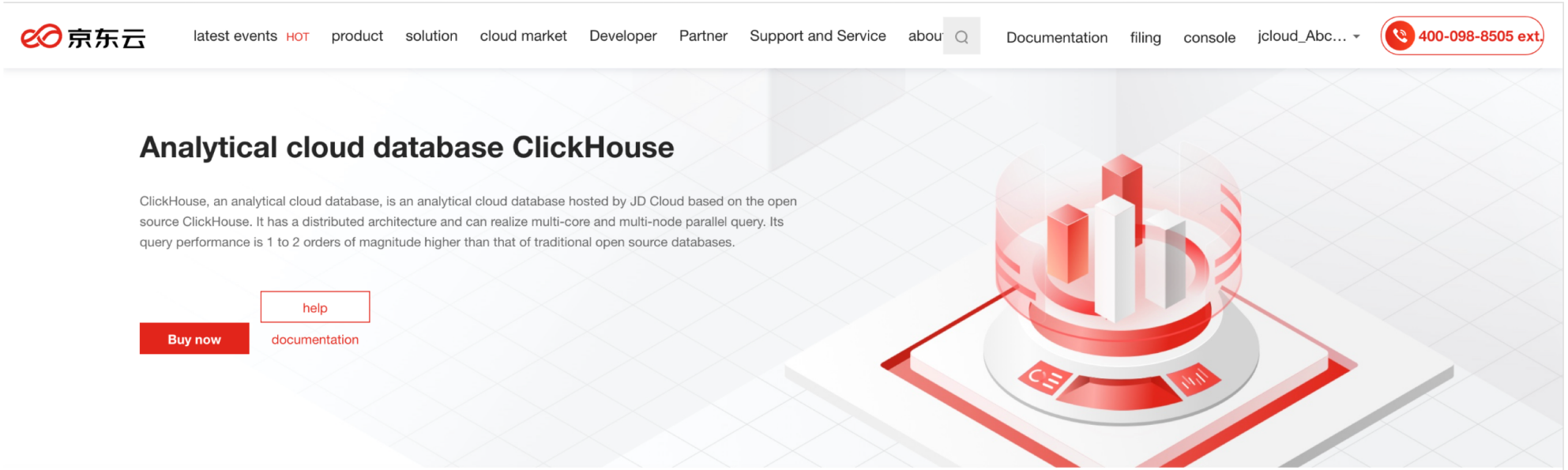
- Open source since 2016, the read and write performance far exceeds the open source competing products in the same period, and the community is very active.

- Large-scale domestic Internet companies have successively introduced OLAP engines on a large scale. In 2019, they have been widely used in JD.com as a general-purpose analytical database.

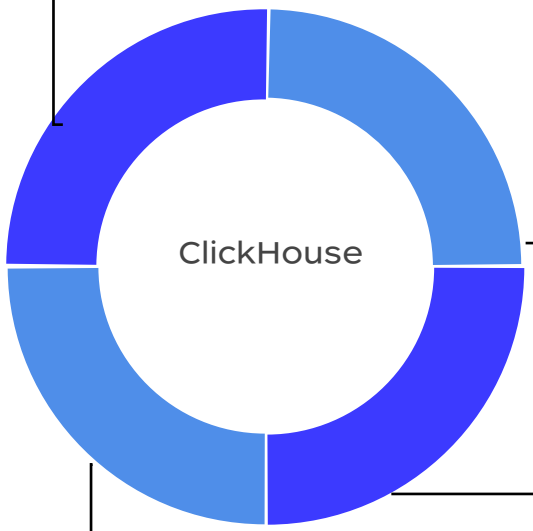
- Manual transmission of super sports car, tens of thousands of servers in JD.com, management and maintenance requires a lot of manpower and material resources.

JD Cloud ClickHouse Home Page

-It is convenient for users to apply for on-demand, one-click hosting



JD Cloud ClickHouse Features



Complete lifecycle management

Create cluster: the currently supported maximum scale is 6 copies and 128 shards; the maximum specification of a single node is 64 cores and 256GB memory; the maximum storage of a single node is 16TB; the cluster creation progress can be viewed during creation; in the case of 100 nodes, the creation time is dozens of minutes.

Delete cluster: Prevent users from accidentally deleting the instance. After the user clicks delete, it will be kept for one day.

Cluster details: You can view all information about this cluster.

cluster scaling: supports vertical scaling; disk supports vertical scaling; supports sharding and replica-level horizontal scaling.

Easy-to-use access to the cluster

Cluster-level LB domain name.

Node-level domain name: The domain name will not change after node restart or high availability.

Open mysql port: Users can connect to clickhouse server through mysql client.

The cluster-level domain name extranet access function can be enabled

Convenient account management

Support various authority account management:
create, query, modify password, delete, etc.

One-click version upgrade

Provide upward batch upgrade service for different LTS versions

One-click parameter modification

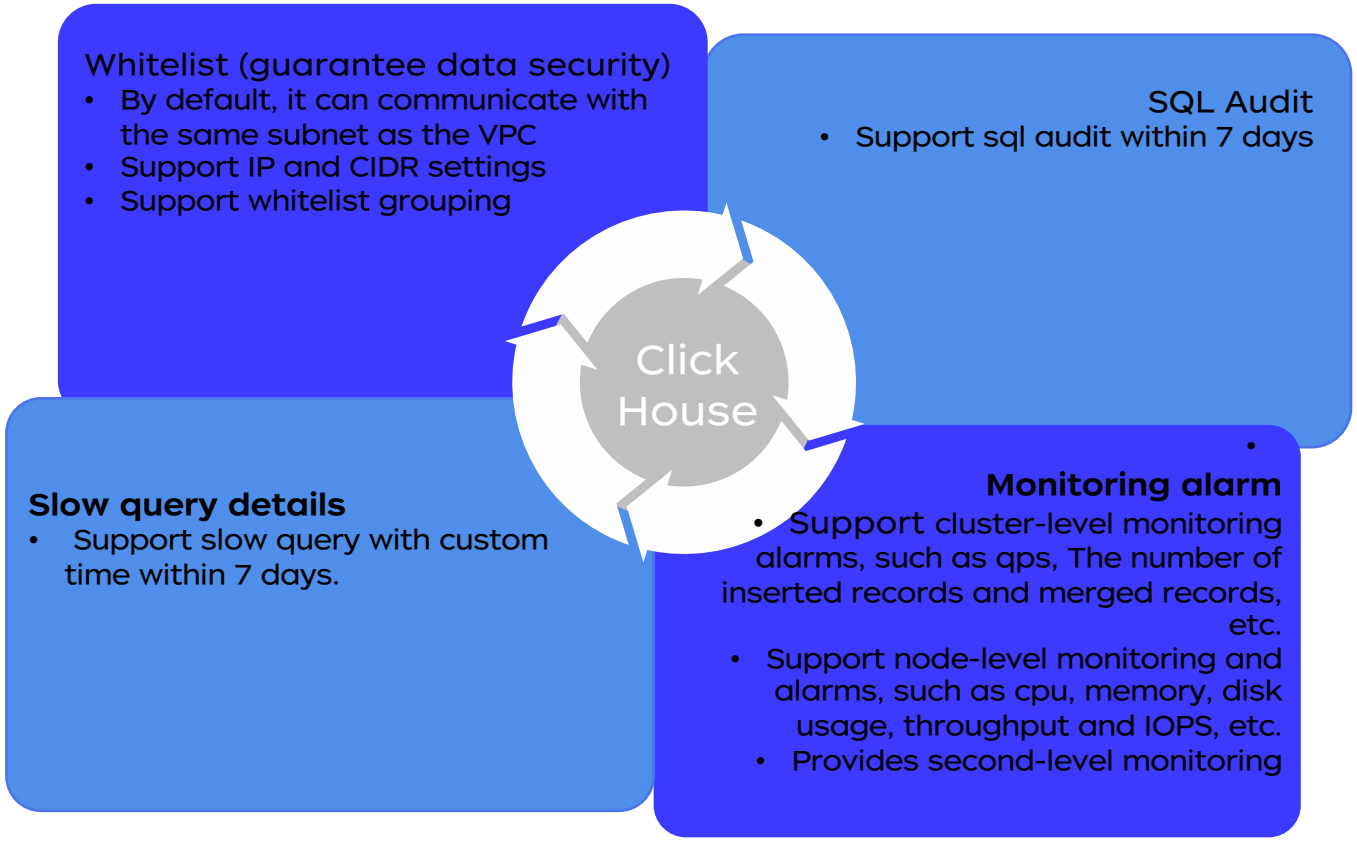
Users can modify parameters directly on the console

JD Cloud ClickHouse Features

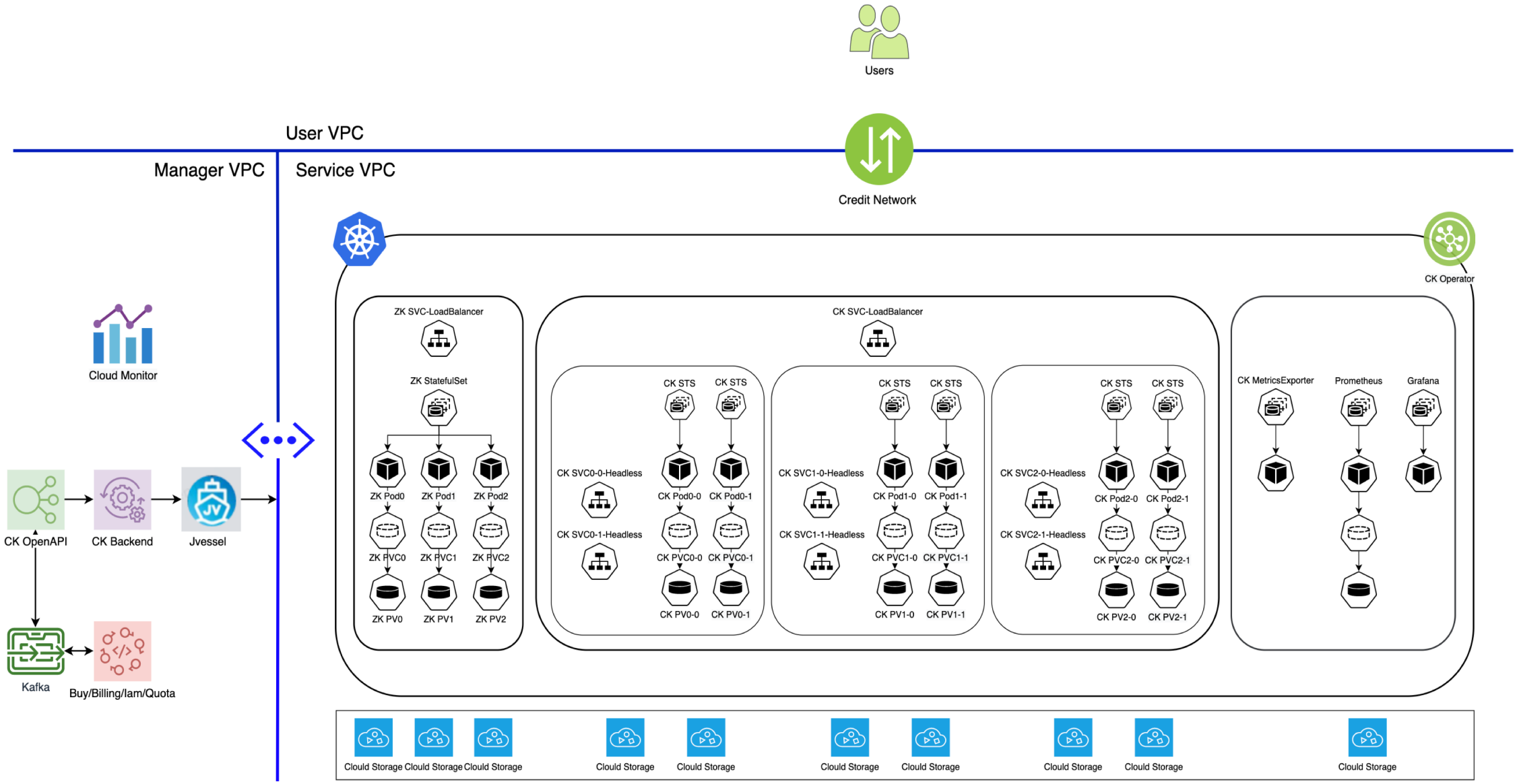
ClickHouse

JDT 京东科技

Security, Auditing, Logging, Monitoring and Alerting



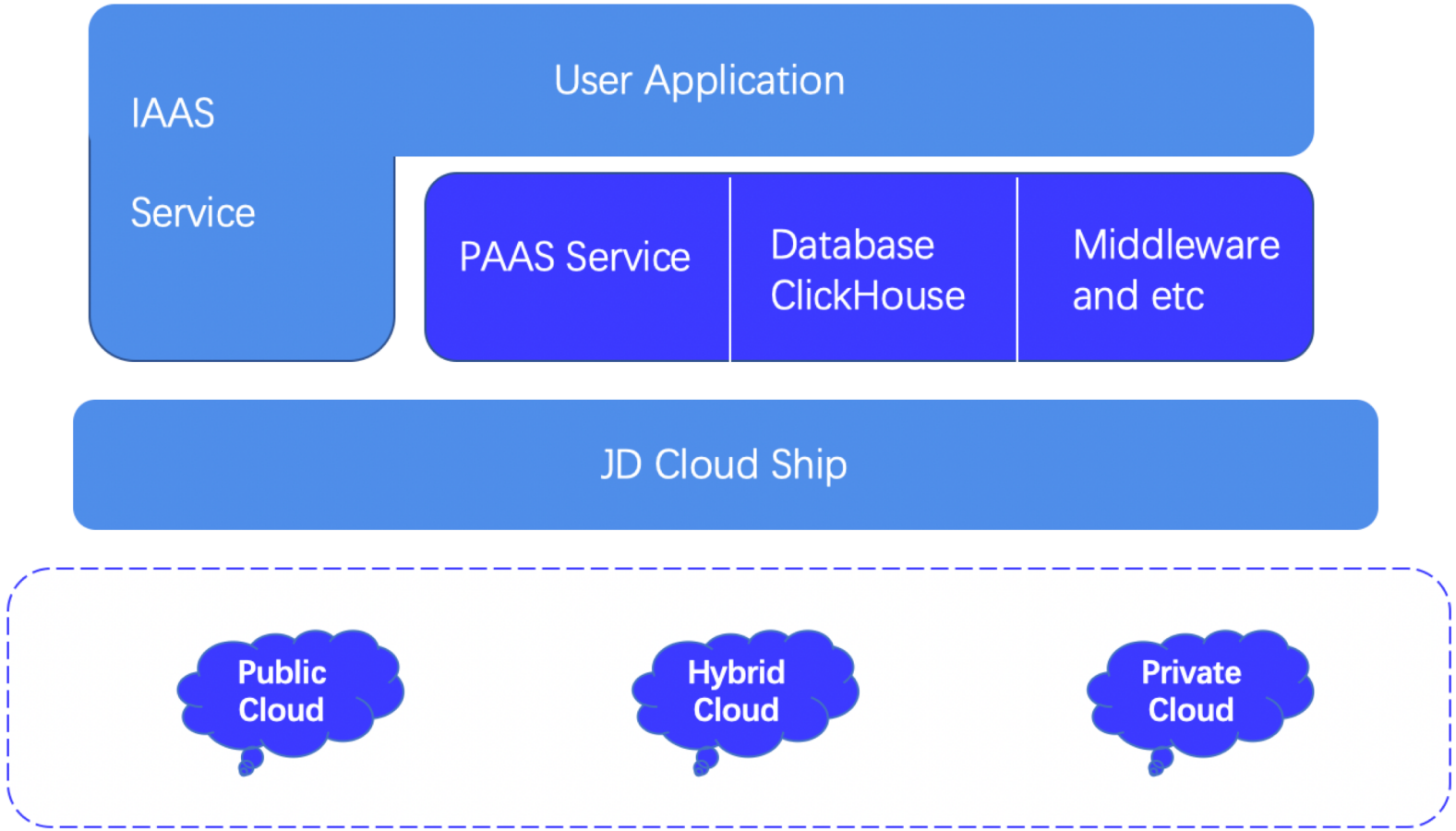
JD Cloud ClickHouse Architecture



JD Cloud Cloud Ship

ClickHouse

JDT 京东科技



JD Cloud ClickHouse Based on ClickhoHouse-operator

JDT 京东科技

Based on the benefits of k8s using clickhouse-operator deployment management, a key step to achieve cloud native. A key advantage of Kubernetes is the automation of much of the operational work that comes with applications and infrastructure. The Kubernetes Operator is considered an important means to achieve this advantage.

- Operators make large-scale Kubernetes deployments practical by automating difficult, error-prone work that would otherwise be performed manually
- Custom container orchestration across multiple hosts.
- Make more full use of hardware and improve resource utilization (more flexible than manual whole machine deployment).
- Effectively manage and automate application deployment and updates.
- Mounting and adding storage is more convenient.
- Scale applications and their resources quickly and on-demand.
- Ensures that deployed applications always run as deployed.
- Implement health checks and self-healing for applications.

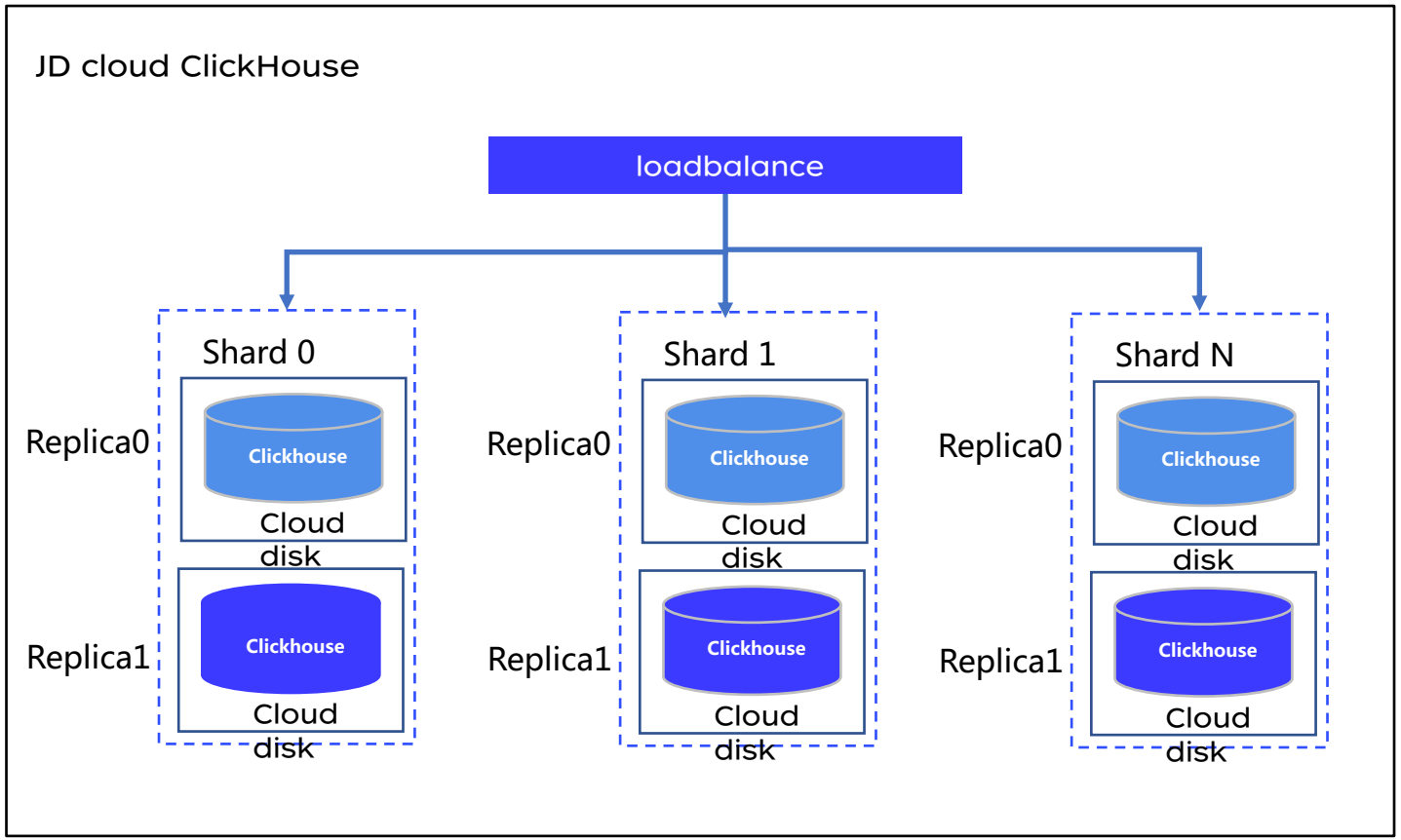
→ Bring the ultimate experience to users

High Availability of Services and High Reliability of Data

JDT 京东科技

ClickHouse

- ✓ Users can customize the number of replicas. The replicas have no master-slave concept. Each replica can be read and written. The downtime of one replica will not affect the read and write of the entire cluster, except for distributed DDL.
- ✓ Fault self-healing, when the probe detects that the node is abnormal, it will automatically restart
- ✓ When the physical machine is unavailable, using the cloud disk only needs to unload and load it to a new machine without re-synchronizing data
- ✓ Cloud disk data is highly reliable, with 0 data loss
- ✓ Node scheduling strategy, different copies of the same shard will not fall on the same physical machine, and different replicas of the same shard can be deployed across availability zones



JDT 京东科技

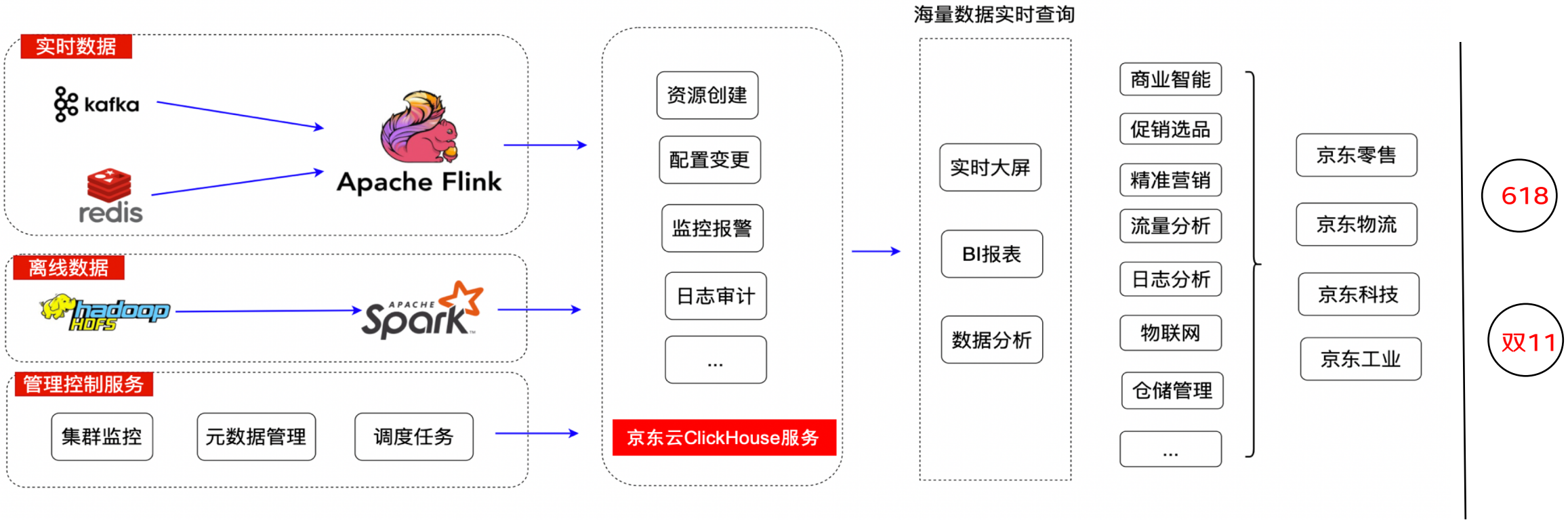
02

JD Cloud-Native ClickHouse Application Practice

JD Cloud-Native ClickHouse Application Practice

JDT 京东科技

Hundreds of businesses such as promotion selection、advertising real-time bidding、logistics analysis、report analysis、Internet of Things、business intelligence and etc.

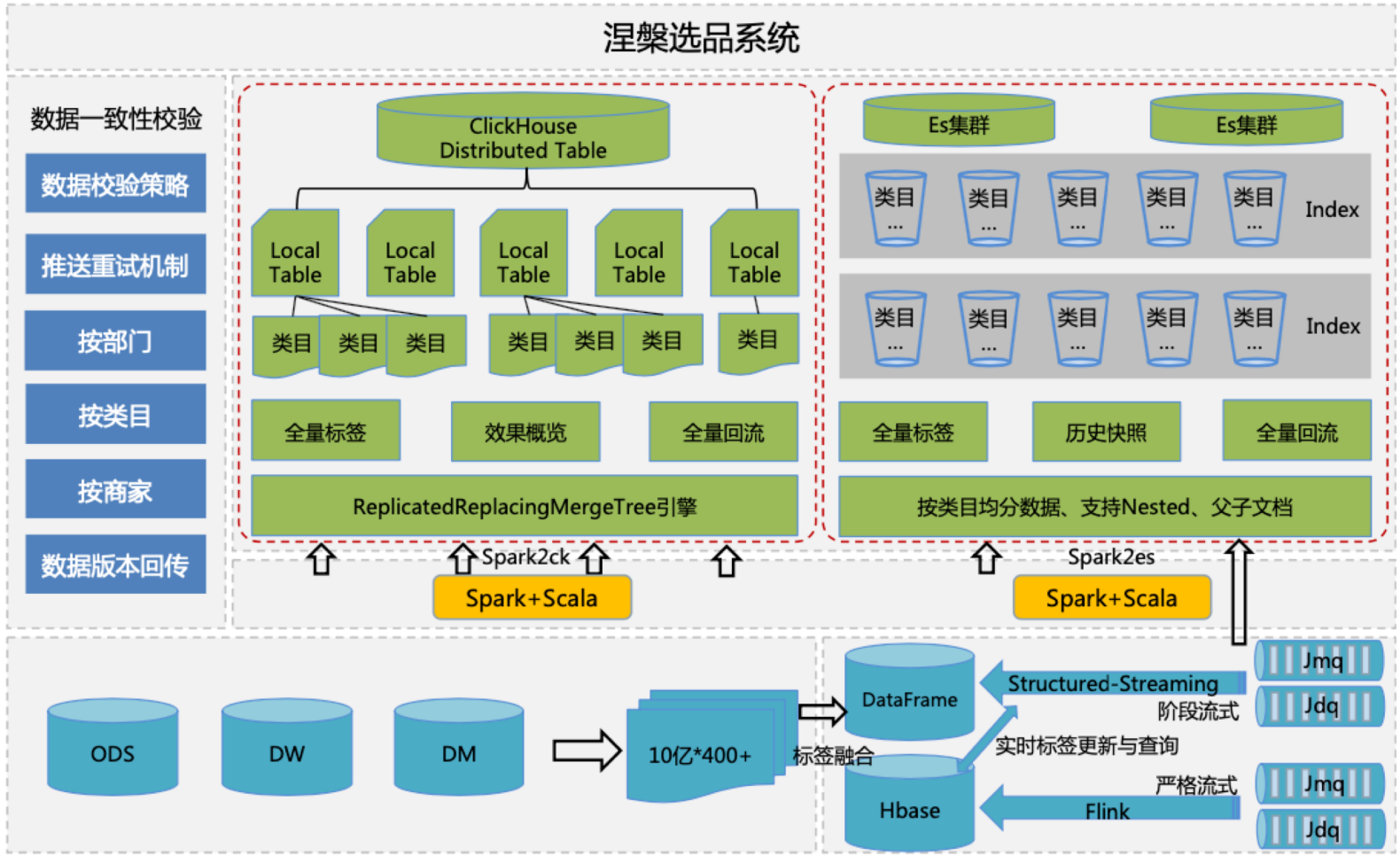


JD Cloud-Native ClickHouse Application Practice

JDT 京东科技

ClickHouse

Practice of retail nirvana product selection system:Build the underlying capabilities of products, open up the reporting and delivery process、realize the online、regular and intelligent selection of products; through multi-party collaborative Inventory、fully express multi-will of marketing、category、operation/purchasing and etc.



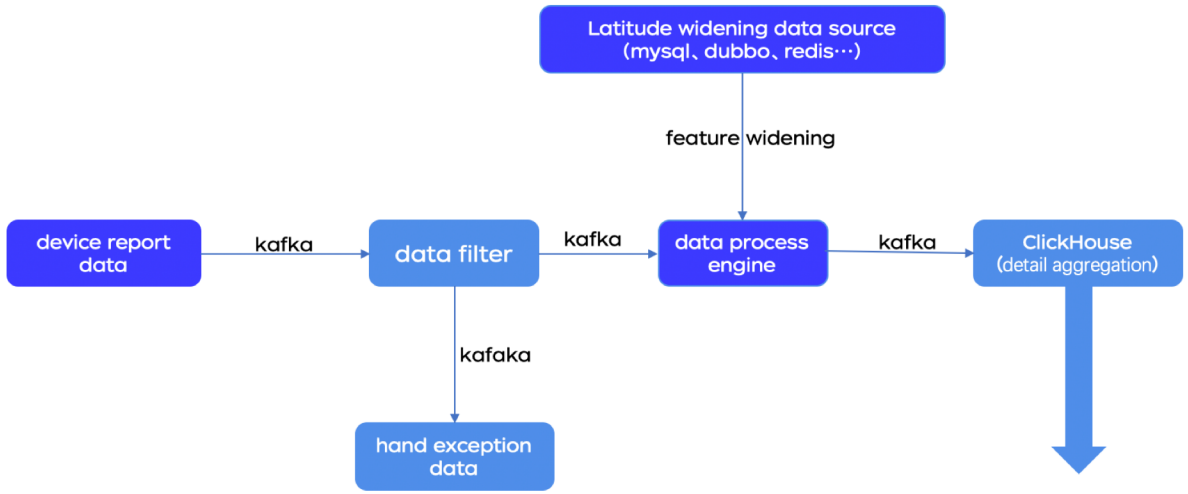
Solve business pain points

1. The business needs to check the coverage of the target commodity pool category, category, merchant, etc. in seconds. More than 400 columns of data about 10 billion are updated daily;
2. High performance requirements, concurrency of 200, and tp99 in milliseconds.
3. The query statement is complex, and the multi-dimensional sales volume, PV and other indicators analysis of commodities in the commodity pool need to be screened.

solution

1. Use CH distributed table + local table to write concurrently at the same time, the writing speed is fast, and billions of data are written in minutes;
2. CH performance meets the requirements for writing and querying
3. CH supports complex scene queries such as diversity, topN, and window sorting, and supports multiple statistical requirements such as multi-dimensional aggregation, drill-up, drill-down, and deduplication.

JD Cloud-Native ClickHouse Application Practice



ClickHouse

JDT 京东科技

Kafka engine table + materialized view + CH local table + CH distributed table

Business Features

Industrial Internet platform, used to import real-time on-site industrial data into the platform for analysis, and do data intelligence work

Challenges

- large-scale customers have hundreds of thousands of instruments and equipment, and the reporting frequency ranges from 1 to 60 per minute
- High real-time query requirements: most customers use large-screen real-time applications as real-time dashboards, and keep track of the dashboards at any time such as adding a new field.
- The customer globally updates the device, resulting in the need to introduce new fields.

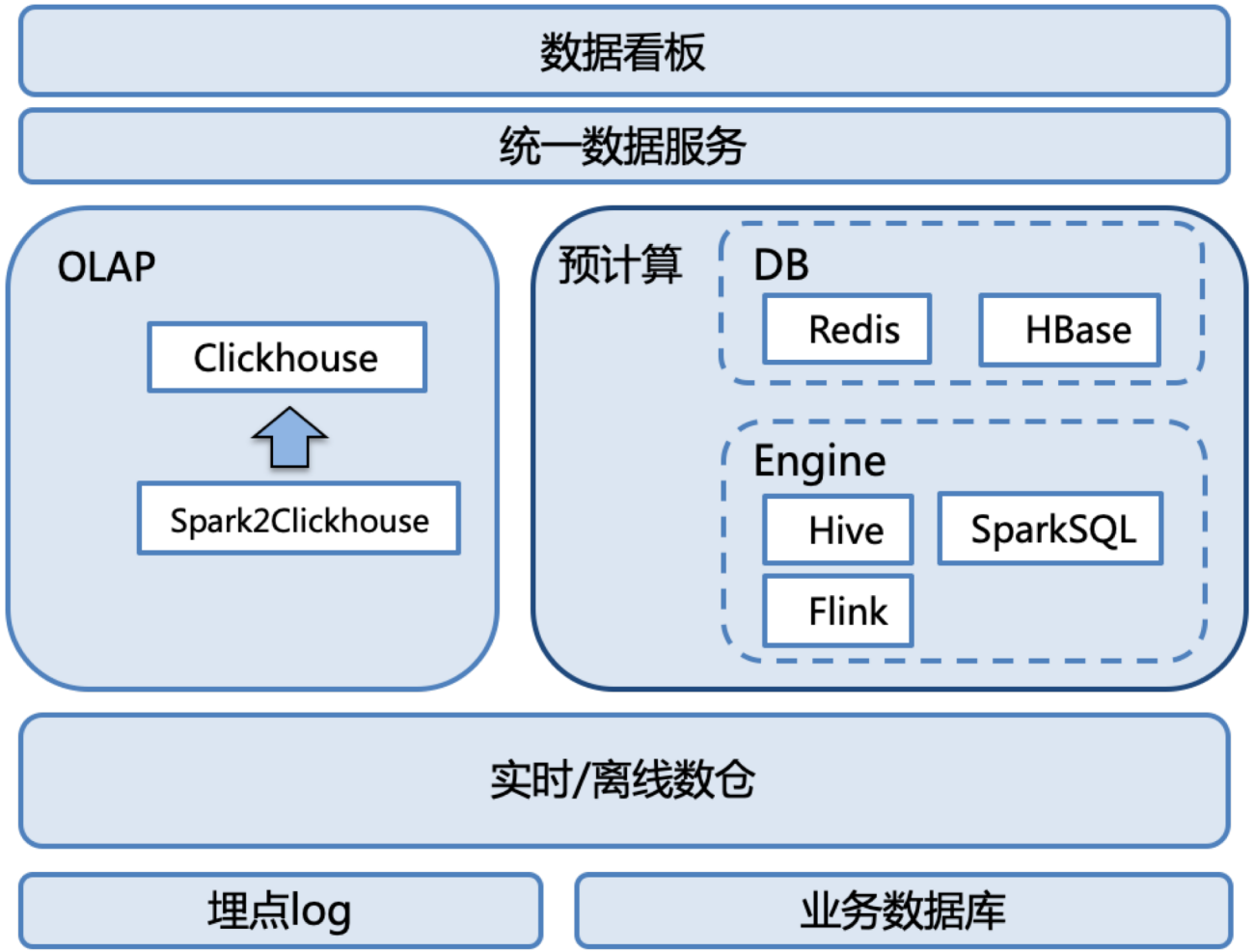
solution

- Support stable and high-throughput data writing, and meet the basic storage requirements for middle-stage construction
- In the face of massive data, the data query speed is extremely fast and the data compression rate is extremely high, saving disk storage
- Columnar storage has natural advantages for addition and subtraction fields

JD Cloud-Native ClickHouse Application

JDT 京东科技

JD Golden Eye Data Display Board: Golden Eye is a data Kanban product for the internal procurement and sales system of JD.com. It provides procurement and sales/operation personnel with data in various dimensions such as main site business, international site business, vertical business traffic, commodities, and transactions, and supports business parties. Assess the current situation of sales, identify sales problems in time, and formulate marketing strategies.



Challenges

- Purchasing and sales personnel need to analyze commodity sales data in the past two years, and analyze dozens of dimensions including stores, commodities, transactions, traffic, finance, etc., and the query time must be responded within seconds;
- The organizational structure of the department to which the SKU product belongs is often changed, and the purchasing and sales appeal can see the operation of the sold product after the change in time. It is necessary to retrospect and update the historical data of the relevant fields of the department to which the SKU belongs on a daily basis;

solution

- In view of the large amount of commodity detail data, a materialized view based on CK cluster is designed to improve query efficiency. CK has better support for multi-dimensional analysis based on large-width tables, and supports query data in any time range.
- Data dictionary + real-time table solves the problem of regular batch refresh of massive data

JDT 京东科技

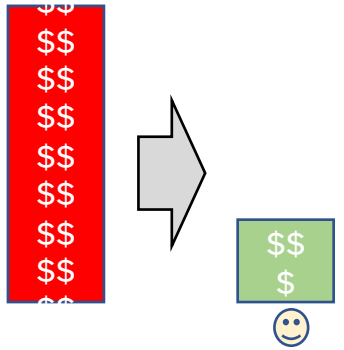
03

JD Cloud-Native ClickHouse Phase II and Roadmap

JD Cloud-Native ClickHouse 2.0

JDT 京东科技

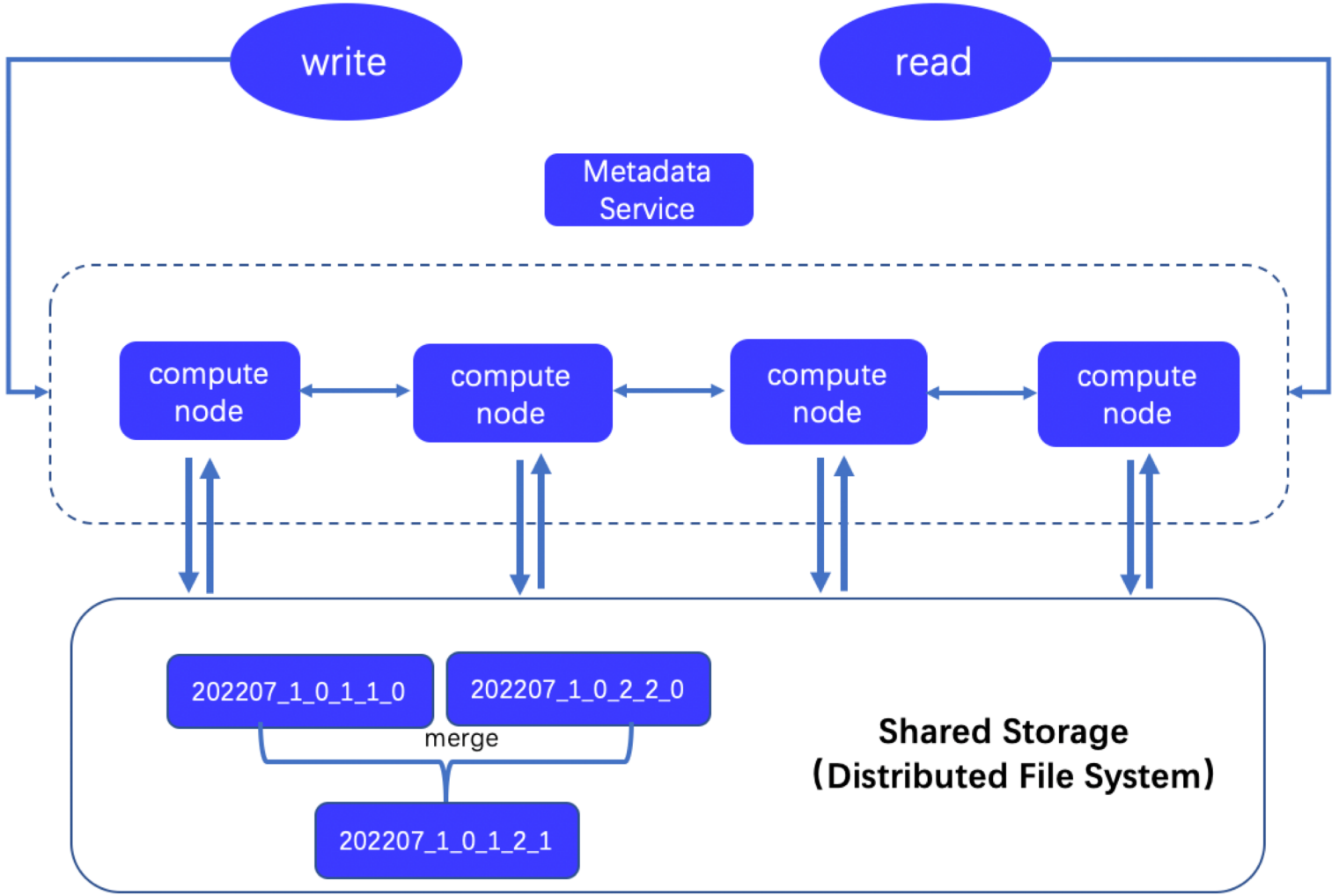
- High total cost of ownership, cost is a priority for customers
- The cost of expansion and contraction is high, storage and computing are coupled together, and data redistribution is required
- If one shard is down and the entire cluster will unavailable
- During the big promotion period, the technical challenges, high costs and risks of emergency expansion
- The separation of cloud-native and storage-computing has become very popular in recent years. Based on the open source ClickHouse, drawing on the design ideas of systems such as Snowflake and BigQuery, a more cloud-native real-time data warehouse has been launched.



JD Cloud-Native ClickHouse 2.0

JDT 京东科技

Architecture



JD Cloud-Native ClickHouse 2.0

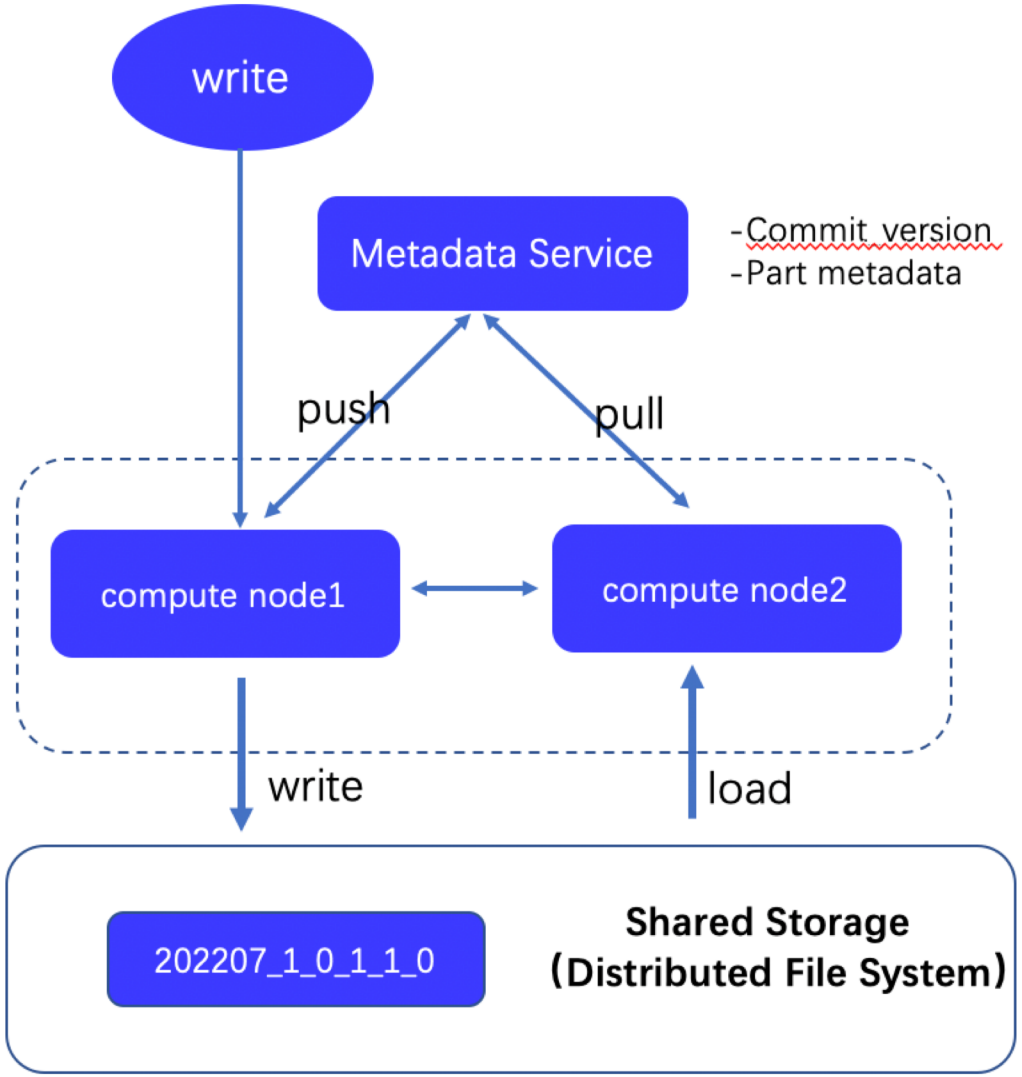
JDT 京东科技

- Distributed write、distributed read (distributed cache Consistency)
- Distributed background tasks
- Extreme elastic expansion and contraction
- Cluster management and failover
- Compatibility support for native clickhouse usage
- Optimization of Distributed File System

JD Cloud-Native ClickHouse 2.0

JDT 京东科技

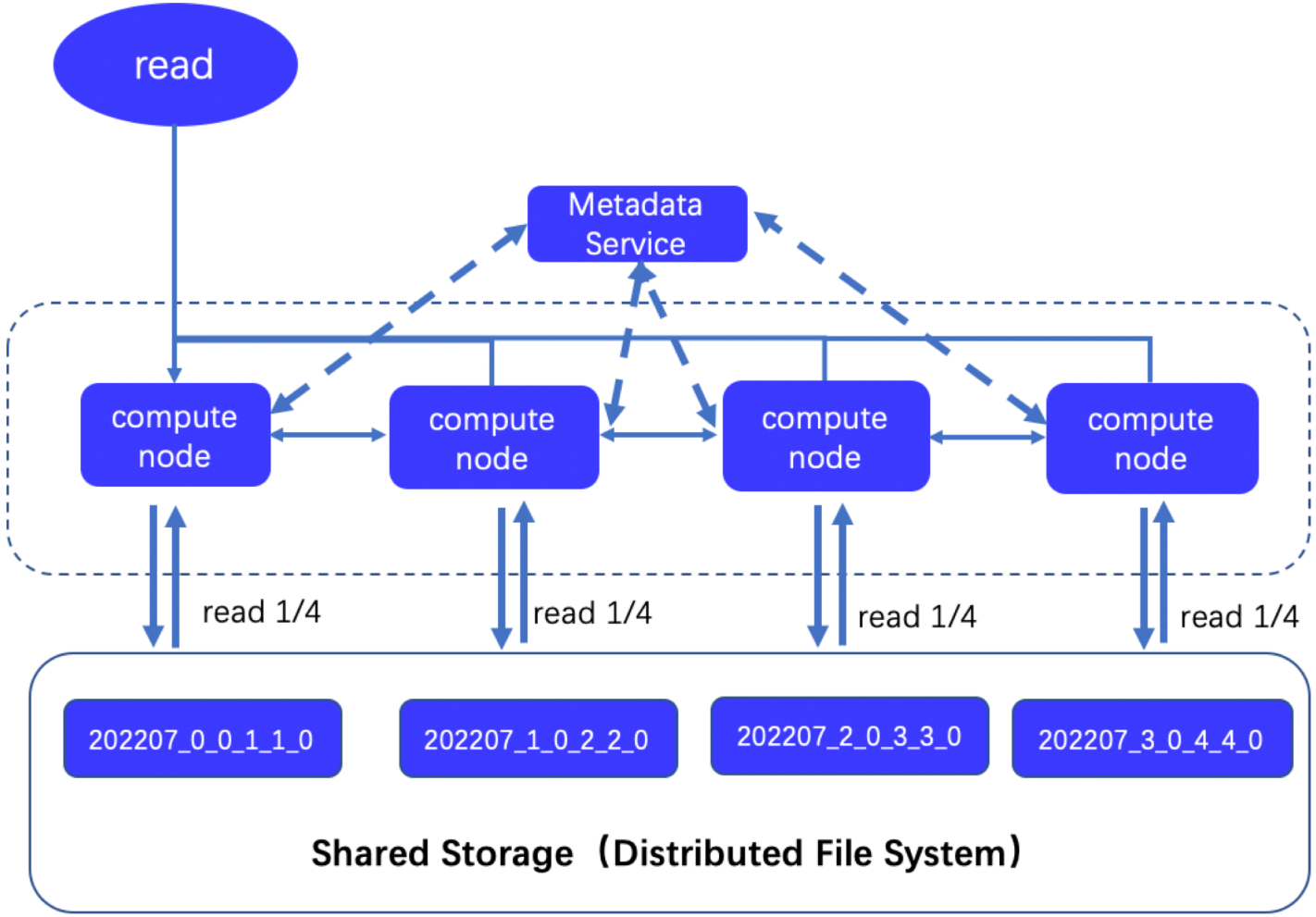
distributed write



JD Cloud-Native ClickHouse 2.0

JDT 京东科技

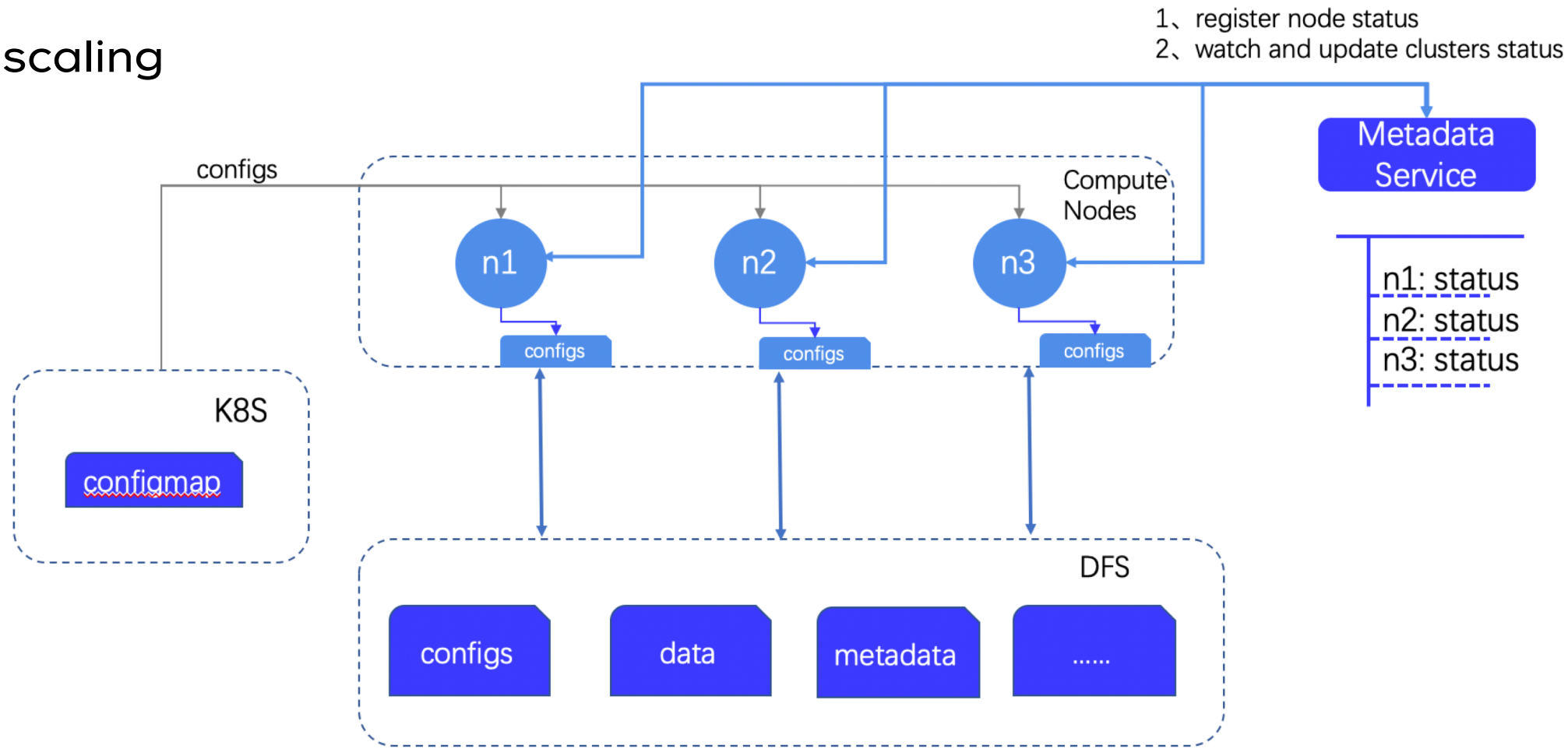
distributed read



JD Cloud-Native ClickHouse 2.0

JDT 京东科技

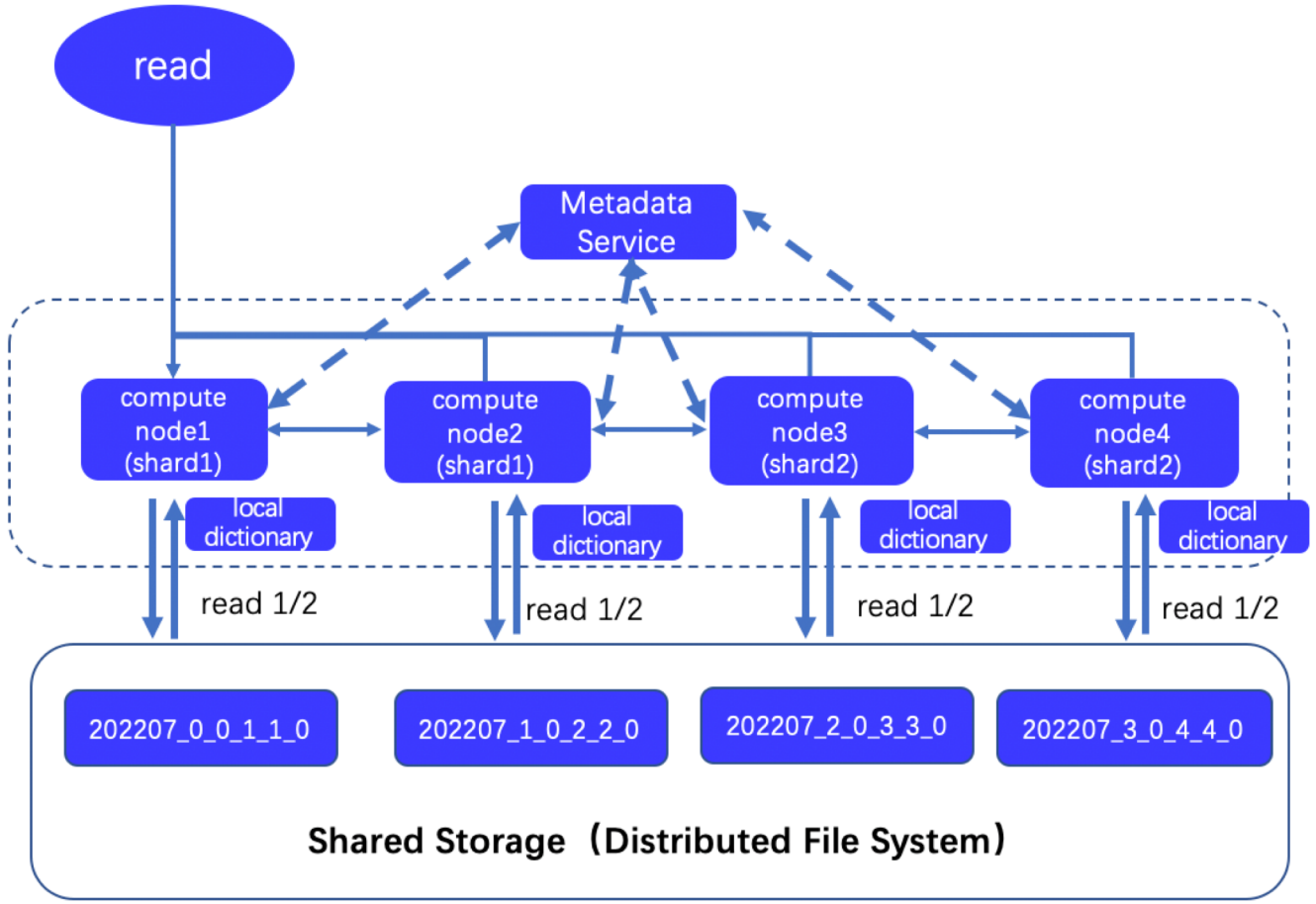
elastic scaling



JD Cloud-Native ClickHouse 2.0

JDT 京东科技

Local memory table support, here takes the local dictionary table as an example



JD Cloud-Native ClickHouse 2.0

JDT 京东科技

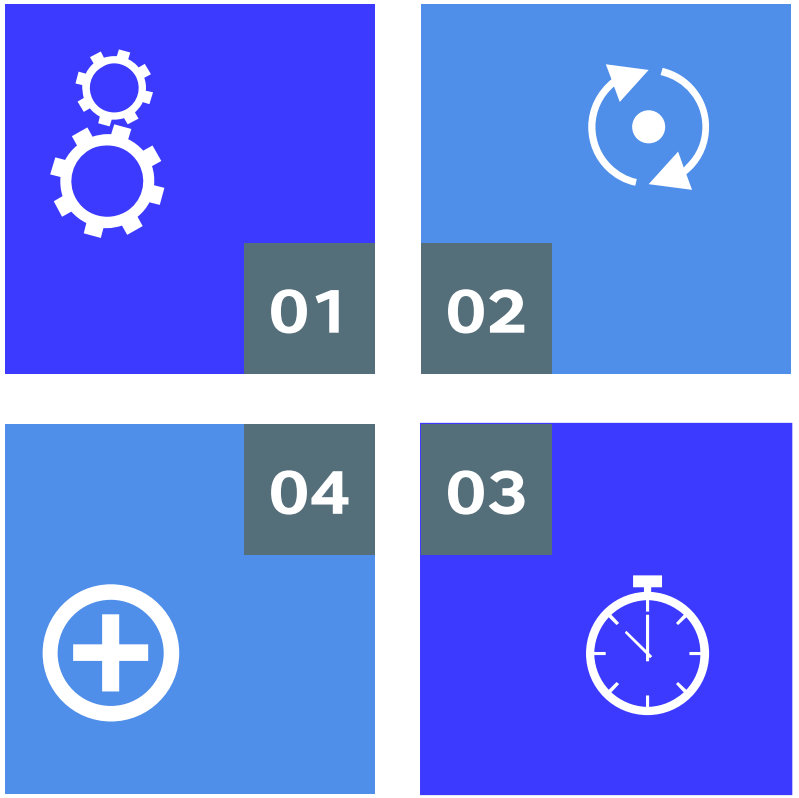
2.0 features

High resource utilization and low cost

Storage costs are reduced by at least 50%, and computing nodes can be expanded and contracted at will with business peaks and lows, greatly improving resource utilization and reducing resource pre-occupation requirements. There is no need to create a large-scale cluster at the beginning. Shared storage is charged according to usage and automatically expands.

Highly available and scalable

The computing nodes are stateless, and storage and computing can be elastically expanded separately.



easy to maintain

Increase cluster state management, unified shared storage (no need to worry about data reliability and data security), provide fault recovery mechanism, and simple process of expansion and contraction (shared metadata)

Compatible with open source, reuse ultra-high performance

Compatible protocol, syntax, database storage format

JD Cloud-Native ClickHouse 2.0

JDT 京东科技

Future plan

Computing resource isolation

Different users define and isolate computing resources without affecting each other, which can effectively ensure the execution efficiency of important services and high-priority users.

Performance optimization and improvement

Calculations are pushed down to the data storage layer, such as predicates and aggregations, to achieve near-storage computing acceleration. Data from storage to computing network acceleration, distributed join optimization

Support multi-cloud deployment

It can support multi-cloud deployment and privatization deployment, and users do not need to worry about being bound by a cloud vendor.

Perfect ecological tool

Support data synchronization and extraction of various mainstream data sources. Data migration can be completed with one click;

JDT 京东科技

Thanks

JD ClickHouse

Wechat