DTCC 2010-2022

第十三届中国数据库技术大会

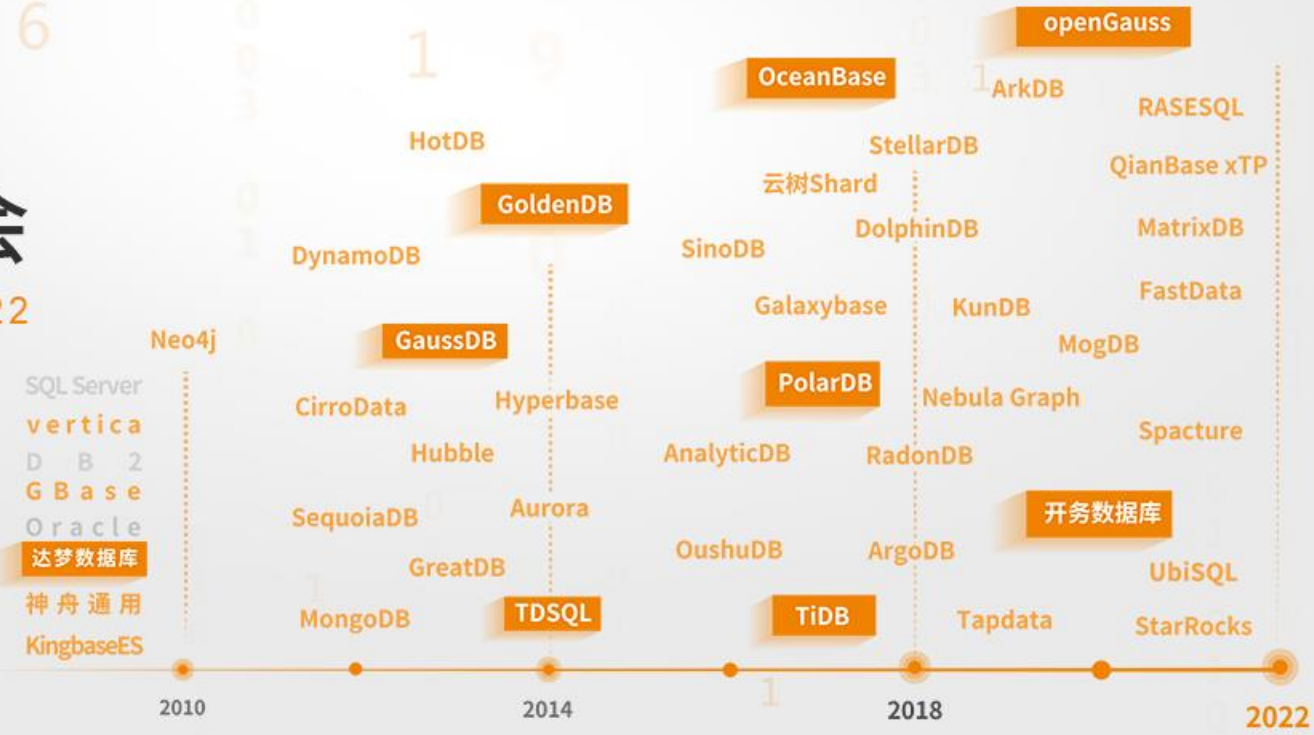
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

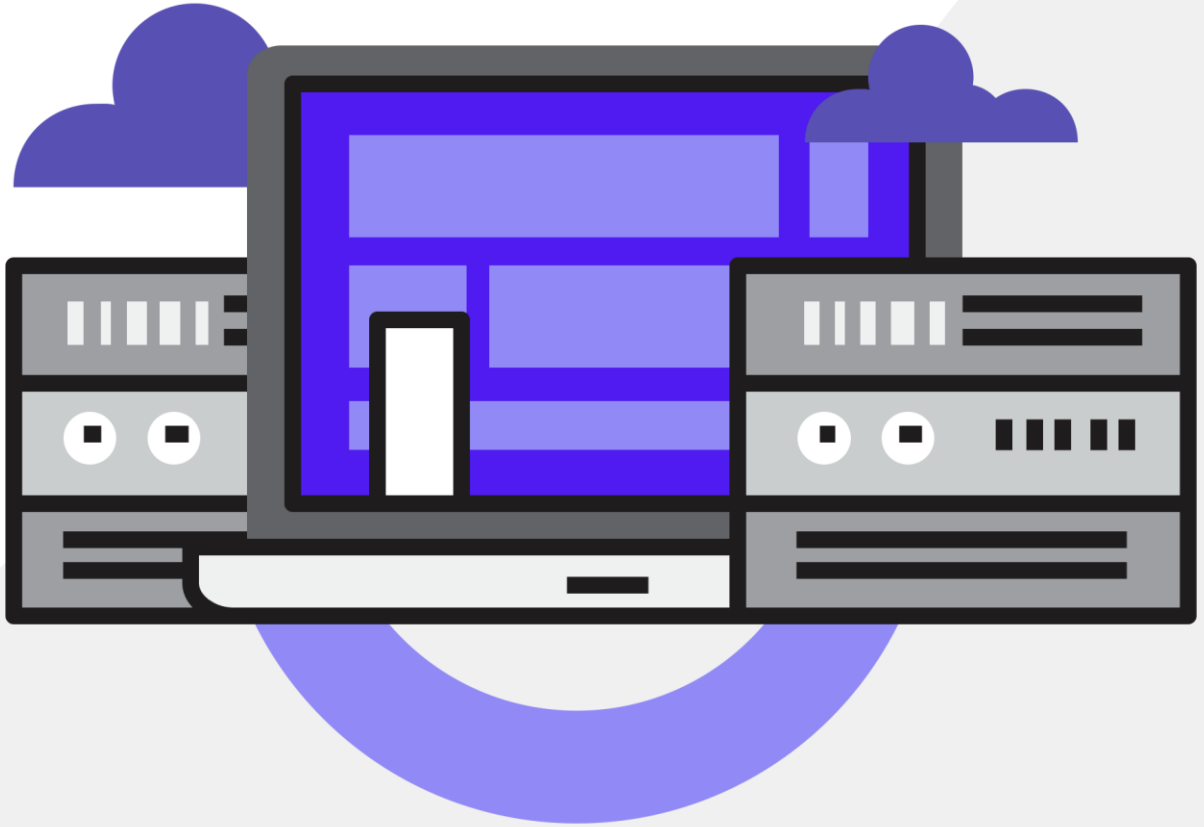
数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间





弹性经济
云上ClickHouse的架构思路

沃趣科技 产品架构师 & 联合创始人

罗春

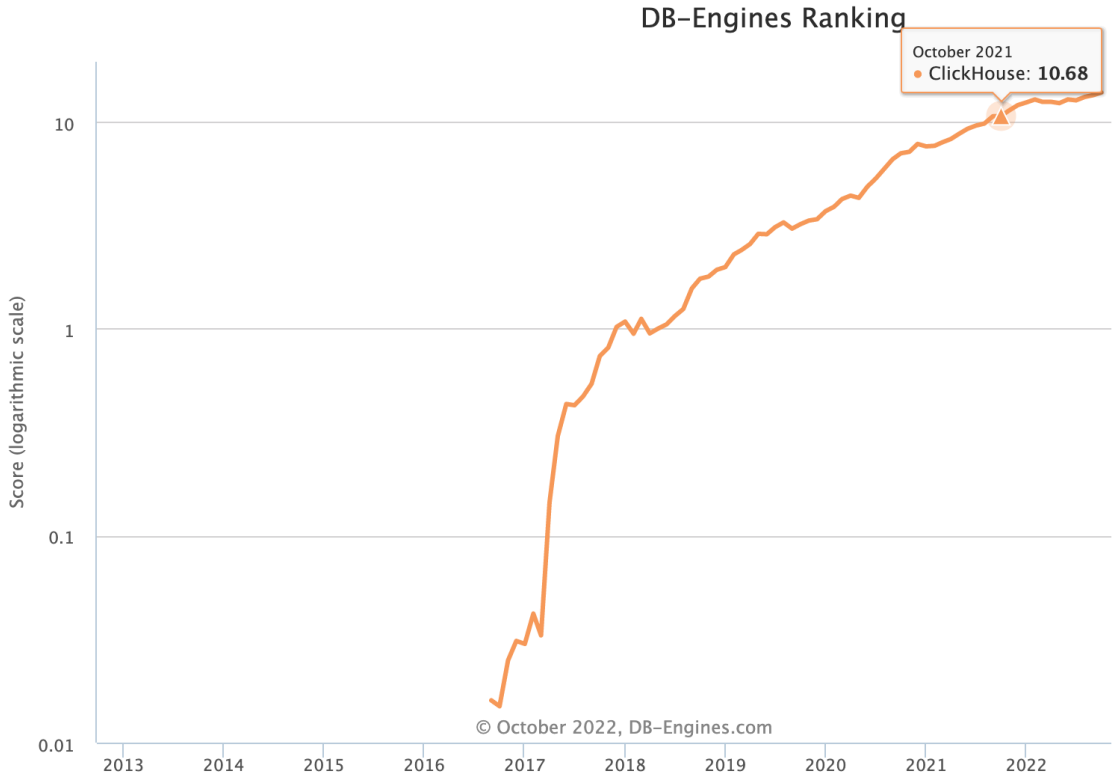
关于我
About Me

罗春 @沃趣科技

- 在阿里巴巴上班，做一名DBA
- 在沃趣科技创业，做一名杂役
- 在沃趣科技再创业，做一名大杂役

大数据MPP界的黑马 ClickHouse

ClickHouse趋势图



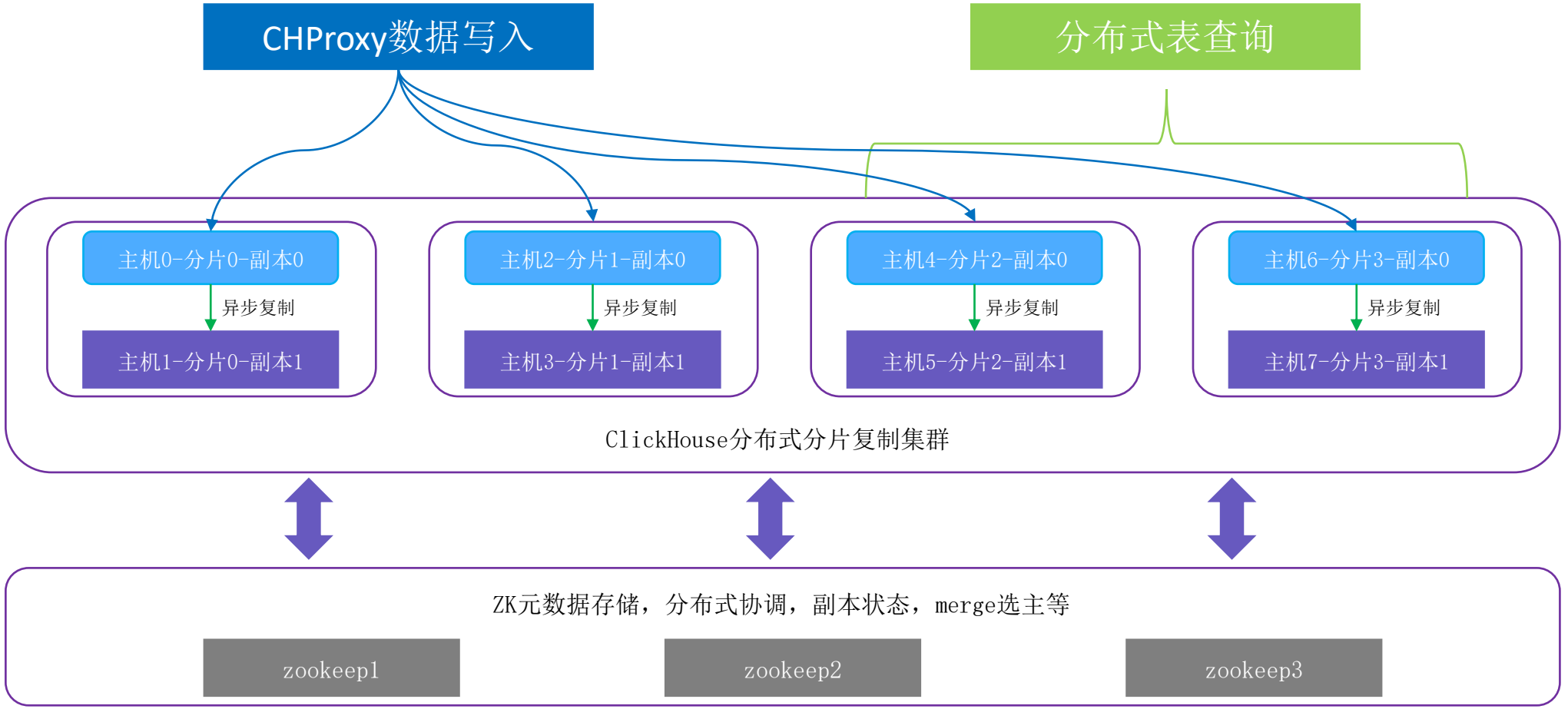
ClickBench跑分表

	ClickHouse (tuned) (c6a.metal, 500gb gp2)	ClickHouse (c6a.metal, 500gb gp2)	SelectDB (c6a.metal, 500gb gp2)	ClickHouse (m5d.24xlarge)	StarRocks (c6a.metal, 500gb gp2)	Redshift (4×ra3.16xlarge)
Load time:	**137s (×1.00)**	138s (×1.01)	369s (×2.70)	208s (×1.52)	484s (×3.54)	1829s (×13.36)
Data size:	13.57 GiB (×1.46)	13.57 GiB (×1.46)	15.95 GiB (×1.71)	**9.31 GiB (×1.00)**	16.53 GiB (×1.78)	24.52 GiB (×2.63)
Q0.	0.00s (×1.08)	0.00s (×1.08)	0.00s (×0.98)	0.00s (×1.08)	0.04s (×4.92)	0.02s (×3.28)
Q1.	0.04s (×2.25)	0.01s (×1.15)	0.02s (×1.50)	**0.01s (×1.00)**	0.03s (×2.00)	0.03s (×1.88)
Q2.	0.04s (×2.45)	0.02s (×1.40)	0.04s (×2.50)	0.01s (×1.25)	0.05s (×3.00)	0.04s (×2.31)
Q3.	0.03s (×1.75)	0.02s (×1.55)	0.06s (×3.50)	0.02s (×1.40)	0.05s (×3.00)	0.06s (×3.44)
Q4.	**0.09s (×1.00)**	1.77s (×17.99)	0.15s (×1.62)	1.88s (×19.07)	0.25s (×2.63)	0.11s (×1.20)
Q5.	**0.12s (×1.00)**	0.78s (×6.05)	0.28s (×2.23)	0.85s (×6.64)	0.29s (×2.31)	0.18s (×1.43)
Q6.	0.02s (×2.15)	0.01s (×1.92)	0.01s (×1.54)	0.02s (×2.08)	0.05s (×4.62)	0.03s (×2.98)
Q7.	0.02s (×1.65)	0.02s (×1.70)	0.02s (×1.50)	0.02s (×1.55)	0.03s (×2.00)	0.03s (×2.19)
Q8.	0.14s (×1.03)	0.26s (×1.87)	0.33s (×2.38)	0.36s (×2.57)	0.31s (×2.24)	**0.13s (×1.00)**
Q9.	0.27s (×1.01)	**0.27s (×1.00)**	0.34s (×1.27)	0.39s (×1.45)	0.35s (×1.30)	1.41s (×5.16)
Q10.	**0.06s (×1.00)**	0.12s (×1.80)	0.11s (×1.71)	0.13s (×1.99)	0.13s (×2.00)	0.09s (×1.41)
Q11.	**0.06s (×1.00)**	0.09s (×1.47)	0.10s (×1.67)	0.13s (×2.05)	0.16s (×2.58)	0.10s (×1.71)
Q12.	**0.14s (×1.00)**	0.14s (×1.04)	0.23s (×1.63)	0.20s (×1.41)	0.16s (×1.16)	0.16s (×1.15)
Q13.	**0.18s (×1.00)**	0.18s (×1.05)	0.75s (×4.09)	0.27s (×1.53)	0.32s (×1.77)	0.31s (×1.70)
Q14.	**0.15s (×1.00)**	0.16s (×1.03)	0.30s (×1.90)	0.22s (×1.42)	0.26s (×1.66)	0.18s (×1.18)
Q15.	0.12s (×1.03)	**0.12s (×1.00)**	0.16s (×1.33)	0.19s (×1.56)	0.26s (×2.11)	0.14s (×1.20)
Q16.	0.32s (×1.02)	**0.32s (×1.00)**	0.50s (×1.56)	0.57s (×1.78)	0.45s (×1.41)	0.34s (×1.07)
Q17.	**0.10s (×1.00)**	0.24s (×2.28)	0.20s (×1.94)	0.37s (×3.48)	0.35s (×3.33)	0.39s (×3.66)
Q18.	0.79s (×1.54)	0.74s (×1.46)	1.06s (×2.07)	1.26s (×2.46)	0.71s (×1.40)	0.66s (×1.31)
Q19.	0.01s (×2.17)	0.01s (×2.17)	0.00s (×0.98)	0.03s (×3.94)	0.01s (×1.97)	0.03s (×3.89)
Q20.	0.26s (×3.35)	0.14s (×1.81)	0.60s (×7.63)	0.26s (×3.38)	0.22s (×2.88)	0.34s (×4.32)
Q21.	0.25s (×2.01)	0.13s (×1.08)	0.34s (×2.69)	0.30s (×2.41)	**0.12s (×1.00)**	0.36s (×2.88)
Q22.	0.45s (×1.47)	0.32s (×1.07)	0.37s (×1.23)	0.68s (×2.21)	**0.30s (×1.00)**	1.03s (×3.36)

ClickBench — a Benchmark For Analytical DBMS

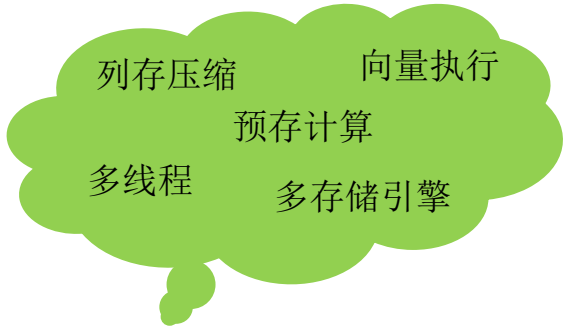
https://benchmark.clickhouse.com

标准ClickHouse架构



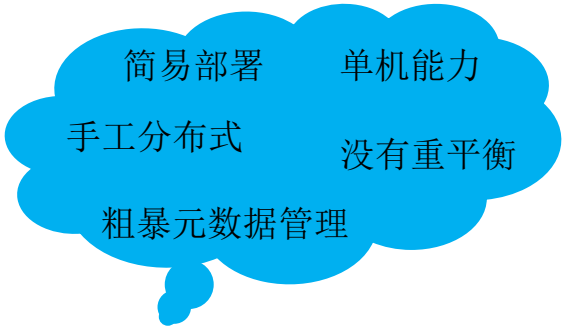
ClickHouse，一个极致性能，“简单的”SQL引擎

单机性能爆表



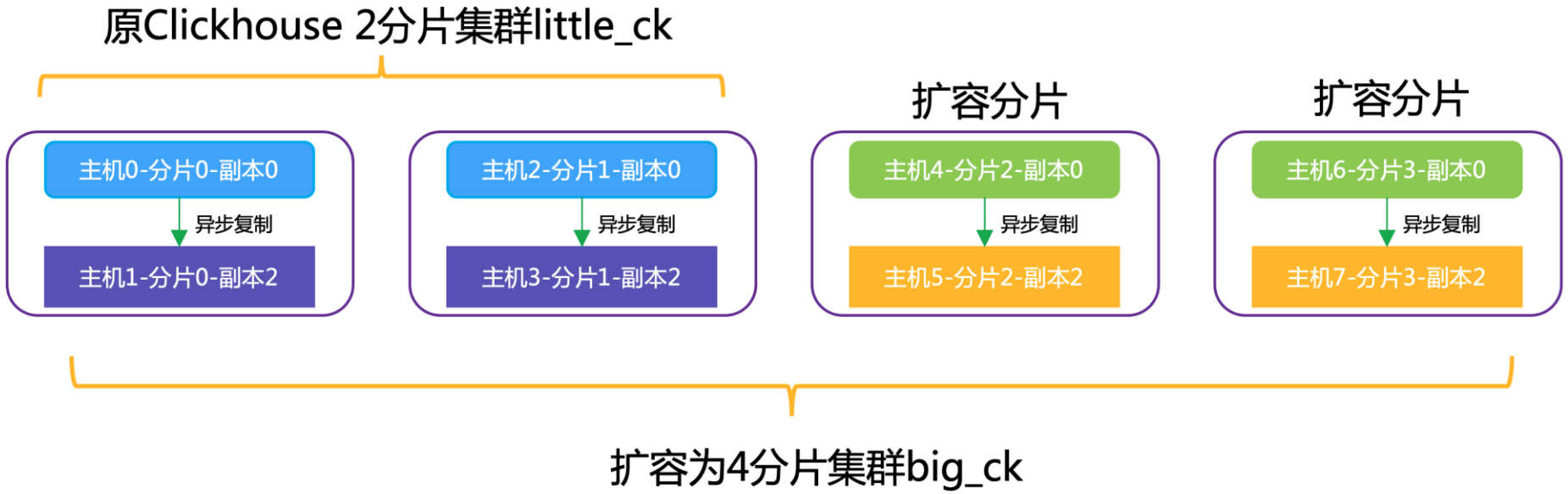
SQL亲民

简单，复杂



ClickHouse的一些槽点

- 扩容无法重平衡
- 无集中元数据管理
- 数据入口尴尬
- Join不友好
- 集群部署复杂



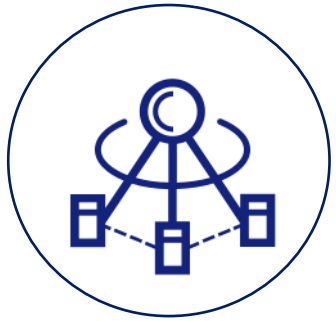
借力云上的弹性资源

云主机

弹性计算，按需申请
丰富的vCPU与内存组合
按量/包月/竞价实例

块设备

高性能云盘，块设备存储
IOPS/吞吐量可量化
自带容错，按需扩容



负载均衡

流量分发，跨区域故障保护
包月/流量多模式灵活计费

S3对象存储

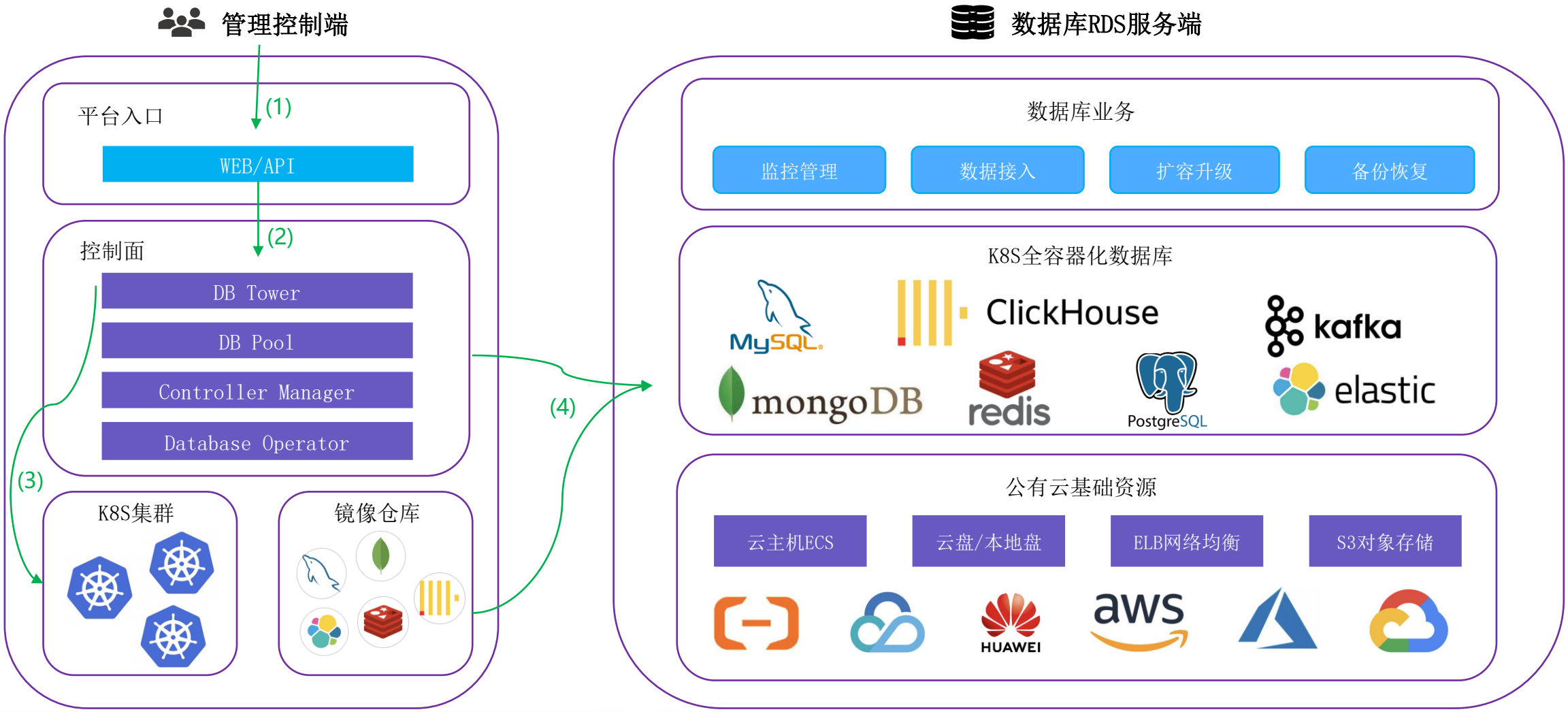
数据存储可靠性9个9，
全网灵活访问
高吞吐量，成本低廉

> 云原生技术有利于各组织在公有云、私有云和混合云等新型动态环境中，构建和运行可弹性扩展的应用。云原生的代表技术包括容器、服务网格、微服务、不可变基础设施和声明式 API。这些技术能够构建容错性好、易于管理和便于观察的松耦合系统。

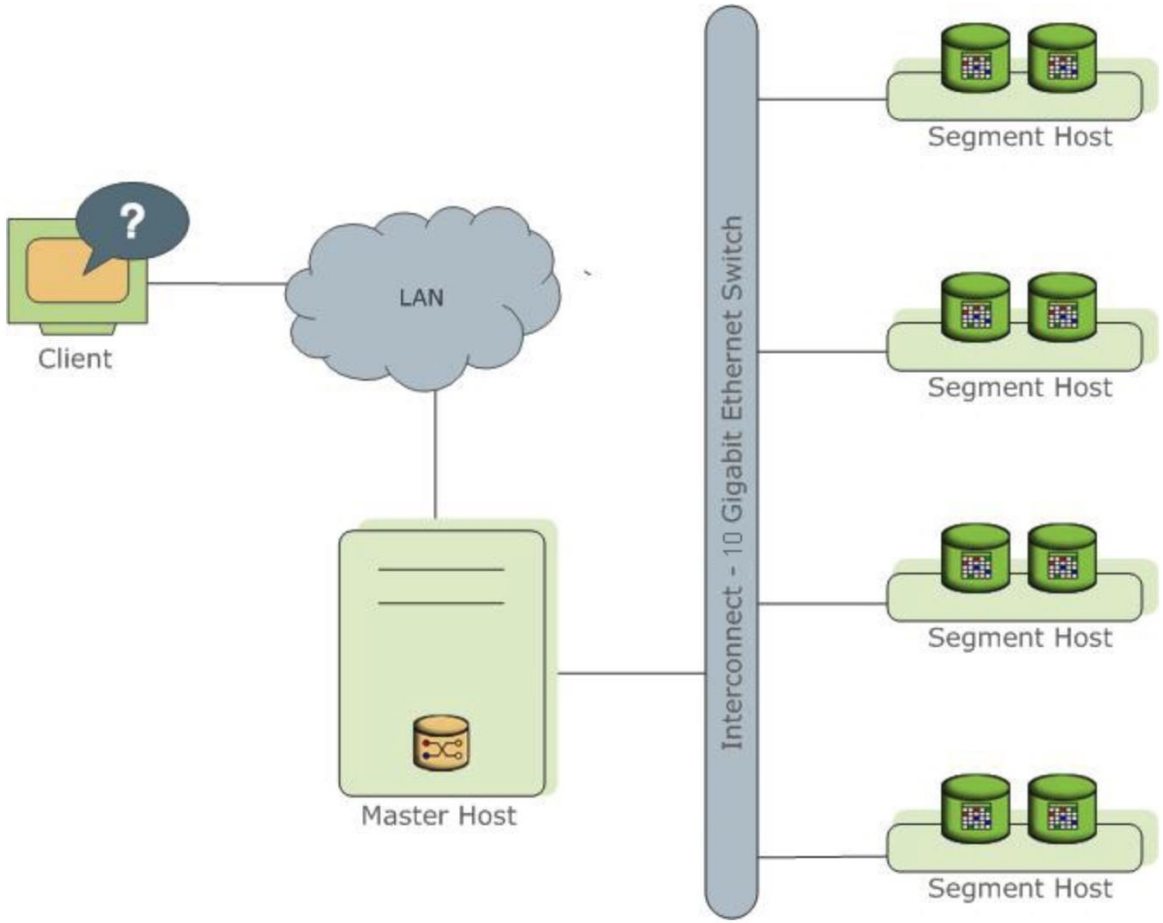
K8s容器化+弹性云资源构建云原生ClickHouse


```
# 云主机定义
module "tf-instances" {
 source                  = "alibaba/ecs-instance/alicloud"
 region                  = "cn-beijing"
 number_of_instances     = "3"
 vswitch_id              = alicloud_vswitch.vsw.id
 group_ids               = [alicloud_security_group.id]
 private_ips             = ["172.16.0.10"]
 image_ids               = ["ubuntu_18_20G_alibase.vhd"]
 instance_type           = "ecs.n2.small"
 ...
 password                = "User@123"
 system_disk_category    = "cloud_ssd"
 data_disks = [
  {
    disk_category = "cloud_ssd"
    disk_name     = "my_module_disk"
    disk_size     = "50"
  }
 ]
}
```

K8s容器化+弹性云资源构建云原生ClickHouse



我眼中的云原生数据库弹性扩展

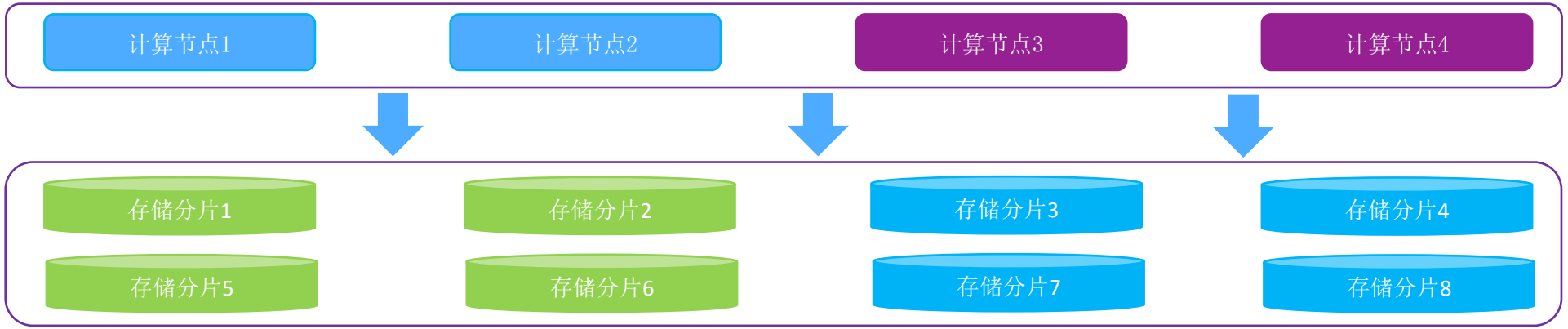


MPP是share-nothing架构，扩展节点容易

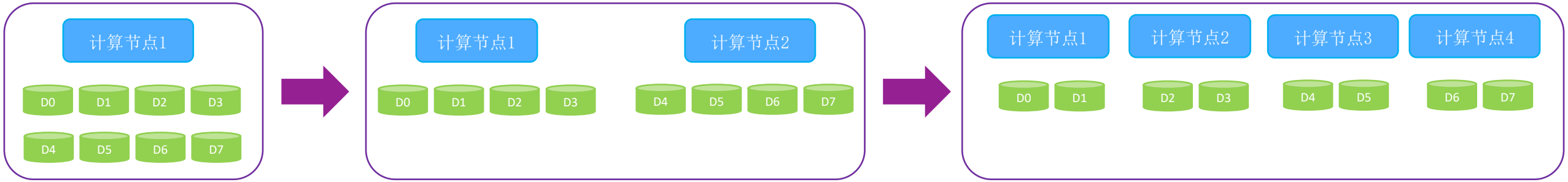
计算和存储紧耦合，数据重分布是弹性扩展最大的障碍

我眼中的云原生数据库弹性扩展

计算资源池

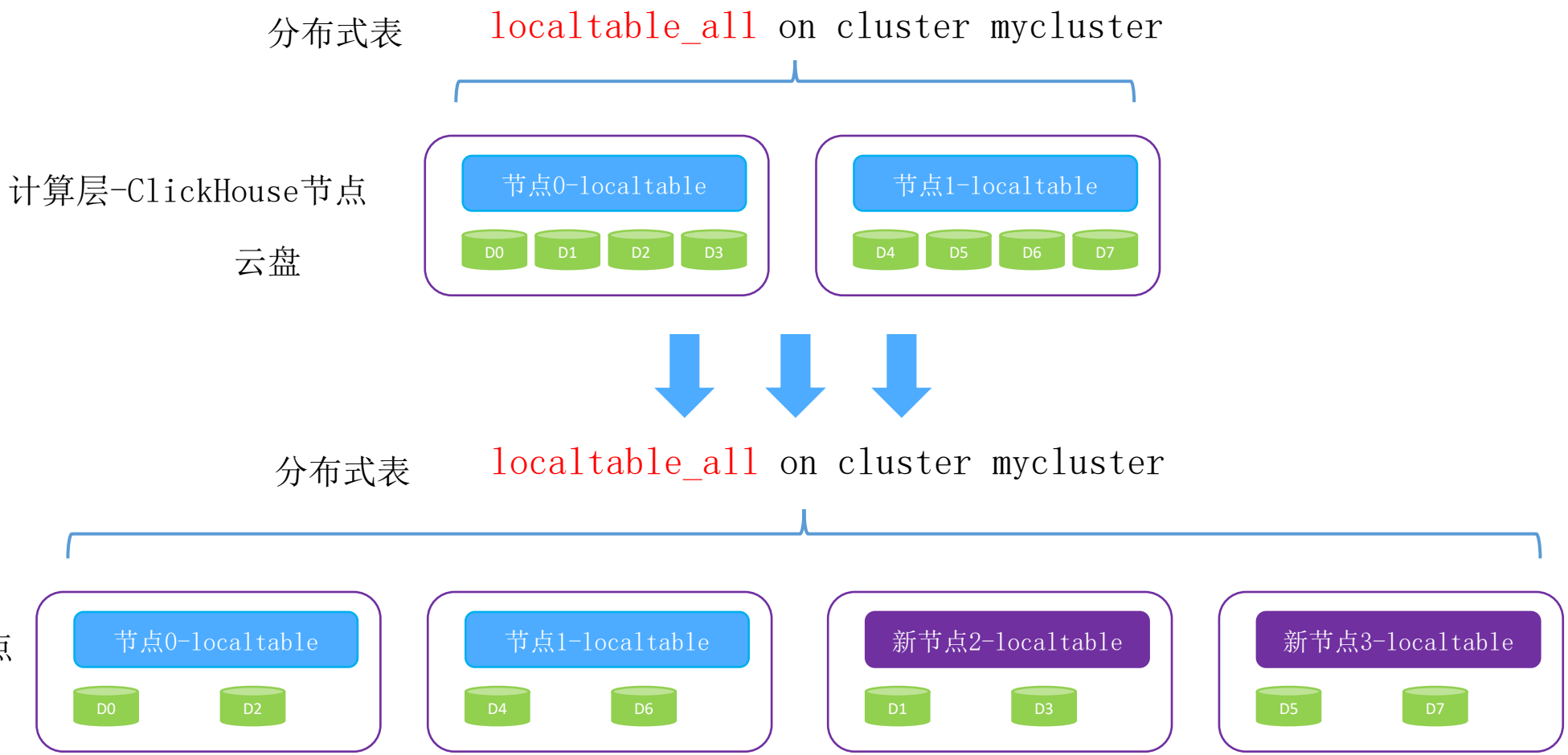


存储资源池

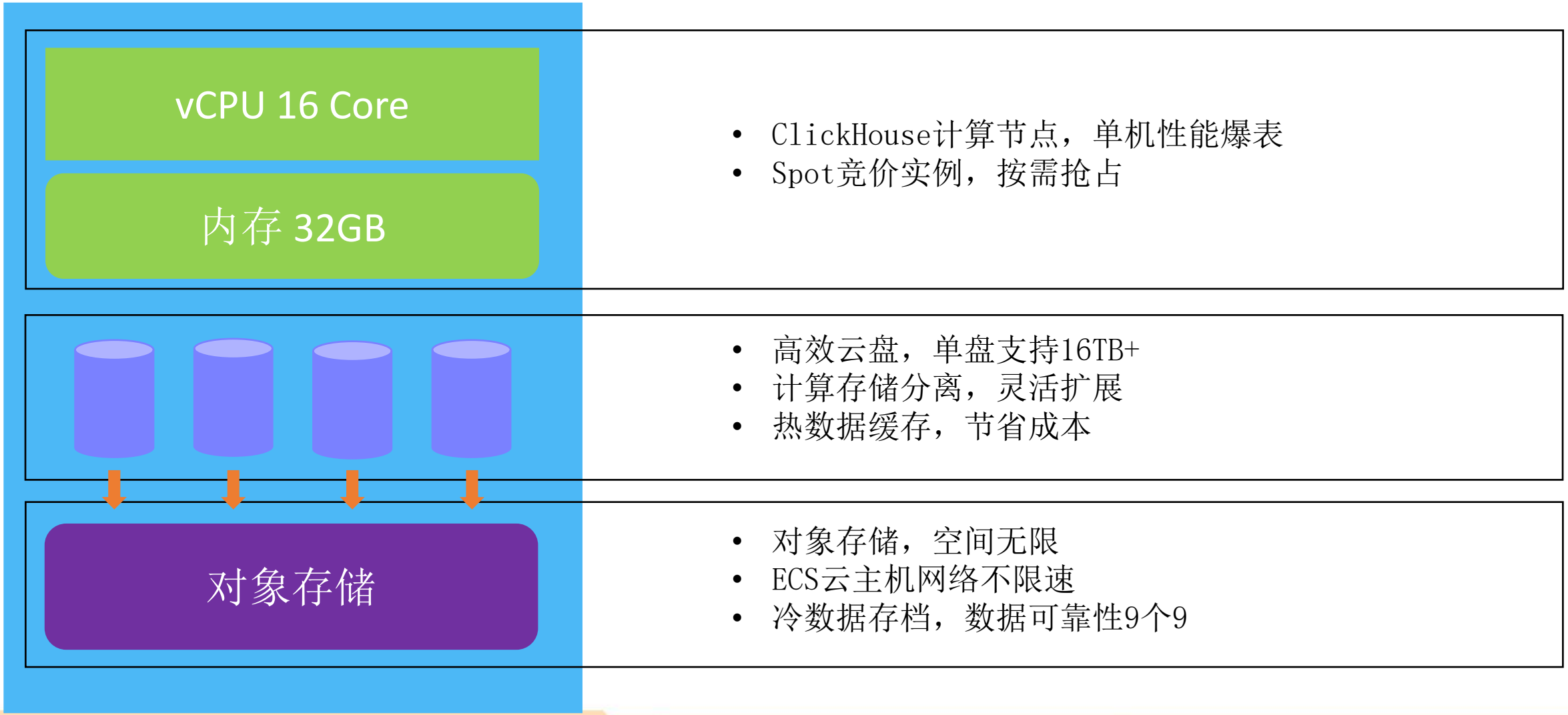


存算分离，计算与存储的关系编排

ClickHouse存算分离-ZeroCopy弹性伸缩



云上的极简ClickHouse使用姿势



- ClickHouse计算节点，单机性能爆表
- Spot竞价实例，按需抢占

- 高效云盘，单盘支持16TB+
- 计算存储分离，灵活扩展
- 热数据缓存，节省成本

- 对象存储，空间无限
- ECS云主机网络不限速
- 冷数据存档，数据可靠性9个9

ClickHouse数据接入-数据库实时增量

ClickHouse的数据写入特点

- 批量插入，害怕零散数据
- 列存压缩，不擅长更新删除

同步方案

- 控制迁移批次数据量，分批插入
- 所有都是插入，引入DML和版本标识
- ReplacingMergeTree单Part自动去重
- 异步任务处理删除数据

```
CREATE TABLE employees
(
    `emp_id` UInt16 COMMENT '员工id',
    `name` String COMMENT '员工姓名',
    `work_place` String COMMENT '工作地点',
    `age` UInt8 COMMENT '员工年龄',
    `depart` String COMMENT '部门',
    `salary` Decimal(9, 2) COMMENT '工资',
    `__version@@` Nullable(Int64),    同主键，版本自增
    `__event_type@@` Nullable(String) DML类型，I，D，U
)
ENGINE = ReplacingMergeTree
ORDER BY (emp_id, name)
SETTINGS index_granularity = 8192
```

Query id: f209d302-6925-49c9-b7df-e3eb25716a95 同步表数据

emp_id	name	work_place	age	depart	salary	__version@@	__event_type@@
1	顾玲	上海	41	丰盛律师部	90000.00	0	I
1	顾玲	上海	41	丰盛律师部	100000.00	1	U
1	顾玲	上海	41	总监俱乐部	150000.00	2	U
2	王多余	上海	32	丰盛律师部	20000.00	0	I
2	王多余	上海	32	丰盛律师部	20000.00	1	D

ClickHouse数据接入-实时查询视图

原始同步数据的困扰

- 业务不想修改SQL逻辑
- 不想等待后台merge

应对方案

- 窗口分析函数
- 自动新增一张视图
- 很土，但管用

```
CREATE OR REPLACE VIEW dbck.__employees
(
    `emp_id` UInt16,
    `name` String,
    `work_place` String,
    `age` UInt8,
    `depart` String,
    `salary` Decimal(9, 2)
) AS
SELECT
    emp_id,
    name,
    work_place,
    age,
    depart,
    salary
FROM
(
    SELECT
        emp_id,
        name,
        work_place,
        age,
        depart,
        salary,
        `__event_type@@`,
        rank() OVER (PARTITION BY emp_id, name ORDER BY `__version@@` DESC) AS __rank__
    FROM dbck.employees
    SETTINGS allow_experimental_window_functions = 1
)
WHERE (__rank__ = 1) AND (`__event_type@@` != 'D')
```

字段拼接

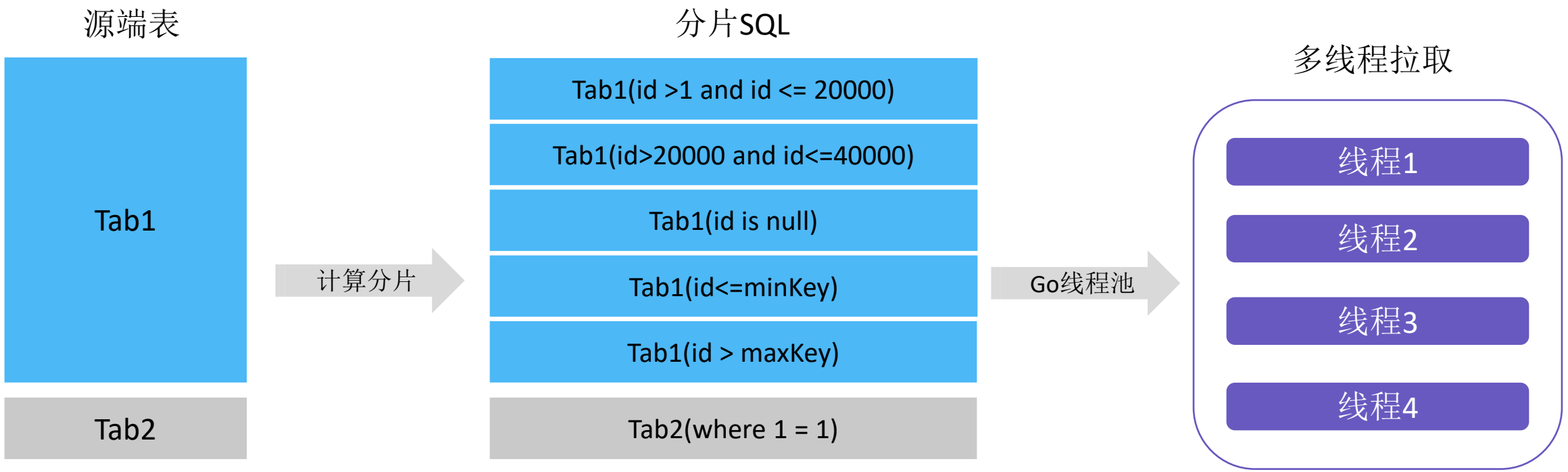
字段拼接

主键字段

版本倒排

外层取最大版本和非D删除的数据

ClickHouse数据接入- 全量并发拉取



分片逻辑

表级并发 vs 表内并发
默认主键，ID数值型字段，时间字段
不符合条件不做分片

参数控制

--work-threads 工作线程数
--split-rowcount 单分片行数限制
--repeat-read一致读
--commit-batchsize 批次提交行数
--fetch-batchsize fetch批次行数

ClickHouse数据接入- 全量断点续传

迁移的状态
- 多久可以迁移完成，进度如何
- 是否有失败的对象
- 迁移对象的状态维护

迁移的异常控制
- 增量断点续传
- 如何应付网络较差的数据库
- 全量也要断点续传？？？

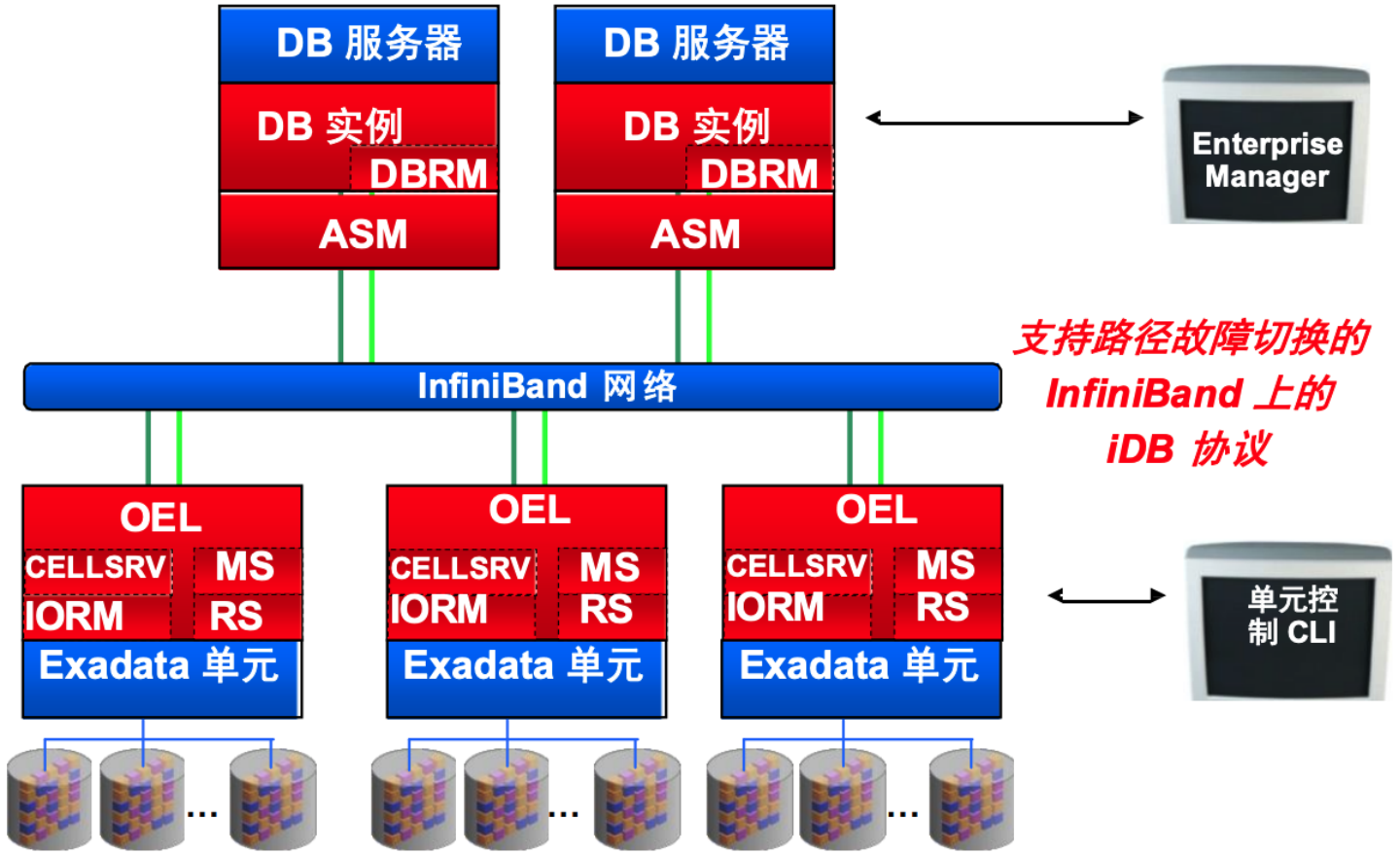
优化前

表名	迁移状态	估算行数	已迁移行数
saledb.tab1	failed	50000	49000
orderdb.order	success	3000	3000

优化后

表名	分片拉取SQL	分片迁移状态	已迁移行数
saledb.tab1	id >1 and id <= 20000	success	19999
saledb.tab1	id > 20000 and id <= 40000	failed	300
saledb.tab1	id <=1	success	1
saledb.tab1	id >= 40000	success	0
orderdb.order	Where 1 = 1	success	3000

从Oracle一体机看大数据技术



Thanks

Your Data, Your Cloud, We Integrate!

罗春 @Squids

15906620338