

# 基于ClickHouse企业版构建可观测平台

晓屿 | 产品经理 | 数据库产品与事业部

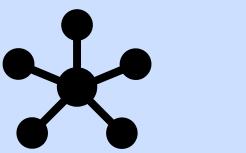
# 可观测技术：现代运维理念的核心基石

可观测（Observability）是现代运维理念，强调系统在运行时应具备全面的、深入的、可理解的状态获取能力。通过收集和分析系统的各种可观测数据，构建全方位、立体化的监控与分析体系，实现快速问题定位、预防性维护以及持续优化。

## 指标 ( Metric)

对系统或程序运行情况的量化描述，例如：

- 交易量、访问量
  - CPU、内存使用率、网络带宽
  - 响应时间、访问延迟
- 特点：**
- 随时间变化（具备时序特征）
  - 数据量相对较小
  - 需要实时监控



## 链路 ( Trace)

一次请求在执行过程中，调用链路上每个节点的处理情况记录，一般用于：

- 查看请求调用路径
- 分析服务间耗时

**特点：**

- 每条记录有明确的链路id
- 数据量相对较大
- 需要能快速定位到对应记录



## 日志 ( Log)

系统运行产生的带时间戳的离散记录。例如：

- 系统运行日志
- 功能开发调试日志
- 埋点行为日志

**特点：**

- 数据格式多样，结构化数据&json混杂
- 数据量很大



## 互联网应用开发

应用开发日志、运维指标监控



## 社媒&广告行业

行为日志分析、日活监测



## 金融行业

K线、交易日志、风控平台搭建



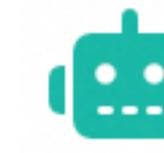
## 汽车制造业&智驾

车控链路追踪、客户行为日志分析



## 游戏行业

日活/留存/转化监控，行为日志分析



## AI大模型&应用开发

模型调用情况监控、模型&应用调试

# 数据库：支撑可观测能力的核心基础设施

数据库作为支持可观测能力的核心基础设施，业务层面对数据库的性能、成本、易用性、运维等方面都提出了一系列的要求，而 ClickHouse 借助其在性能、压缩率和对各开源技术栈高度兼容的优势，成为了可观测场景最受欢迎的技术栈之一。

- 支持海量时序数据批量高速写入
- 毫秒级查询响应能力
- 支持百TB~PB级数据低成本存储

**高性能**

- 兼容市面上大多数常见的采集和可视化工具
- 易于部署

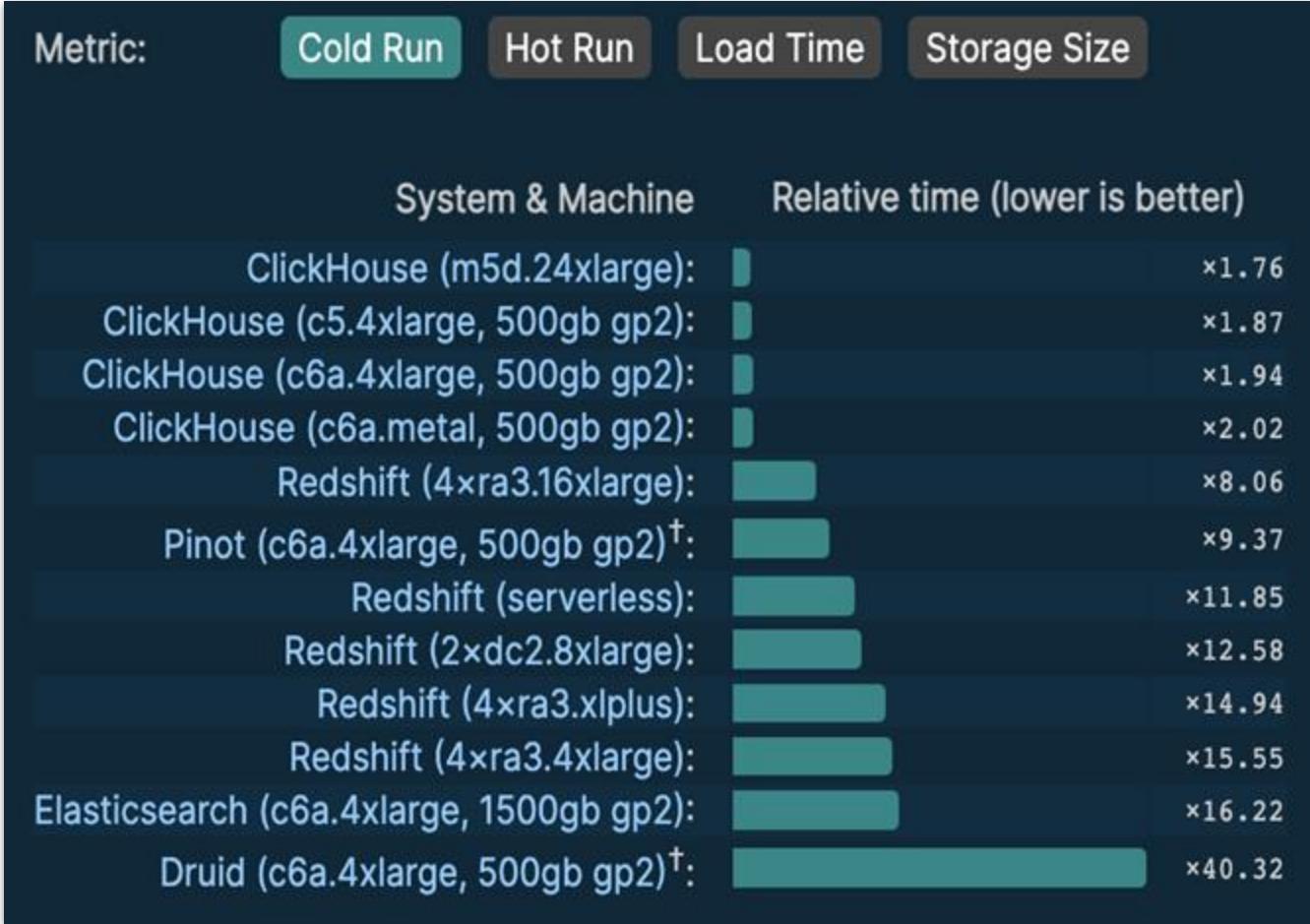
**易于部署**

- 灵活水平扩展、动态资源调配
- 多租户资源隔离

**灵活管理**

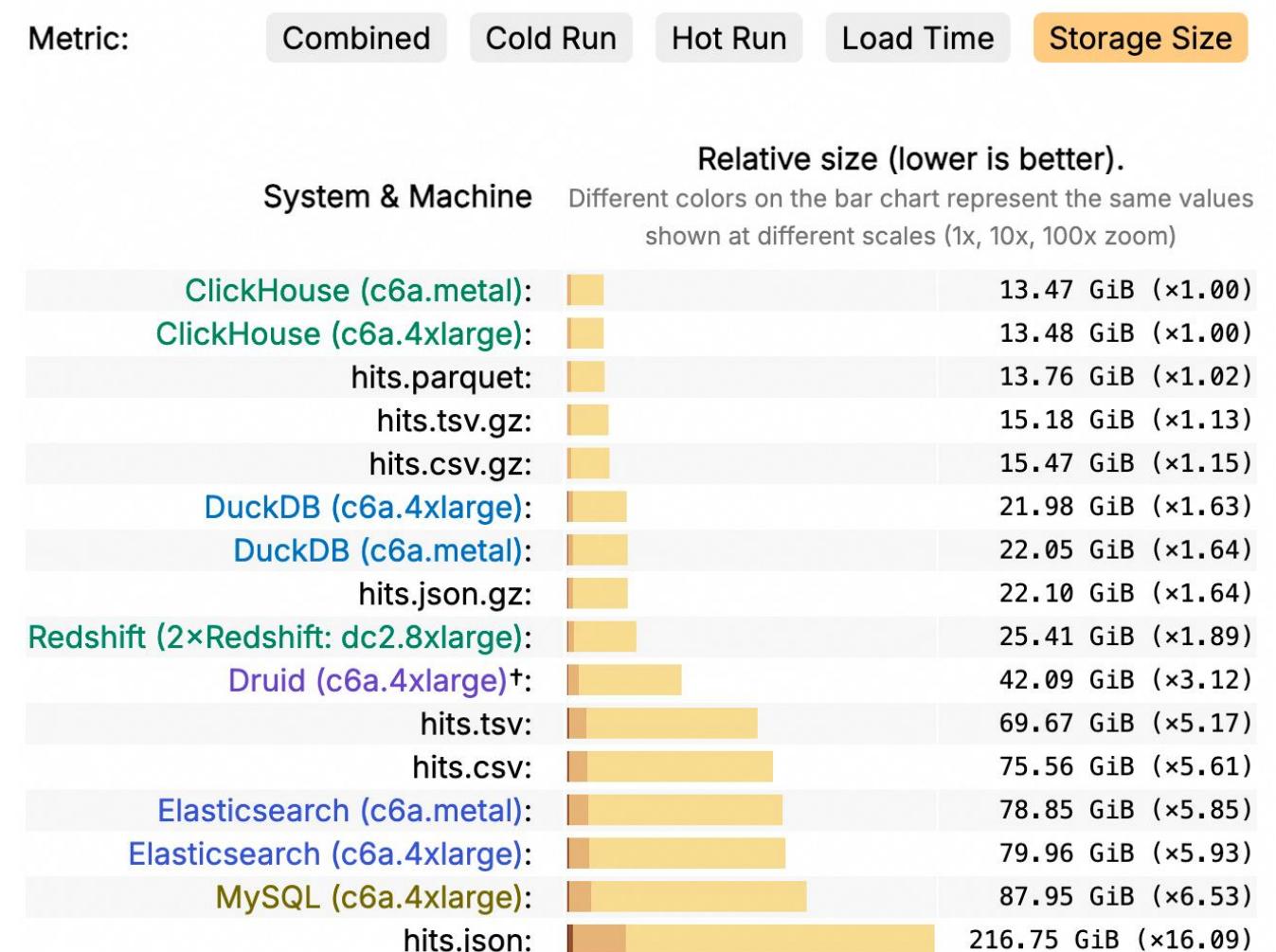
## ClickHouse：超高读写性能

1. 超高查询性能：900+倍于MySQL, 50+倍于druid
2. 稳定高效的批量写入性能：单节点50-200MB/s, 随节点数线性增加



## ClickHouse：超高数据压缩率，超低数据存储成本

1. 支持LZ4、ZSTD等多种数据压缩算法
2. 超高数据压缩率：5+倍于ElasticSearch, 10+倍于MySQL



# 数据库：支撑可观测能力的核心基础设施

数据库作为支持可观测能力的核心基础设施，业务层面对数据库的性能、成本、易用性、运维等方面都提出了一系列的要求，而 ClickHouse 借助其在性能、压缩率和对各开源技术栈高度兼容的优势，成为了可观测场景最受欢迎的技术栈之一。

- 支持海量时序数据批量高速写入
- 毫秒级查询响应能力

**高性能**

- 支持百TB~PB级数据低成本存储

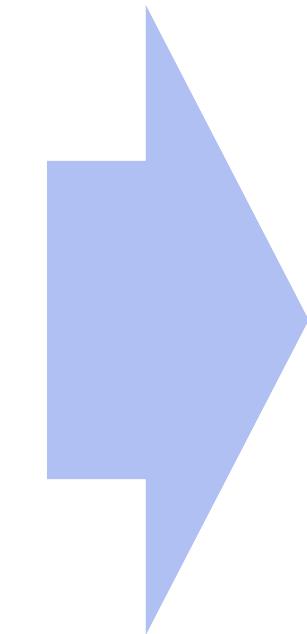
**低成本**

- 兼容市面上大多数常见的采集和可视化工具
- 易于部署

**易于部署**

- 灵活水平扩展、动态资源调配
- 多租户资源隔离

**灵活管理**



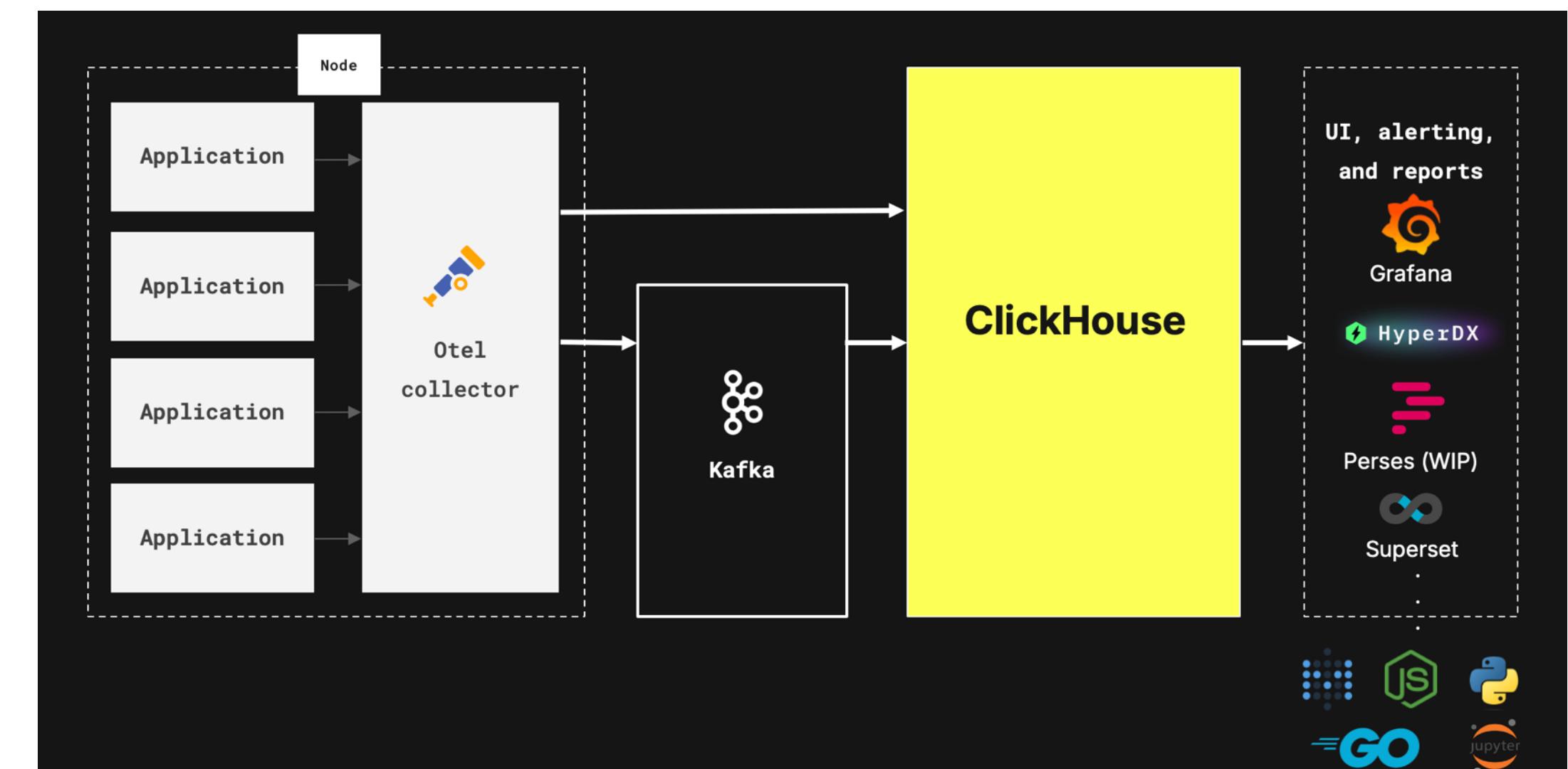
## ClickHouse：超高读写性能

1. 超高查询性能：900+倍于MySQL, 50+倍于druid
2. 稳定高效的批量写入性能：单节点50-200MB/s, 随节点数线性增加

## ClickHouse：超高数据压缩率，超低数据存储成本

1. 支持LZ4、ZSTD等多种数据压缩算法
2. 超高数据压缩率：5+倍于ElasticSearch, 10+倍于MySQL

## ClickHouse：兼容多种数据源&目标端



# 01 / 利用CLICKHOUSE企业版为可观测业务提效

## 更好的资源可扩展性

- 存算分离，增删节点无需停读停写，分钟级完成
- 支持计算和存储资源serverless，资源使用量随业务波动秒级自动弹升弹降
- 集群的写入性能和并发查询qps上限随节点数增加线性增长

## 更高的容灾能力

- 集群默认3keeper&至少双server架构，保障元数据安全
- 存算分离，单server故障不影响集群其他server正常对外提供服务，不影响整体可用性
- 支持多可用区部署架构，提供跨机房的故障容灾能力

## 更高的性价比

- 计算资源基于负载秒级弹升弹降，可节省约50%计算资源
- 存储资源基于实际使用量付费，无需预留存储空间，可节省约20%存储资源采购量
- 共享对象存储，存储采购单价相比自建云盘版降低80%+

## 更稳定&便捷的链路支持

- 支持flink、kafka同步数据
- 支持oss外表
- 支持通过dts从rds、polar同步数据
- 支持通过dts从sls投递数据（邀测中）
- 控制台内嵌的一站式可观测能力（ClickObserve）

# 阿里云和 ClickHouse, Inc 战略合作打造 ClickHouse 企业版

阿里云智能集团  
ALIBABA CLOUD INTELLIGENCE GROUP

国内独家  
唯一可提供存算分离闭源内核的国内云厂商

- 2023年3月：
  - 阿里云和ClickHouse Inc.达成**国内独家**战略合作
- 2023年10月23日：阿里云ClickHouse企业版公测上线**独家**
  - SharedMergeTree引擎 – 存算分离架构
  - Parallel replica
  - Lightweight update **闭源独占1.5年**
- 2024年4月26日：阿里云ClickHouse企业版正式商业化**独家**

## ClickHouse, Inc. and Alibaba Cloud Announce a New Partnership

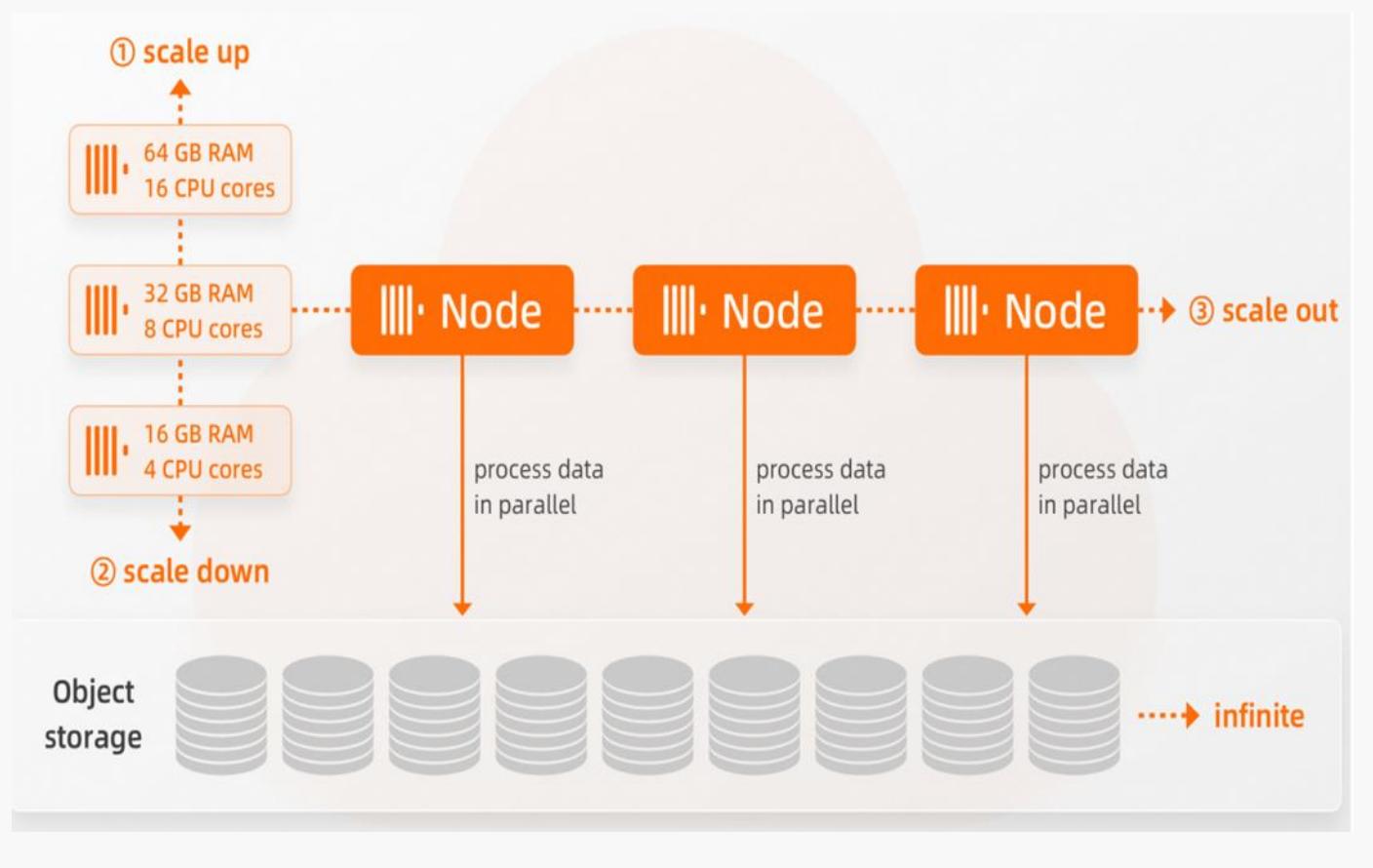


Nick Peart  
Mar 24, 2023

SAN FRANCISCO - March 24, 2023 - Today, ClickHouse, Inc. creators of ClickHouse online analytical processing (OLAP) database, and Alibaba Cloud, the digital technology and intelligence backbone of Alibaba Group, announced a partnership that will enable Alibaba Cloud to offer ClickHouse as an enterprise, first-party service on its platform. This partnership is an exclusive agreement between ClickHouse, Inc. and Alibaba Cloud in mainland China to offer a joint first-party enterprise service in APAC.

联合研发  
客户第一，共同推动客户需求落地

- 联合研发：
  - 先进内核：原厂研发存算分离内核
  - Serverless：阿里自研秒级serverless
  - 国内云生态：对接阿里云上下游生态，支持基于DTS的从MySQL/SLS的数据导入
- 定期syncup
  - 同步高风险内核缺陷，及时修复
  - 针对云环境的内核参数调整与优化



定向支持&生态共建  
阿里&ClickHouse Inc.联合提供专家服务

- 专家服务
  - 提供覆盖全球的专业技术支持服务
  - 在客户支持上开展专项合作，深入业务场景提供专家级推荐及优化建议
- 生态共建
  - 中文社区
  - 季度维度meetup
  - 云栖大会



# ClickHouse 企业版：云原生存算分离架构

## ➤ Stateless worker

- 基于ClickHouse Keeper实现元数据同步
- 每个计算节点都是**存有全量元数据、可独立对外提供服务的rw节点**

## ➤ Parallel Replica

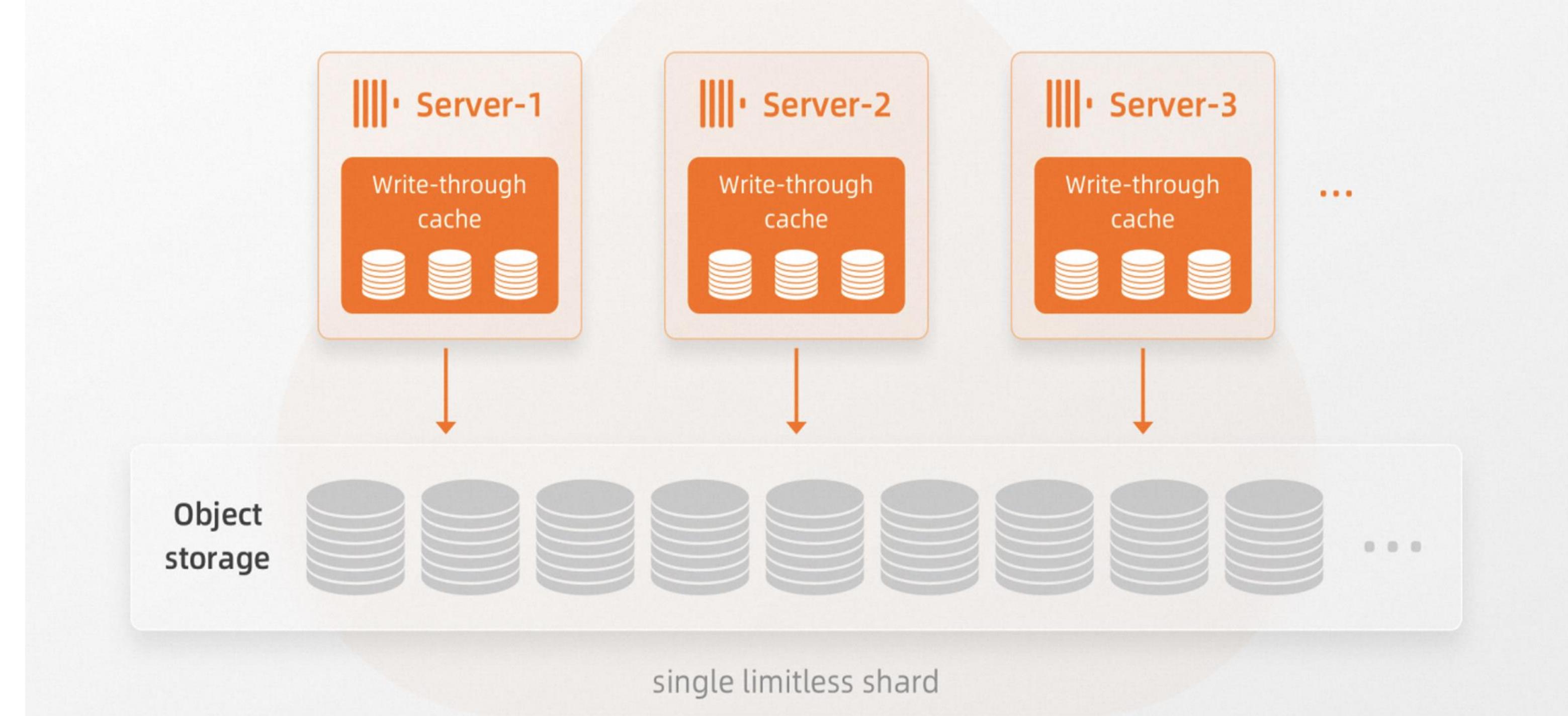
- 可选开启，支持多节点并行查询
- 开启后，优化器自动识别当前sql是否适用

## ➤ 本地缓存

- 相对OSS存储提供2-10倍热查询优化
- lru数据淘汰，缓存大小 = Server内存大小  $\times$  7

## ➤ 共享对象存储

- OSS存储：已上线
- L1、L2云盘：930上线
- L1、L2+OSS冷热分层：on the way



	OSS	ESSD_L1	ESSD_L2	缓存
集群读写带宽上限	北上杭深： 读：100Gb/s; 写：20Gb/s 其他region： 读：5-10 Gb/s; 写：5-10 Gb/s	460*节点数(MB/s)	980*节点数(MB/s)	-
latency	10ms	0.9ms	0.9ms	0.2ms

# ClickHouse 企业版：云原生存算分离架构

## ➤ Stateless worker

- 基于ClickHouse Keeper实现元数据同步
- 每个计算节点都是**存有全量元数据、可独立对外提供服务的rw节点**

## ➤ Parallel Replica

- 可选开启，支持多节点并行查询
- 开启后，优化器自动识别当前sql是否适用

## ➤ 本地缓存

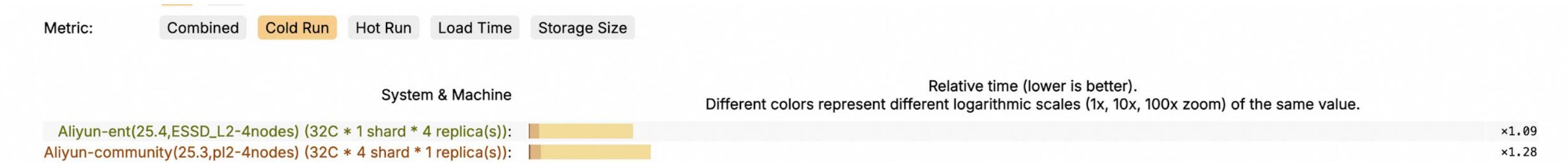
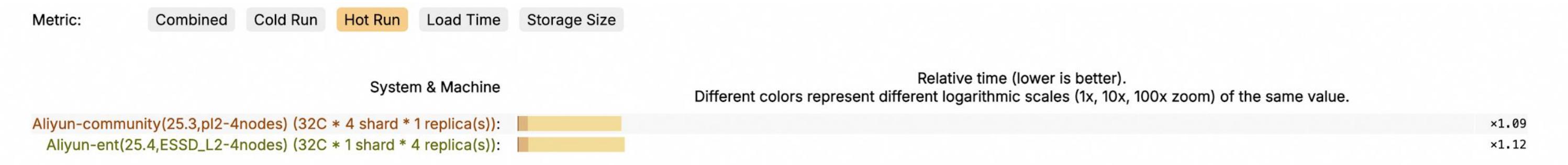
- 相对OSS存储提供2-10倍热查询优化
- lru数据淘汰，缓存大小 = Server内存大小 × 7

## ➤ 共享对象存储

- OSS存储：已上线
- L1、L2云盘：930上线
- L1、L2+OSS冷热分层：on the way

## 高性能对象存储 (9月底公测上线)

ClickHouse企业版增加了云盘存储类型，相对同版本同等级云盘的开源自建ClickHouse，热查询性能持平，冷查询性能高~20%（下图为同版本PL2开源自建ClickHouse和阿里云企业版ESSD\_L2等级存储的clickbench对比）。

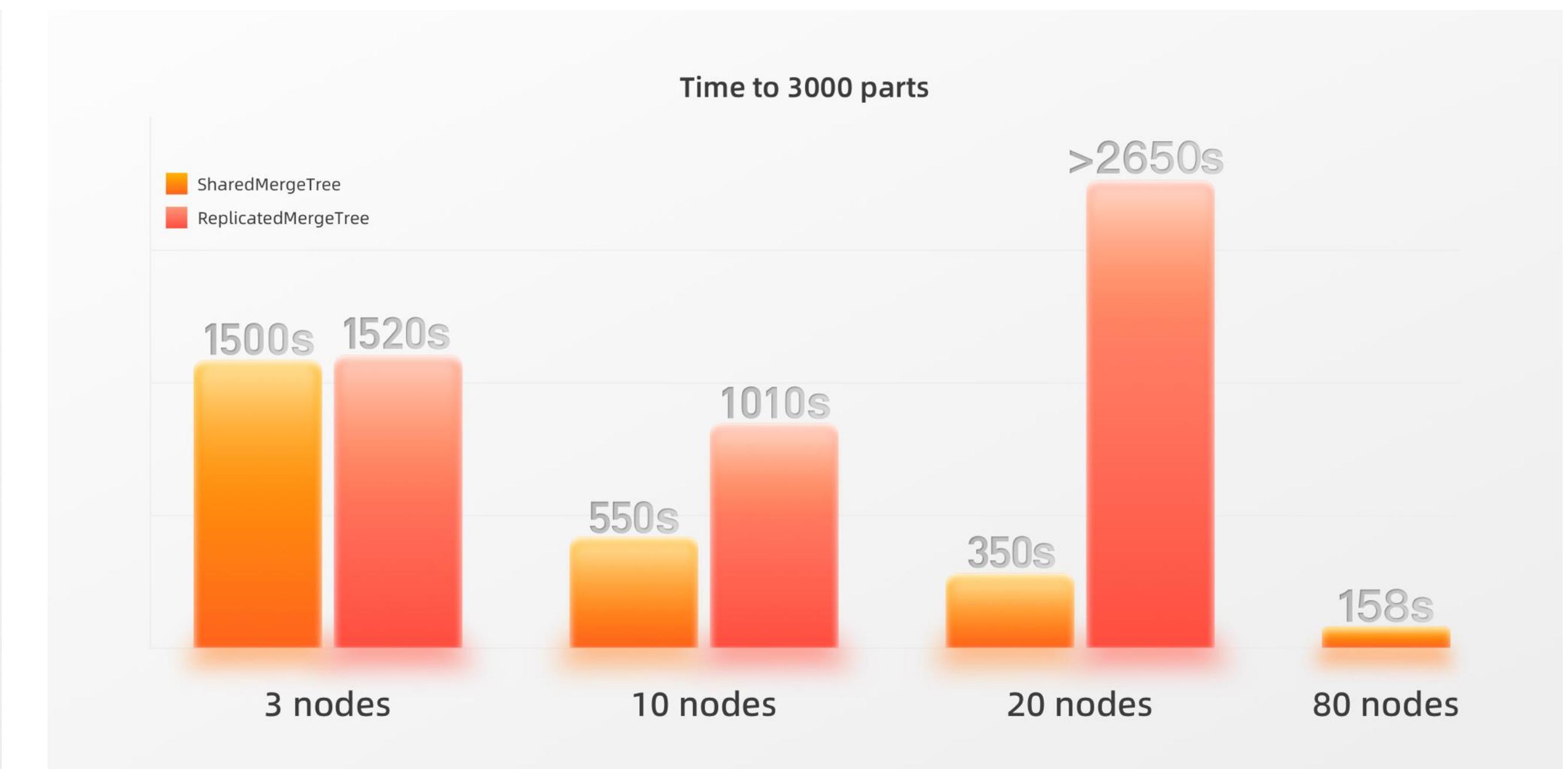
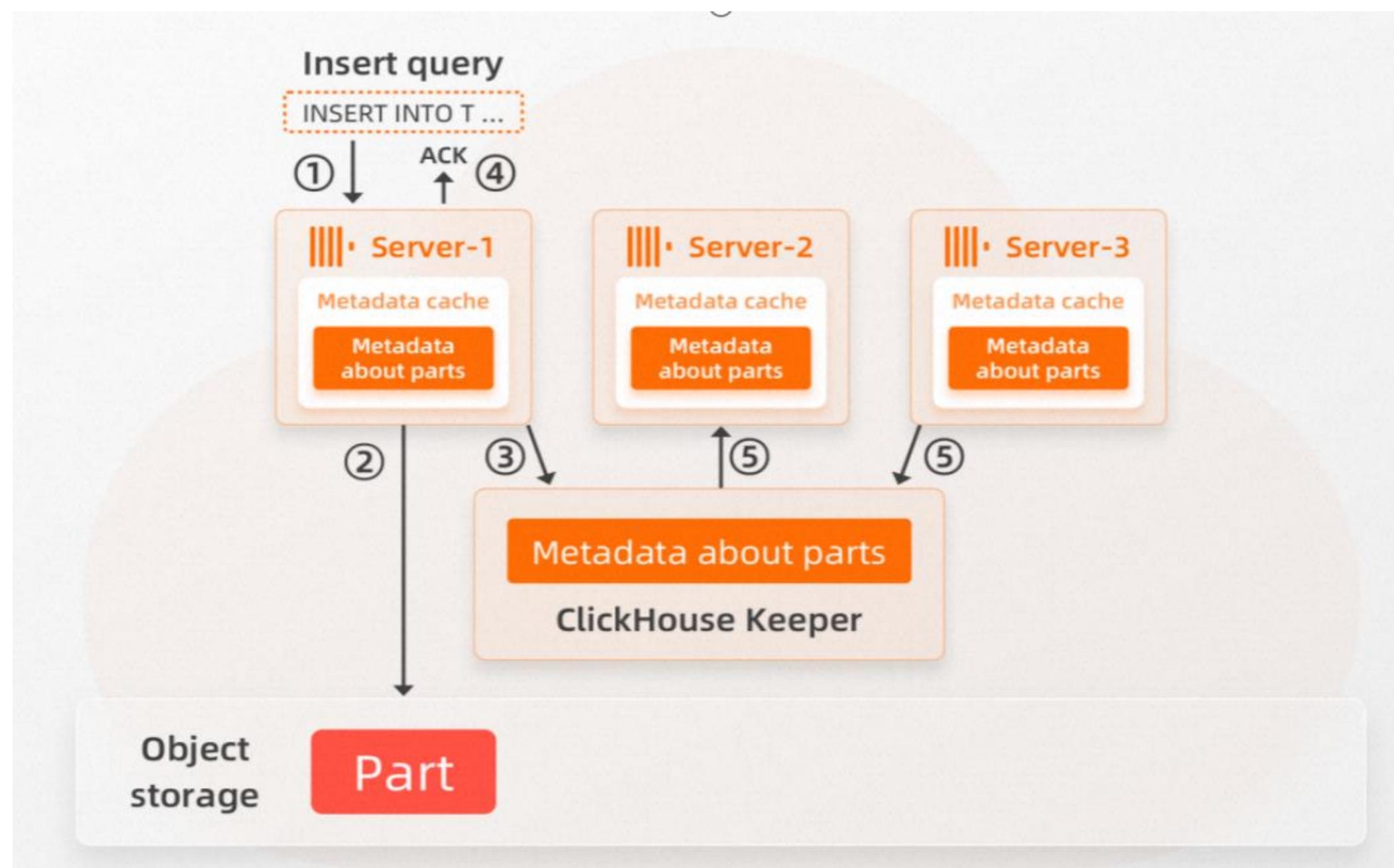


	OSS	ESSD_L1	ESSD_L2	缓存
集群读写带宽上限	北上杭深： 读：100Gb/s; 写：20Gb/s 其他region： 读：5-10 Gb/s; 写：5-10 Gb/s	460*节点数 (MB/s)	980*节点数 (MB/s)	-
latency	10ms	0.9ms	0.9ms	0.2ms

# ClickHouse 企业版：merge性能、集群qps随节点数增加线性扩展

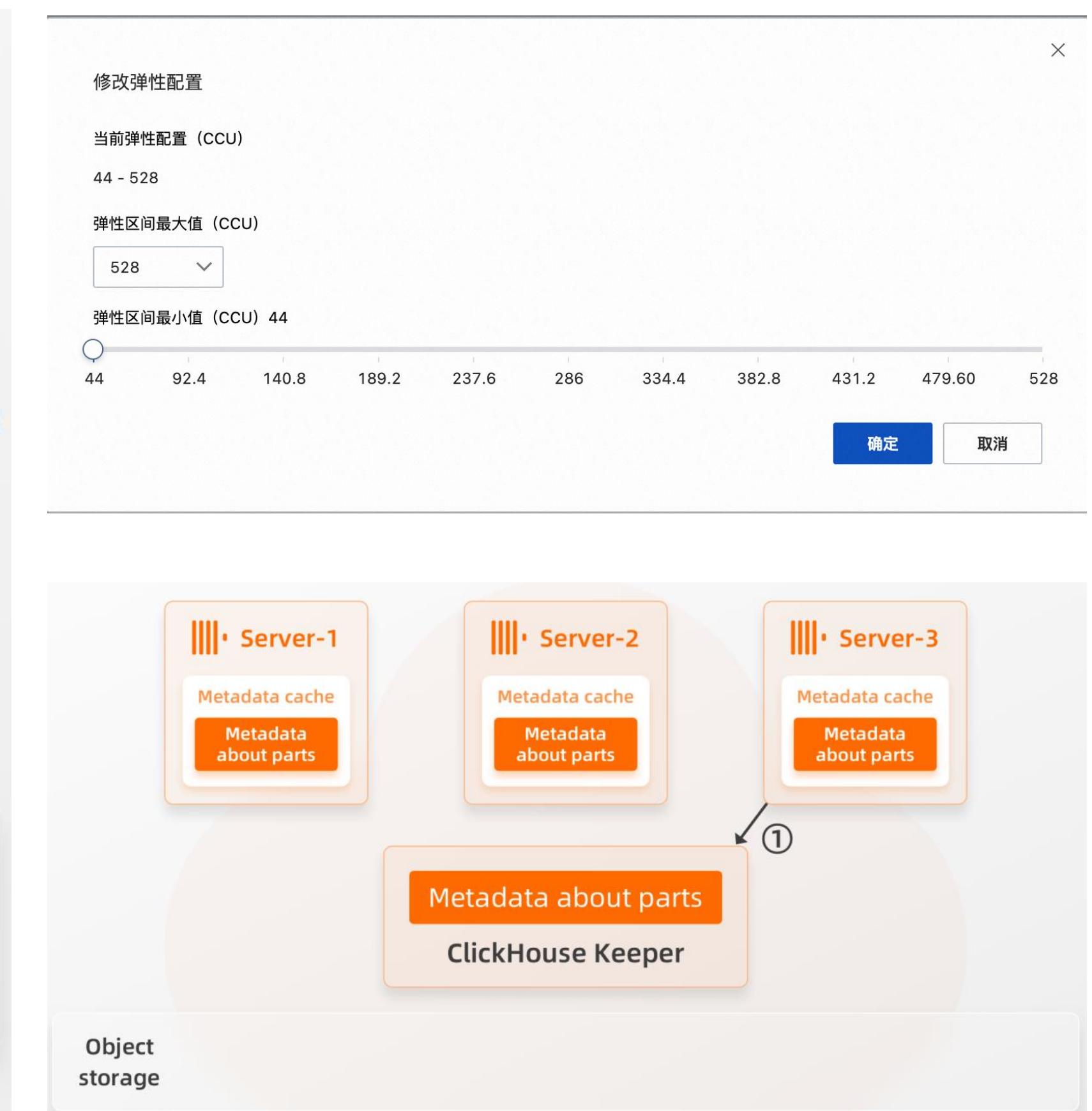
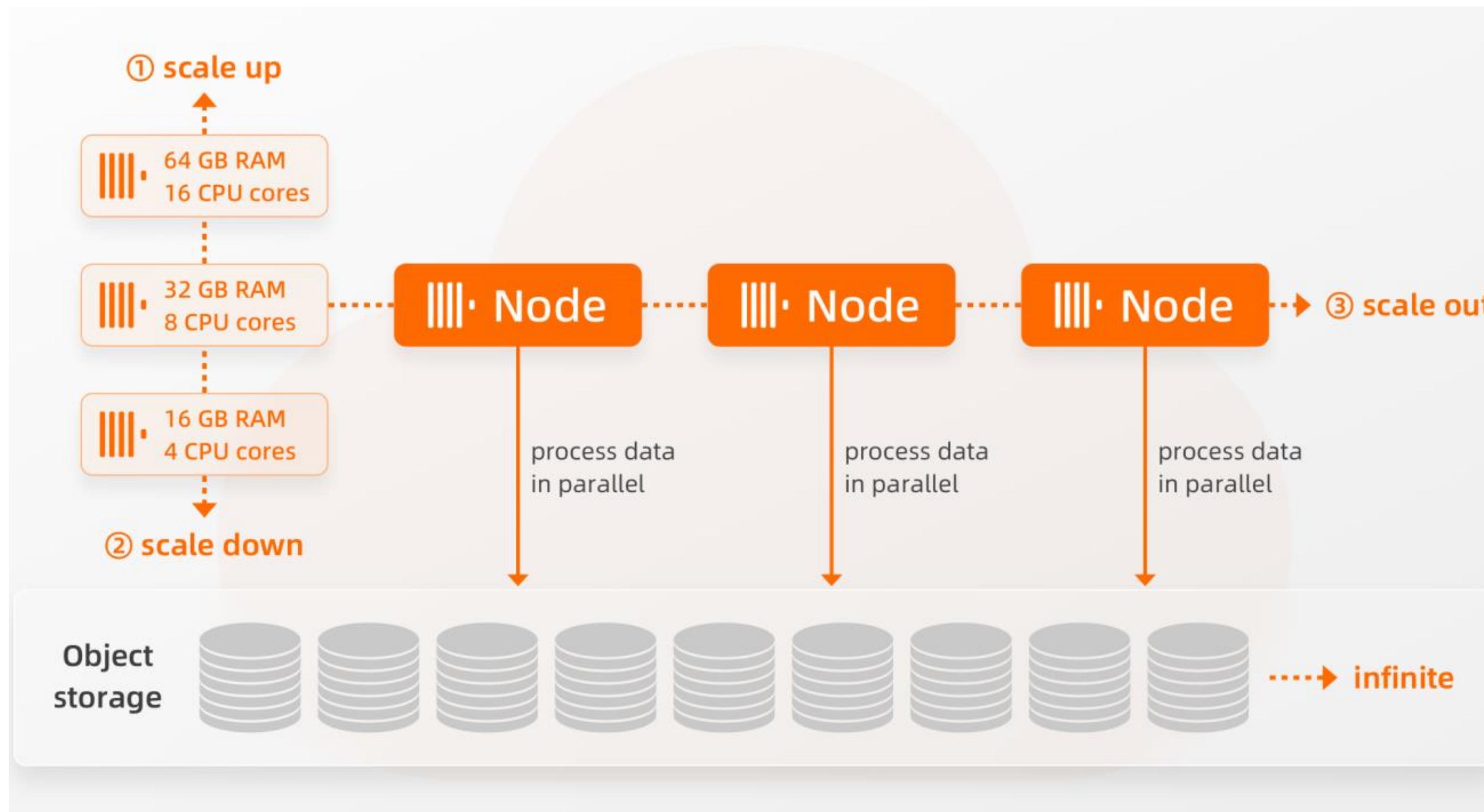
## 性能的可扩展性：

- 各节点数据同步操作更加轻量和高效，数据一致性强，无资源占用
- 更高的 Merge 处理效率：merge性能随节点数的增加线性扩展
- 更好的集群并发qps：单节点~1000qps，集群qps随着节点数增加线性增长（目前线上推荐集群节点数不要超过16个，白名单可开放32个）



# ClickHouse 企业版：分钟级水平扩展，扩展过程业务无感

1. 存储和计算分离，存储和计算的增加都可以独立进行
2. 无分片概念，增加节点无需数据迁移，且新增节点可以立即加入并行计算，扩展过程分钟级完成，扩展效率大大提升
3. 集群的水平扩缩容不影响集群的整体可用性



# ClickHouse企业版：基于负载的秒级serverless能力

- 秒级弹升，跟上业务负载，业务低峰释放资源，Pay-as-You-Go，有效降低企业成本支出。
- 滑动窗口自动检测cpu和内存负载情况，动态调整，降低运维压力。
- 支持保守型和激进型两种弹性策略，用户可根据业务的稳定性和成本诉求自主选择



弹性策略	开启条件	弹升策略	弹降策略
<p>策略类型：保守型。 适用场景：</p> <ul style="list-style-type: none"><li>• 业务低谷时段资源使用率低于30%的场景。</li><li>• 业务高峰和低谷时段资源使用率差距较大的场景。</li><li>• 对业务稳定性要求较高的场景。</li></ul>	<p>在新建集群时，默认采用保守型Serverless策略。</p>	<ul style="list-style-type: none"><li>• 触发条件：(以下条件满足一个即可)<ul style="list-style-type: none"><li>◦ CPU使用率在5s内的平均值超过60%。</li><li>◦ 内存利用率在1s内超过60%。</li></ul></li><li>• 弹升结果：计算资源=MIN(触发弹升的资源实际使用量除以45%，预设的弹性上限)。</li></ul> <p><b>说明</b> 触发弹升的资源实际使用量=MAX(CPU使用率*CCU, 内存使用率*CCU)。</p>	<ul style="list-style-type: none"><li>• 触发条件：(以下条件需均满足)<ul style="list-style-type: none"><li>◦ CPU使用率连续60s小于30%。</li><li>◦ 内存利用率连续60s小于30%。</li><li>◦ 当前集群的资源量高于预设的弹性下限。</li></ul></li><li>• 弹降方式：以1 CCU的步长逐步弹降，不是一次弹降至目标值。</li><li>• 弹降结果：集群资源量保持在MAX(使CPU使用率稳定在30%的CCU量，使内存利用率稳定在30%的CCU量，预设的弹性下限)。</li></ul>
<p>策略类型：激进型。 适用场景：</p> <ul style="list-style-type: none"><li>• 业务低谷时间资源使用率仍处于30%以上的场景。</li><li>• 对降低成本有强烈需求的场景。</li></ul>	<p>支持<a href="#">通过工单白名单</a>开启。</p> <p><b>重要</b> 开启激进型策略后，可能会对集群的性能和稳定性产生影响，谨慎开启。</p> <ul style="list-style-type: none"><li>• 查询性能影响：集群会更频繁地触发计算资源的弹降和节点缓存的加锁释放，影响缓存命中率。</li><li>• 稳定性影响：弹升阈值提高和弹升步长降低，可能导致弹升耗时的增加和MemoryLimitException异常的增多。</li></ul>	<ul style="list-style-type: none"><li>• 触发条件：(以下条件满足一个即可)<ul style="list-style-type: none"><li>◦ CPU使用率在5s内的平均值超过90%。</li><li>◦ 内存利用率1s超过80%。</li></ul></li><li>• 弹升结果：计算资源=MIN(触发弹升的资源实际使用量除以45%，预设的弹性上限)。</li></ul> <p><b>说明</b> 触发弹升的资源实际使用量=MAX(CPU使用率*CCU, 内存使用率*CCU)。</p>	<ul style="list-style-type: none"><li>• 触发条件：(以下条件需均满足)<ul style="list-style-type: none"><li>◦ CPU使用率连续15s小于80%。</li><li>◦ 内存利用率连续15s小于75%。</li></ul></li><li>• 弹降方式：以1 CCU的步长逐步弹降，不是一次弹降至目标值。</li><li>• 弹降结果：集群资源量保持在MAX(使CPU使用率稳定在80%的CCU量，使内存利用率稳定在75%的CCU量，预设的弹性下限)。</li></ul>

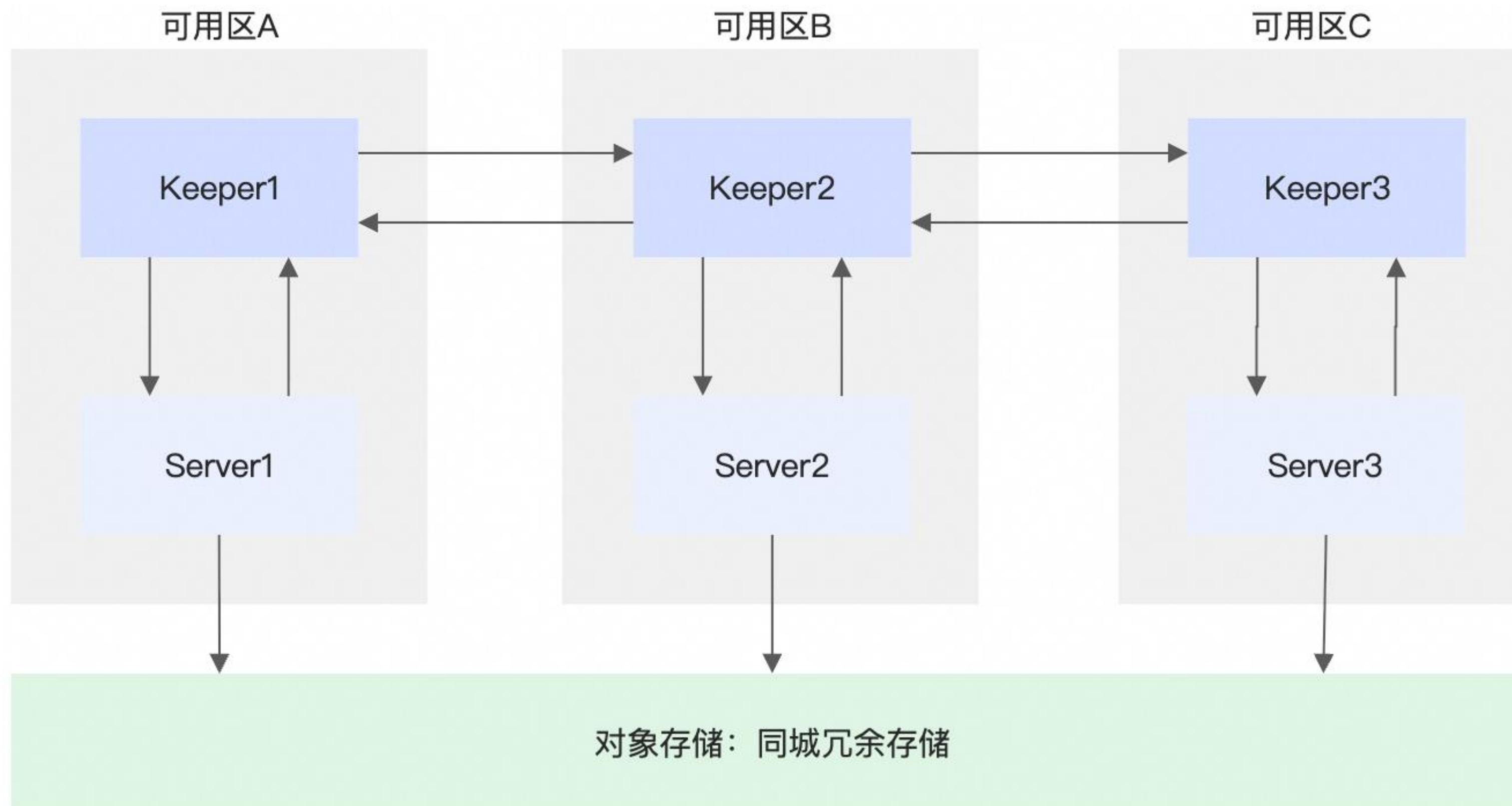
# ClickHouse 企业版：容灾能力提升，SLA>99.9%

## 容灾能力提升：

- 集群默认至少双节点高可用, SLA > 99.9%
- 单节点故障集群RTO=0, RPO=0, 不影响集群整体对外提供读写服务

## 多可用区部署支持：

- SLA > 99.95%
- server和keeper分别部署在3个可用区, 可用区级故障也可保证数据不丢、业务不停



# ClickHouse 企业版：更好的性能&更细粒度的资源管理

阿里云智能集团  
ALIBABA CLOUD INTELLIGENCE GROUP

## 计算组 (10月底上线)

### 资源隔离：

- 在共享存储的基础上提供上层计算资源的隔离，支持客户自定义进行读写、租户级别的隔离
- 支持计算组配置和切换RW、RO能力

### 独立弹性，更精细化的成本管理：

- 支持计算组独立定义弹性范围

## 分布式缓存 (12月底上线)

- 计算-缓存-存储三层解耦
- 所有节点都可以访问同一个分布式缓存中的数据，更进一步提高缓存命中率和集群的整体性能

clickhouse keeper  
(free by default)

clickhouse keeper  
(free by default)

clickhouse keeper  
(free by default)

compute group1 (by default)

clickhouse  
server

clickhouse  
server

compute group2

clickhouse  
server

clickhouse  
server

distributed cache (by default memory:cache=1:6)

shard storage

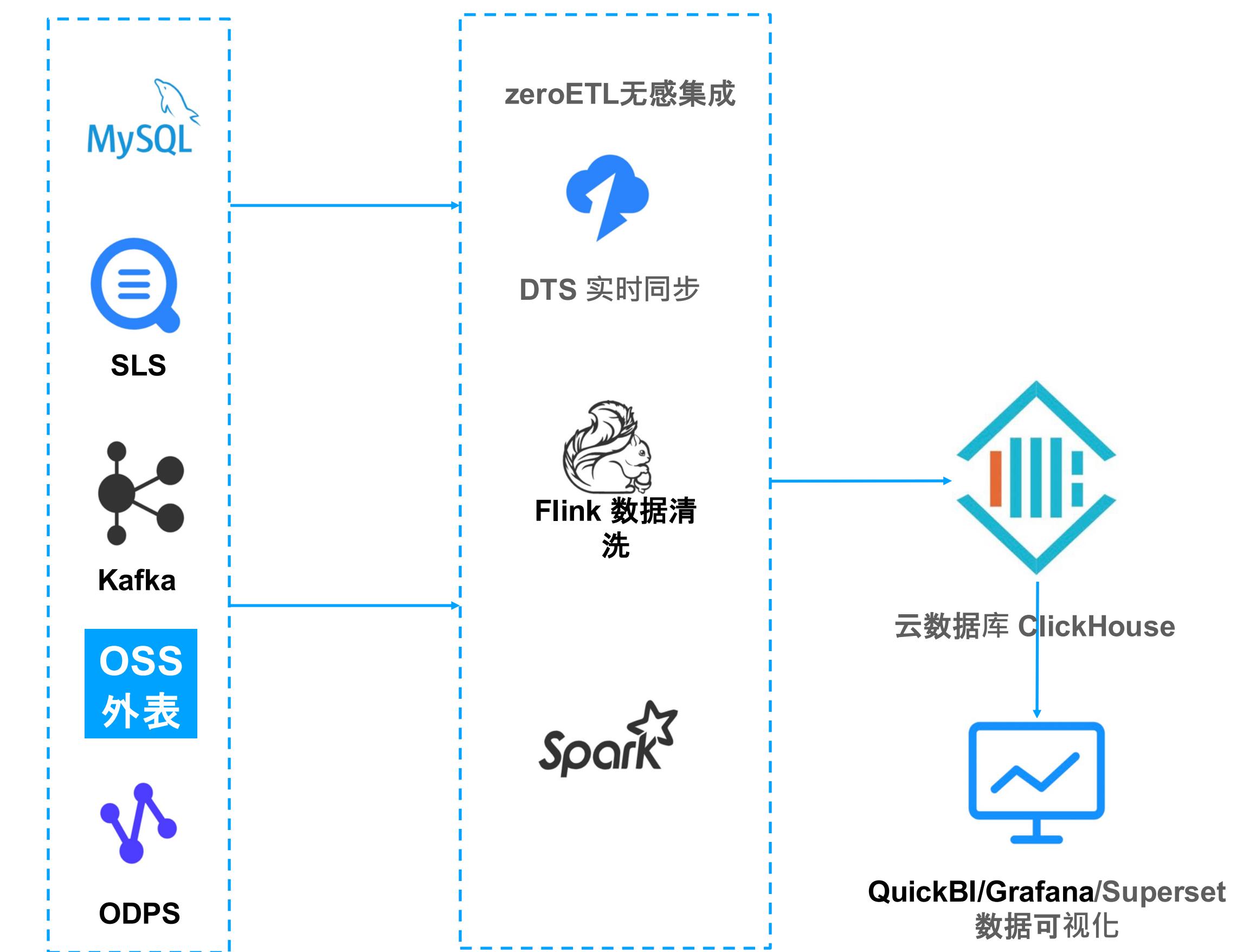
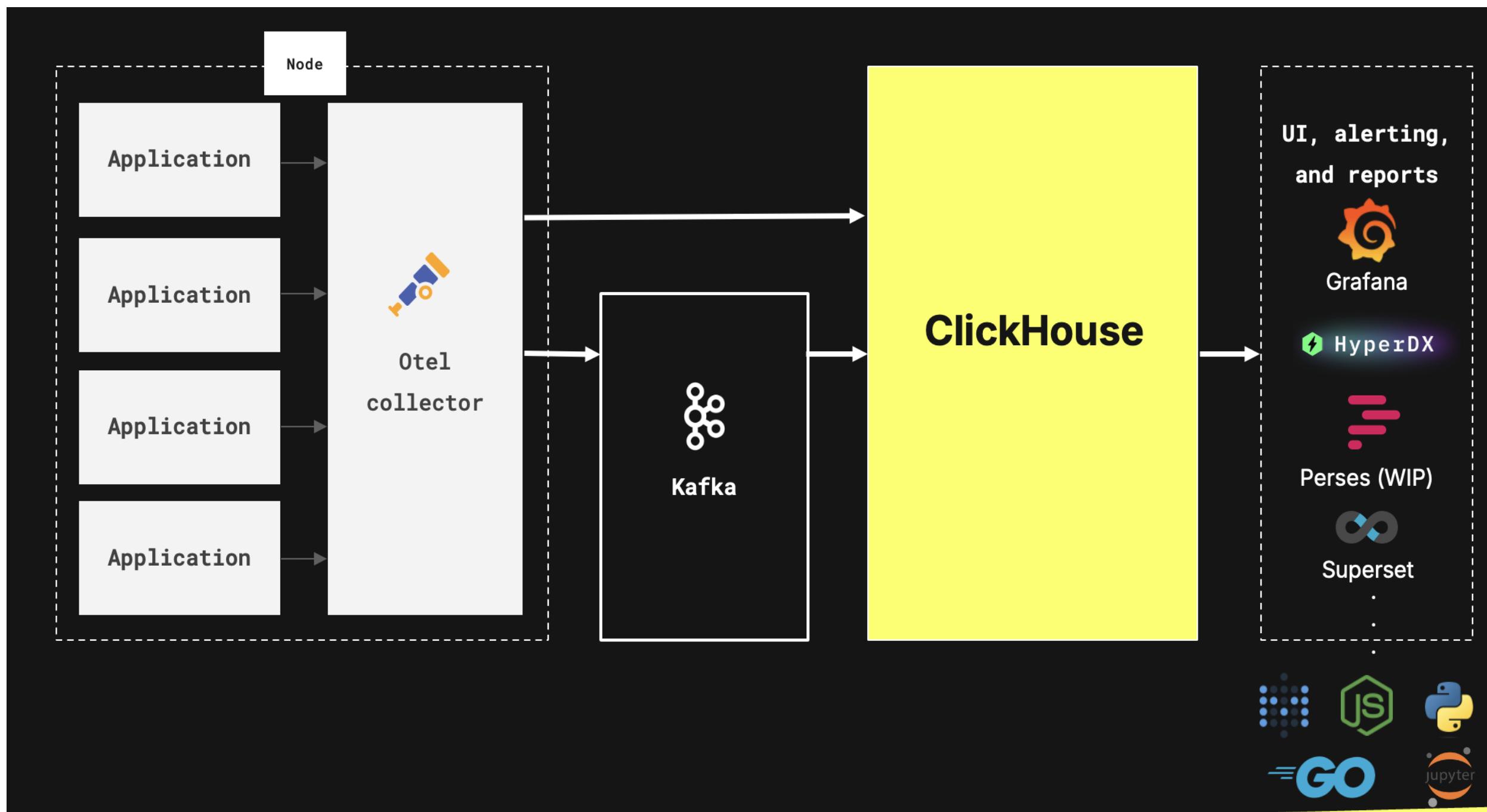
# 云数据库ClickHouse企业版：更完善的链路支持

## 开源能力100%兼容

- 多种外表引擎支持：MySQL、PostgreSQL、Kafka、Hive、OSS/S3、HDFS、Iceberg等
- Flink connector支持
- Otel采集端适配：Otel collector部署后自动采集数据到ClickHouse标准表
- Grafana集成：Grafana支持clickhouse插件，配置后可以在Grafana端使用Query Builder 或 SQL Editor构建查询

## 阿里云生态适配：

- 基于DTS的更稳定的MySQL到ClickHouse数据同步链路
- 基于DTS的从SLS到ClickHouse的数据同步链路



# ClickHouse 企业版：一站式可观测能力集成（ClickObserve）

阿里云智能集团  
ALIBABA CLOUD INTELLIGENCE GROUP

- ClickHouse控制台实例详情内嵌入口，支持一键部署
- 涵盖数据采集、数据存储&查询、数据可视化这三个可观测平台的核心能力
- 兼容Lucene语法（kibana用户可以无障碍迁移，并继续使用自己熟悉的语法）
- 相对ES，性能提升5~10倍，数据存储成本降低30%~60%
- 完善的从ES到CK的数据、[数据采集agent](#)、[sdk](#)迁移方案和详细指导文档

The screenshot shows the Alibaba Cloud ClickHouse observability integration interface. The top navigation bar includes links for Workstation, Region (North China 2 (Beijing)), Search, and various system status indicators. The main page displays a cluster named '阿里云ClickHouse可观测' which is currently '运行中' (Running). On the left, a sidebar lists management categories: Cluster Information, Monitoring & Alerts, Account Management, Database Management, Data Security, Parameter Management, Query Management, Zero-ETL Integration, Backup Recovery, Data Visualization, Data Import, and One-stop Observability. The central content area is divided into sections: Cluster Properties (Cluster ID: clickhouse, Engine: clickhouse, Label: default resource group), Network Information (External IP: 10.128.1.123, VpcId: vpc123, Port: 9004), Cluster Status (Status: Running, Billing Type: Pay-as-you-go), and Configuration Information (Small Version: 24.12.1.18592\_6, Compute Elasticity Range (CCU): 8-16). A blue button labeled '登录数据库' (Login to Database) is located above the '集群状态' section. A sidebar on the right provides AI assistance.

## 02 / 各行业客户在阿里云上的可观测实践

场景介绍：得物全链路日志采集监控&可观测平台，基于PB级Trace数据构建的内部实时日志分析平台

自建开源ClickHouse痛点：成本高、容灾能力差、运维难

- **成本高：**PB级数据的存储带来高额成本压力。
- **容灾能力差：**单副本集群容灾能力差，单节点故障影响整个集群可用性。
- **运维难：**每年六一八双十一都会有**2-3倍的业务读写增长**，存算一体架构下集群扩容需要停写，**业务影响大**；分布式集群需要**业务直连节点进行数据和写入压力的平衡**，运维工作量大。

业务读写优化：企业版内核特性保障业务读写性能

写入优化：

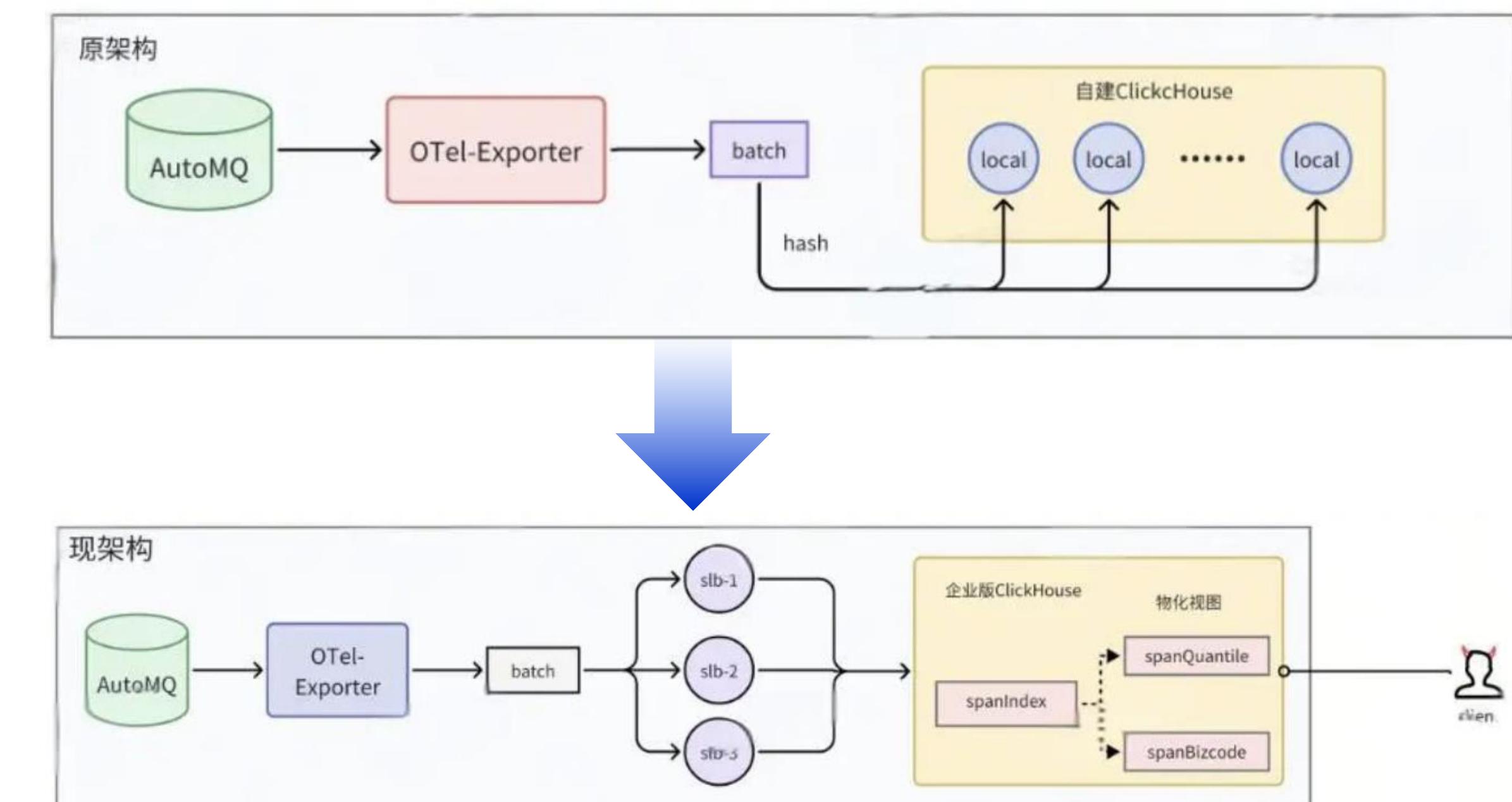
- **稳定性提升：**借助负载均衡（LB），将写入请求均匀分配到多个计算节点，避免单节点过载，提高系统稳定性。
- **性能提升：**利用ClickHouse企业版的Serverless架构，成功支持了分布式链路追踪场景下集群每秒高达**2000万行**的写入操作，单次请求40万行数据写入耗时优化至**1s左右**。

查询优化：

- **慢sql性能优化：**利用Parallel Replica将查询分发至多个节点并行处理，特定场景下，**查询速度提升可达2.5倍**。整体查询效率与自建ClickHouse不相上下。

企业版收益：成本降低60%+、容灾能力提升、运维简化

- **成本降低：**基于计算serverless和oss存储，实现**成本60%+**
- **容灾能力提升：**单节点故障不影响集群正常对外提供服务
- **运维简化：**
  - 双十一业务高峰扩容分钟级完成，无需停读停写
  - 无本地表和分布式表的概念，可直连lb进行读写，**无需业务管理节点间的负载均衡和数据分布**
- **架构切换便捷：**除写入方式从直连节点修改为直连lb外，**无需业务改造**，**企业版可100%兼容社区版语法**



# 互联网行业：道旅行行为日志分析，相比传统数仓成本降低70%，性能提升15倍

## 客户及场景介绍



**客户介绍：**道旅科技是一家全球旅游产品分销服务商，为全球客户提供迅捷且高度灵活的优质旅游产品连接解决方案，涵盖酒店、机票以及用车、门票等碎片化产品领域。

**场景介绍：**道旅核心业务上报的请求埋点数据，经过Kafka、Flink到数据仓库进行业务的实时分析，支持数据分析/算法团队做业务上的运营策略调整；而技术侧的Trace数据，用于给技术侧做开发和问题定位。

### 业务特性：

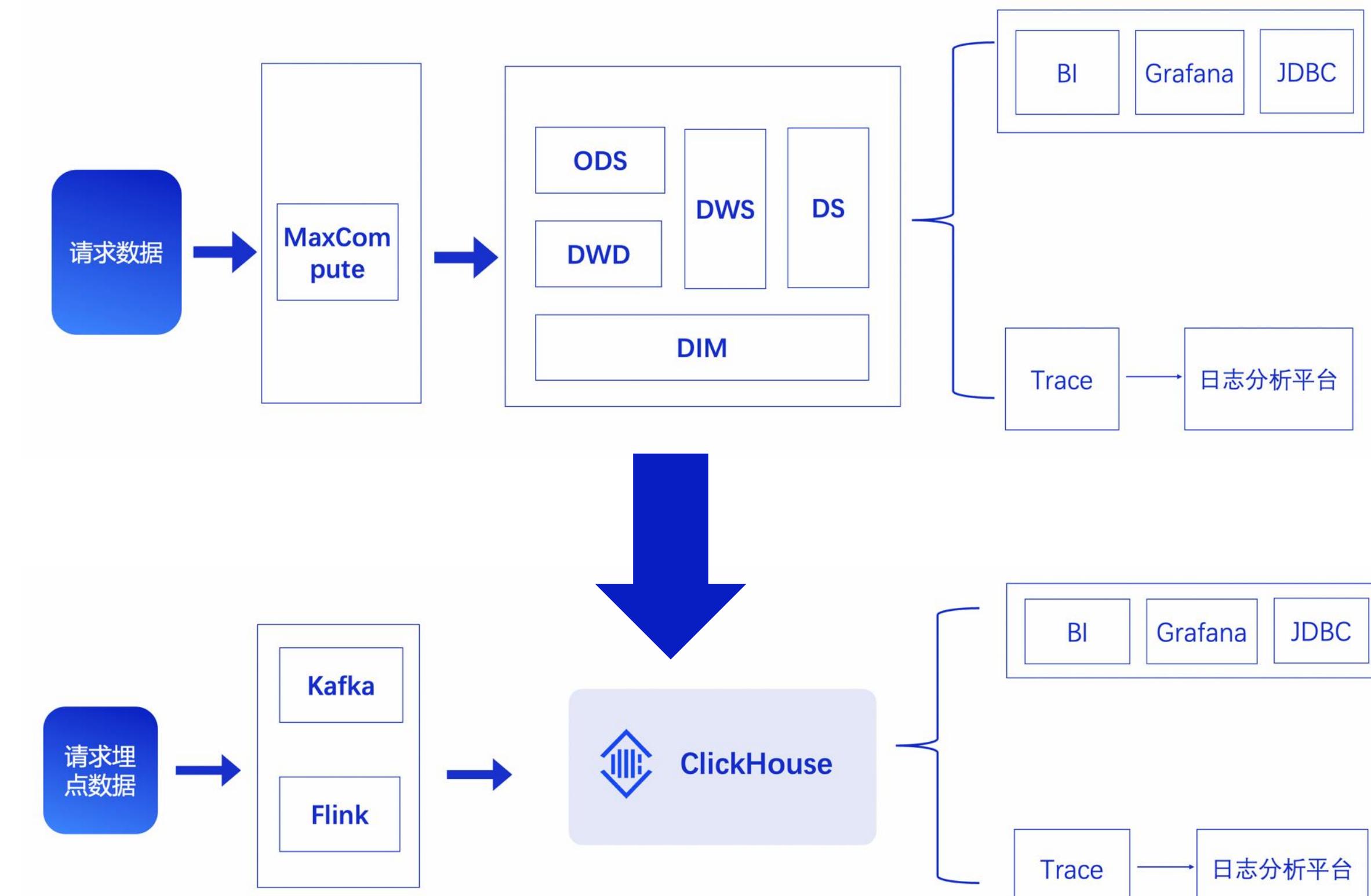
- 写入量大：**作为每天有海量B端访问的全球旅游产品分销服务商，每天有几百TB的埋点数据产生，需要底层的数据库具备较高批量数据写入性能。
- 批量查询性能要求高：**为了更好的支持业务团队进行实时数据分析，做出运营决策，也为了提高算法团队的研发效能，要求日志分析平台能够提供秒/亚秒级的关键指标展示和复杂查询支持。
- 存储成本低：**作为存储了所有业务中心上报的用户埋点行为数据和技术侧Trace数据的大数据平台，累计数据数千万条，规模高达几百TB，要求底层的数据库产品能具备较高的存储压缩率和较低的存储成本。

### 传统数仓使用痛点：性能差、稳定性差、运维成本高

- 性能差：**层层加工导致数据从源头到应用延时较高
- 响应速度慢：**需求变化时，需要多层联动修改，相应速度慢
- 运维成本高：**层级多、表多、需要投入较多人力进行开发和运维

## 升级到ClickHouse：成本降低70%，性能提升15倍

- 查询速度提升15倍：**使用clickhouse内置的物化视图能力和Projection，大幅提升数据分析处理，查询性能提升~15倍
- 存储成本降低80%：**借助ClickHouse的列式存储和超高数据压缩率，提升存储压缩率5倍+
- 稳定性提升：**可以轻松支持每秒百万行数据的写入
- 运维压力降低：**利用物化视图和Projection，降低了业务侧的开发量50%+



# 汽车行业：智驾车控链路追踪，替换自建ElasticSearch，成本降低50%，查询稳定10s内

## 车控链路追踪场景介绍&业务特性

### 场景介绍：

汽车车控链路数据的存储和分析，用于对智能汽车电子架构的车辆控制指令（包括远程开关车门、开关窗、智能驾驶等行为）的全生命周期追踪监控，涵盖从信号输入→决策计算→执行输出的完整闭环。用于新车功能的开发或验收时的车控链路日志查看、以及车控链路异常的排障。

### 业务特性：

- 数据量大：**由于车控链路日志涉及到手机端-云端计算-车控系统的三端交互，链路复杂，非峰值数据日增量100TB，至少保留7天，部分业务需保留1月
- 写入吞吐高：**日常写入流量500w行/s，峰值700-800w行/s
- 查询实时性要求高：**链路追踪和排障分析要求查询秒级返回
- 查询模式固定：**以TopN查询、Trace\_id查询、聚合分析和核心指标计算为主，90%的查询只查当天数据，部分查询跨周末或跨周
- 容灾诉求：**集团要求每年停产运维时间不超2小时
- 开源兼容：**运维中心遵循多云部署策略，在idc、阿里云、百度云均有部署，要求技术栈100%兼容开源



## ElasticSearch痛点：成本高、性能差、稳定性差

- 成本高：**ElasticSearch压缩率4.5，上百TB数据的存储成本高昂
- 性能差：**计算资源已经扩到1600core，但仍存在部分大查询失败无法返回，影响新功能调试和故障排查的效率
- 稳定性差：**已知的系统bug无法修复，影响线上业务稳定性

## ClickHouse企业版：成本降低50%、查询秒级返回、容灾增强

- 成本降低：**
  - 存储成本：**ClickHouse企业版存储压缩率10，是ES的2.5倍
  - 计算成本：**企业版768 核集群即可达到客户性能要求，相对ES集群计算资源降低50%
- 性能提升：**topk日内查询<=5S，部分长周期查询(跨7天)P99在10s以内，排障效率提升
- 容灾能力增强：**集群变配无需停产，单节点故障不影响业务
- 开源兼容：**100%兼容开源ClickHouse，降低多云部署政策下的运维复杂度

## 车控链路追踪场景下的企业版最佳实践：物化视图+并行查询

### 车控链路追踪场景的企业版最佳实践：

- 写入性能：**无需调优可支持每秒千万行数据级别的写入
- 查询性能：**
  - 针对按trace\_id查询场景，创建物化实图，将trace\_id作为主键进行过滤
  - 开启parallel replicas，查询性能随着节点数可以近线性提升
  - 开启异步load mark cache

# THANKS