ClickHouse 在快手的大规模应用与架构改进

李振炜

快手 数据平台部 大数据架构工程师

SPEAKER
INTRODUCE

李振炜 大数据架构工程师

- 2015年硕士毕业之后加入奇虎360
- 2018年加入快手
- 熟悉spark，presto和clickhouse

TABLE OF

CONTENTS 大纲

- ClickHouse 简介
- ClickHouse 在快手的应用现状
- ClickHouse on HDFS
- ClickHouse后续改进计划

ClickHouse 简介

ClickHouse 是俄罗斯Yandex在2016年开源的一个高性能分析型SQL数据库，主要面向OLAP场景。开源之后，凭借优异的查询性能，受到业界的青睐。

ClickHouse 简介

快手

性能对比

Compare: ClickHouse Vertica Vertica (x3) Vertica (x6) InfiniDB MonetDB Infobright Hive MySQL MemSQL Greenplum(x2) Greenplum

Dataset size: 10 mln. 100 mln. 1 bn.

Run number: first (cold cache) second third

Relative query processing time (lower is better):

ClickHouse官网提供，详见https://clickhouse.yandex/benchmark.html

ClickHouse 简介

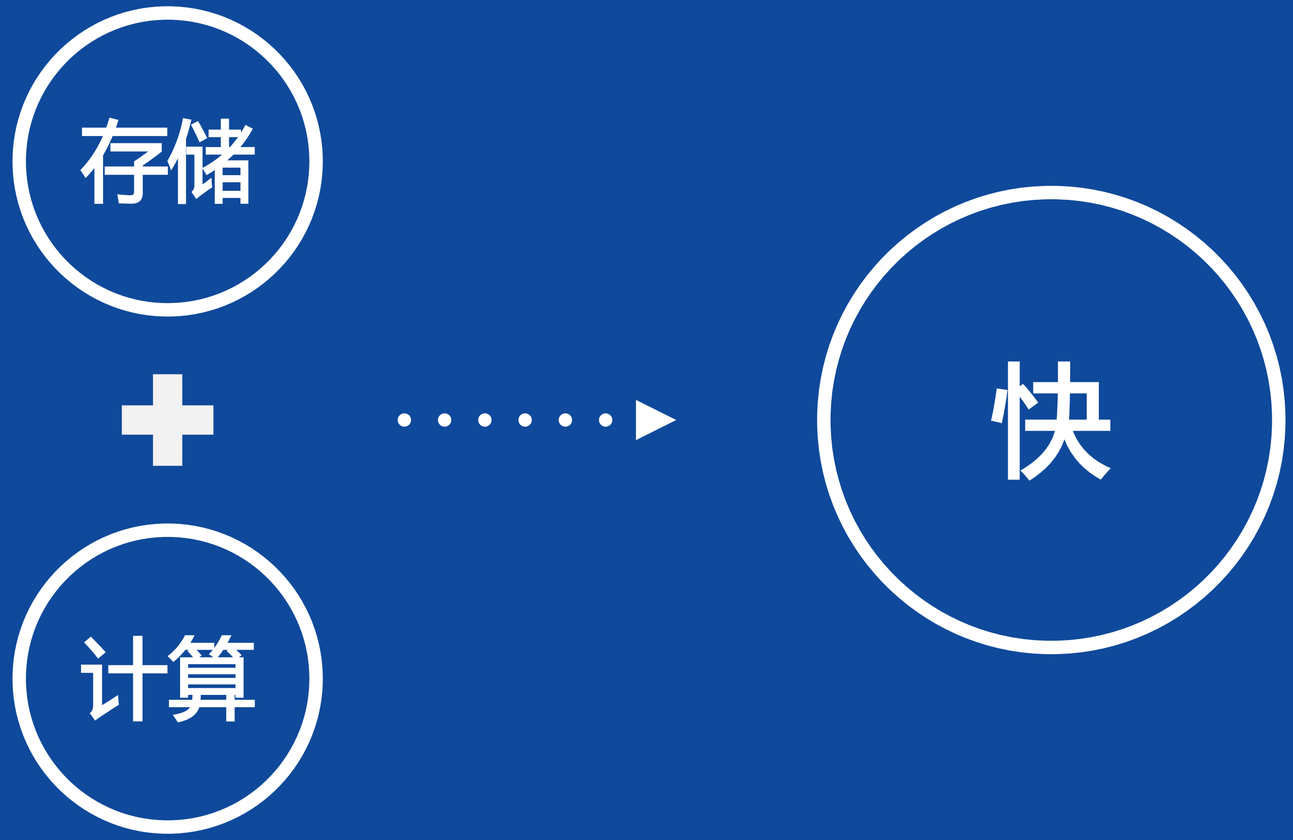
快手

性能对比

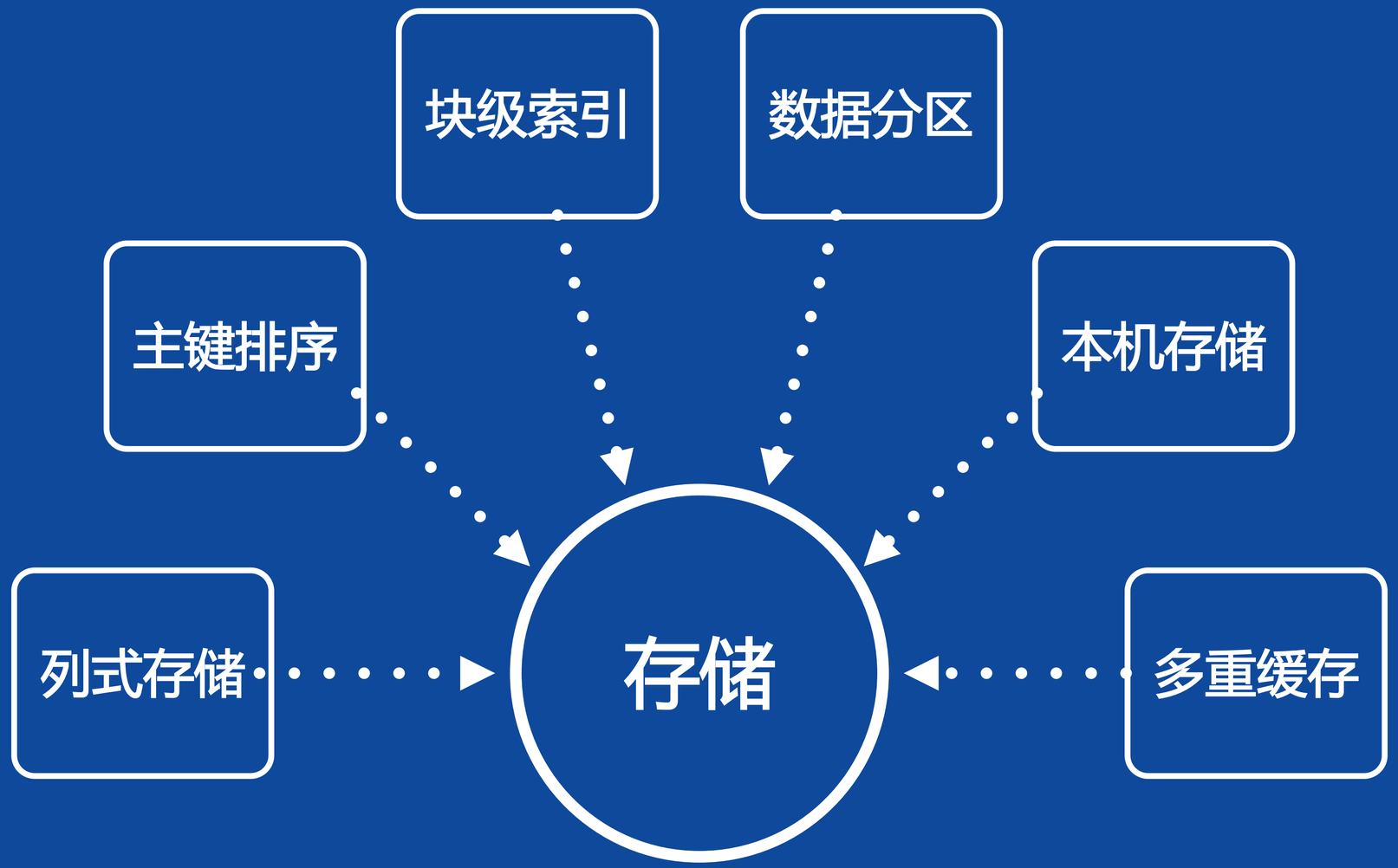
sql语句（单表测试语句）	Hawq	presto(orc格式)	Impala(parquet格式)	spark-sql(orc格式)	ClickHouse	greenplum	hive(orc格式)
sql_01	12.734	1.08	1.53	6.66	0.307	9.018	51.45
sql_02	15.578	2.1	4.04	9.62	0.515	10.887	129.78
sql_03	16.774	3.03	4.85	8.95	0.759	11.247	130.7
sql_04	23.469	5.78	11.59	11.06	0.477	20.137	185.38
sql_05	12.547	3.26	1.32	4.75	0.443	8.694	50.05
sql_06	88.506	29.55	43.16	43.43	12.341	89.75	343.86
sql_07	86.468	28.89	45.16	41.34	12.198	90.318	346.92
sql_08	134.72	68.23	72.32	90.28	19.217	154.77	455.37
sql_09	133.69	54.18	72.45	98.59	39.669	221.782	2402.521
总时间	524.486	196.1	256.42	314.68	85.926	616.603	4096.031

易观测评结果，详见http://www.clickhouse.com.cn/topic/5c453371389ad55f127768ea

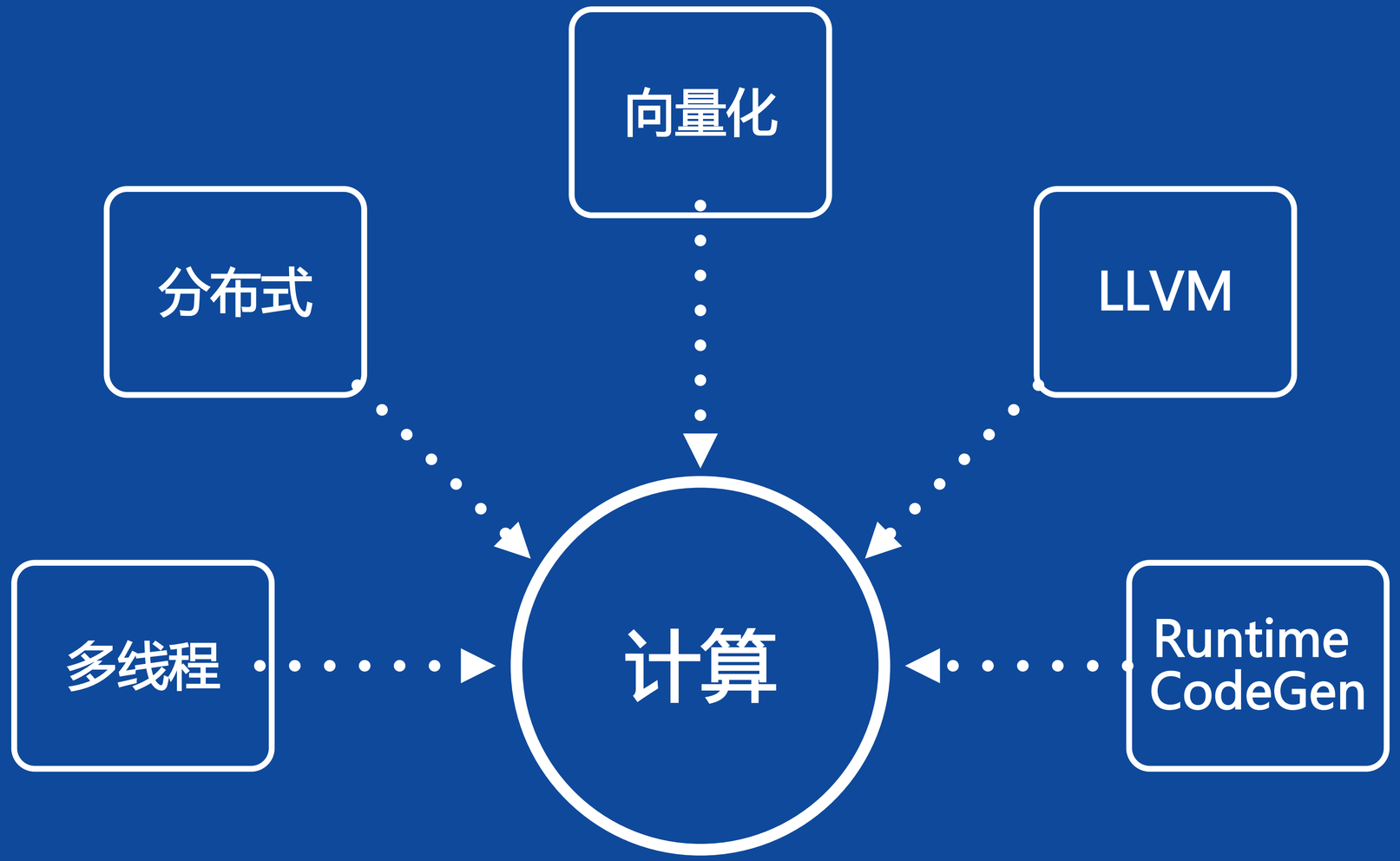
ClickHouse 简介



ClickHouse 简介



ClickHouse 简介

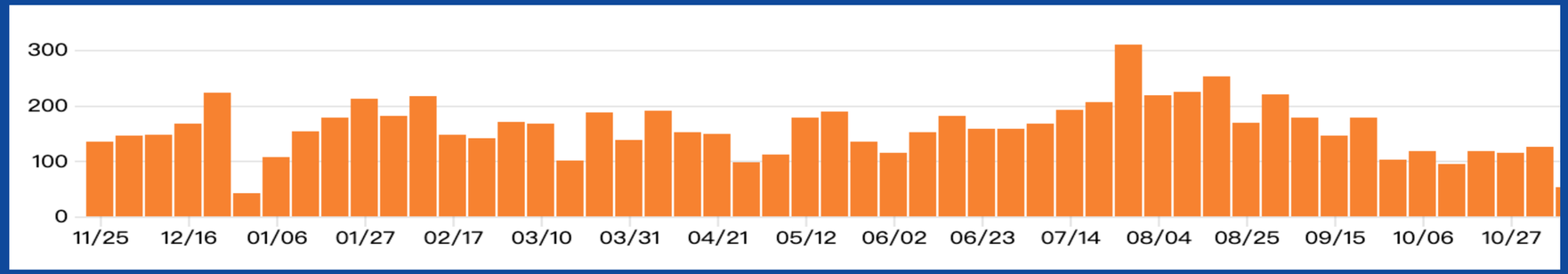


ClickHouse 简介

业界使用

- 国内：快手，头条，腾讯，京东，新浪，携程等，
- 国外： Yandex, CloudFlare, Spotify等

社区活跃



ClickHouse 在快手的应用现状

ClickHouse在快手规模

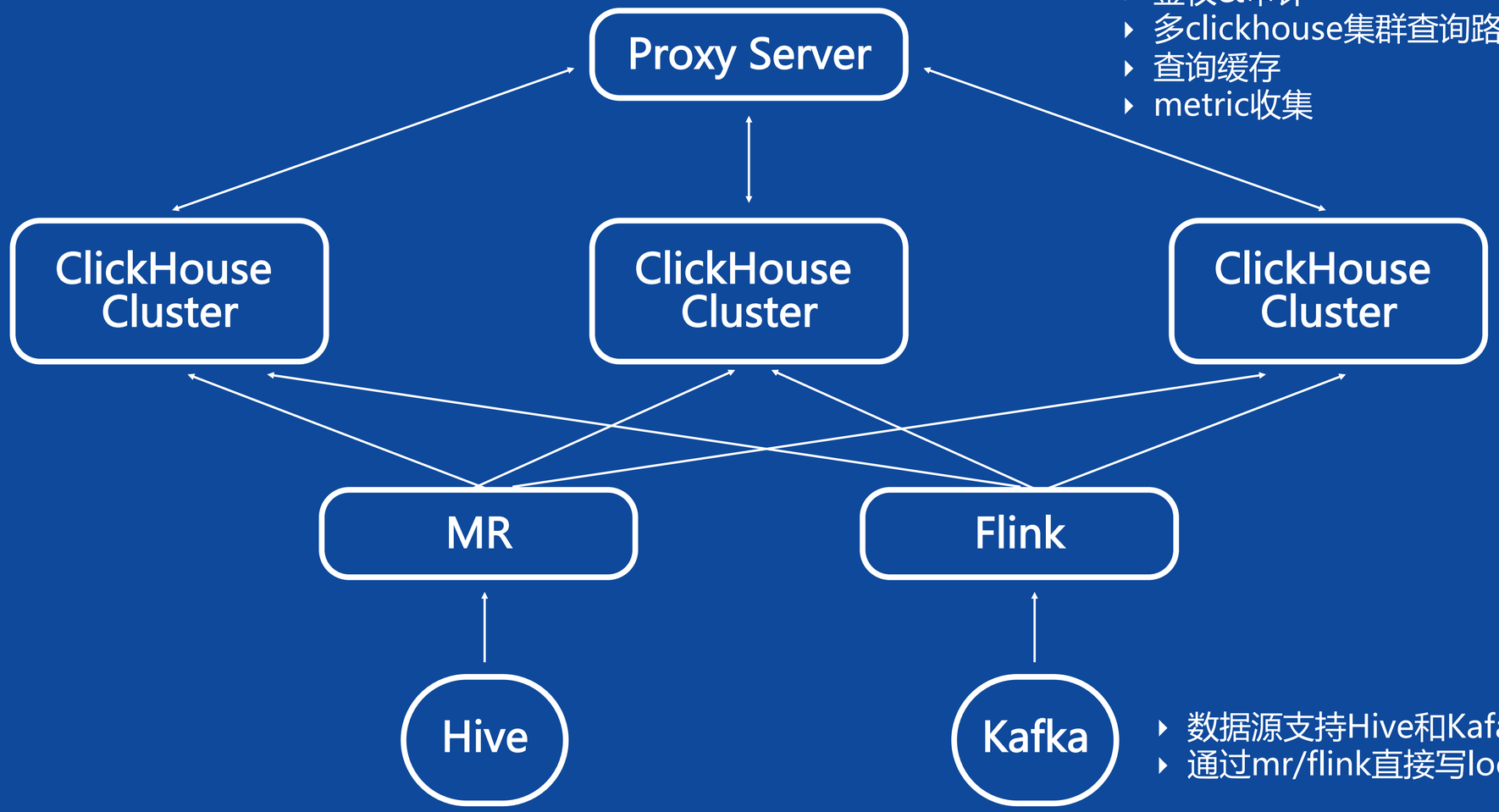
存储数据量	~10PB
每天新增数据量	~200TB
每天查询	~50w
查询P90时延	<3s

(截止2019年10月份数据)

ClickHouse 在快手的应用现状

ClickHouse 在快手的应用现状

ClickHouse在快手部署架构



- 鉴权&审计
- 多clickhouse集群查询路由
- 查询缓存
- metric收集

- 多集群建设，资源物理隔离
- 每个shard有两个副本

- 数据源支持Hive和Kafaka
- 通过mr/flink直接写local表

ClickHouse 在快手的应用现状

CHProxy介绍

ClickHouse 在快手的应用现状

localQuery特性介绍

- table engine
- 语法：localQuery(clickhouse_cluster, db.table, sql)
- 作用：sql全部在单个shard执行，然后返回结果
- 数据：全部按关键列hash分shard处理

ClickHouse 在快手的应用现状

使用经验分享

- 单表分析场景性能最好
- 相关字段sharding,采用local in/join
- 复杂操作，尽量在每个节点完成
- 不建议作为一个事务型数据库使用

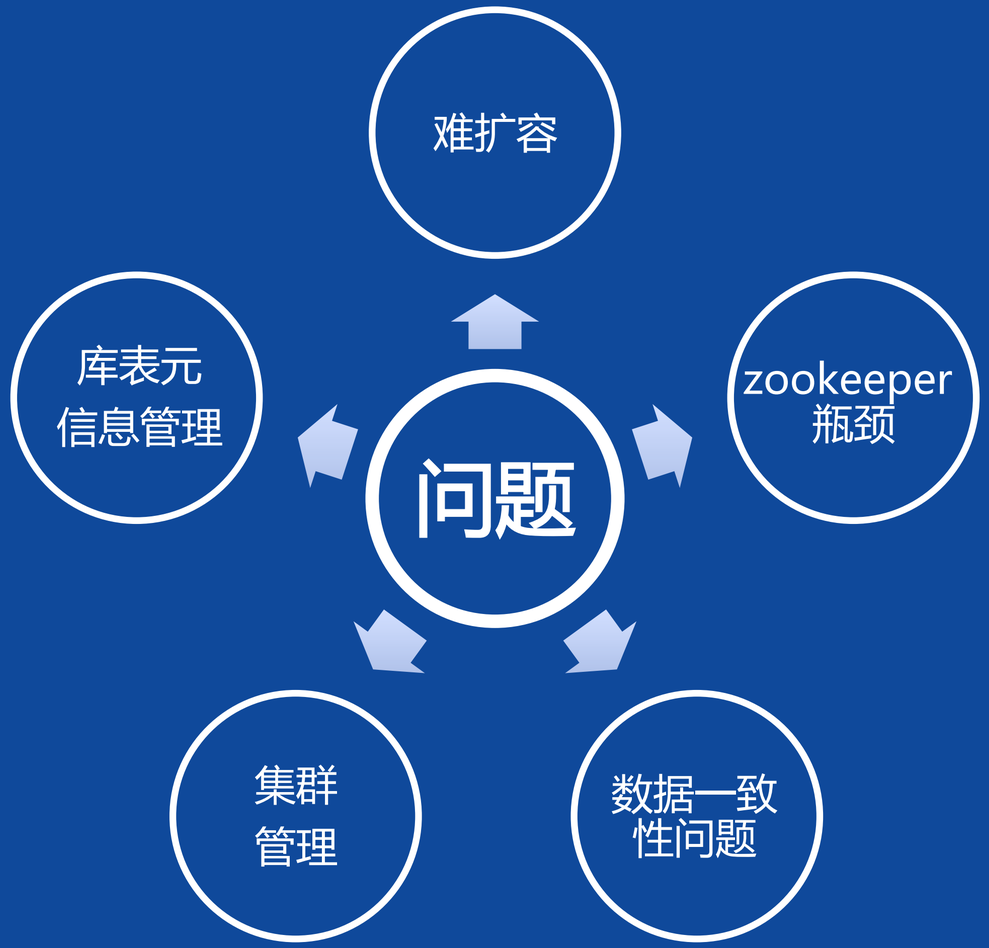
面临的问题与解决方案

面临的问题

解决方案

实测效果

面临的问题与解决方案



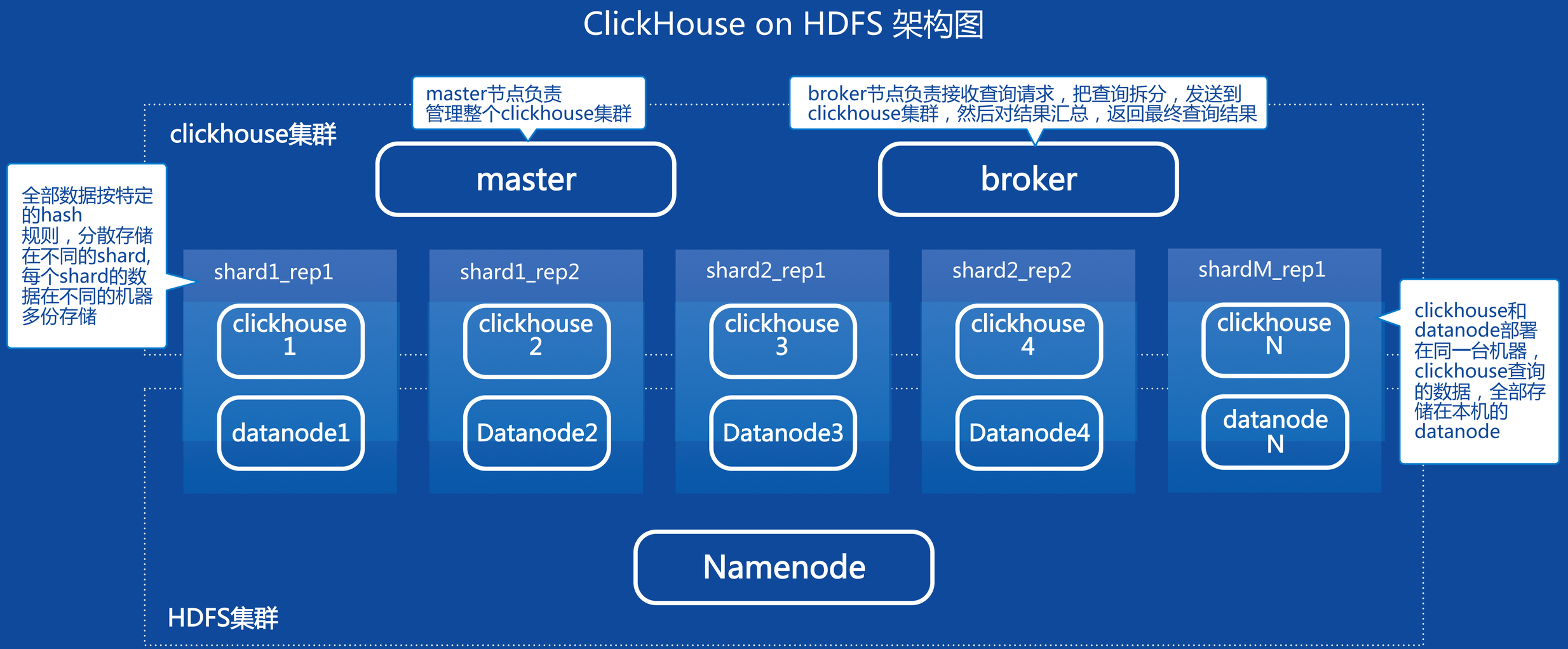
ClickHouse on HDFS

快手

解决思路

存储和计算分离，数据存储管理依托于成熟的平台

ClickHouse on HDFS



ClickHouse on HDFS

面临的问题

小文件问题

文件读取性能问题

ClickHouse on HDFS

每个part下面所有的MergeTree文件合并成一个文件

实现接口，能够从合并之后的文件中读取MergeTree数据

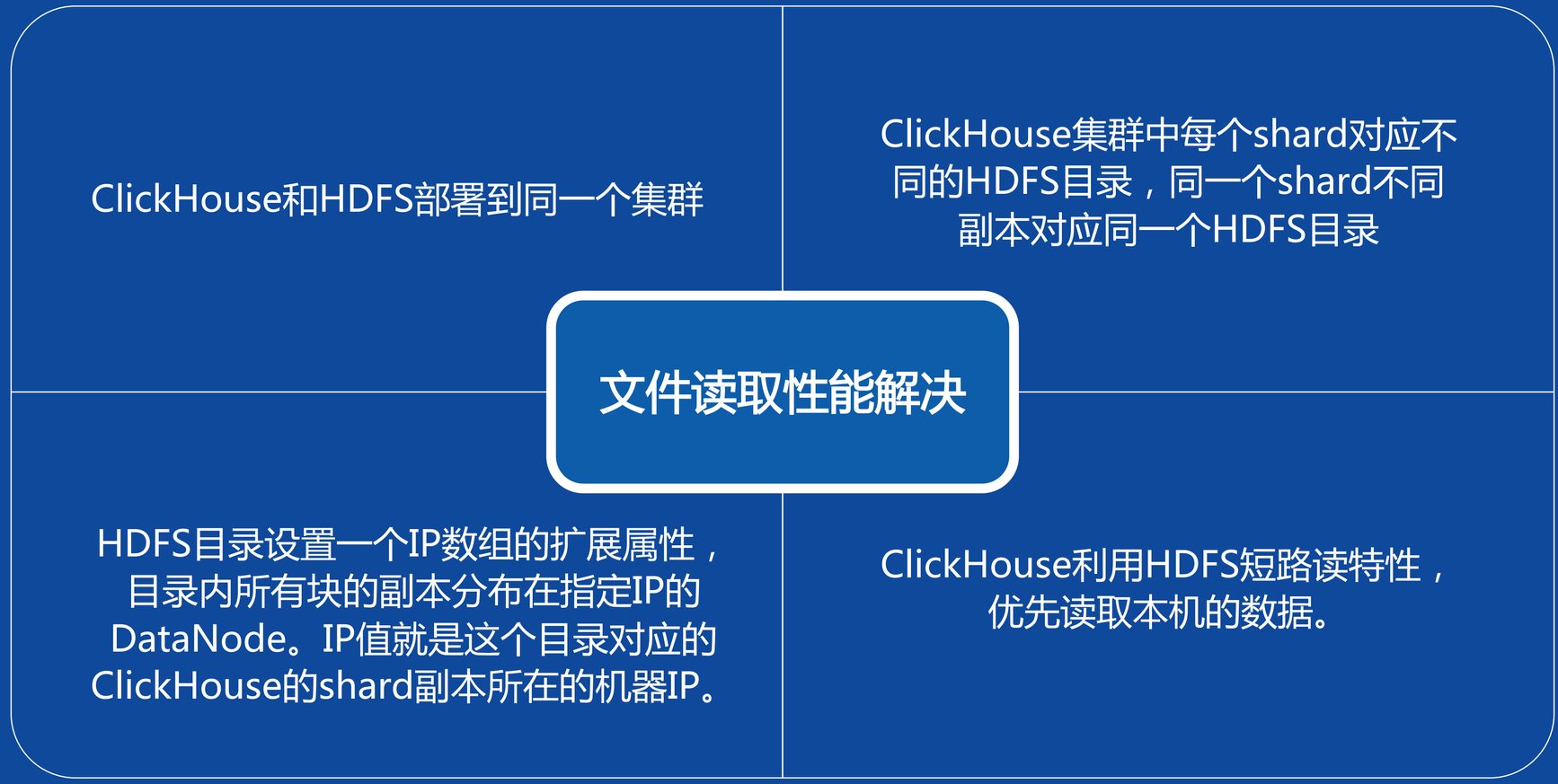
小文件问题解决

ClickHouse on HDFS

HDFSMergeTree

- table engine
- 继承全部MergeTree特性，完全兼容MergeTree操作
- 实现了HDFSMergeTreeReader和HDFSMergeTreeReaderStream
- 语法：HDFSMergeTree(/home/${clickhouse_cluster})

ClickHouse on HDFS



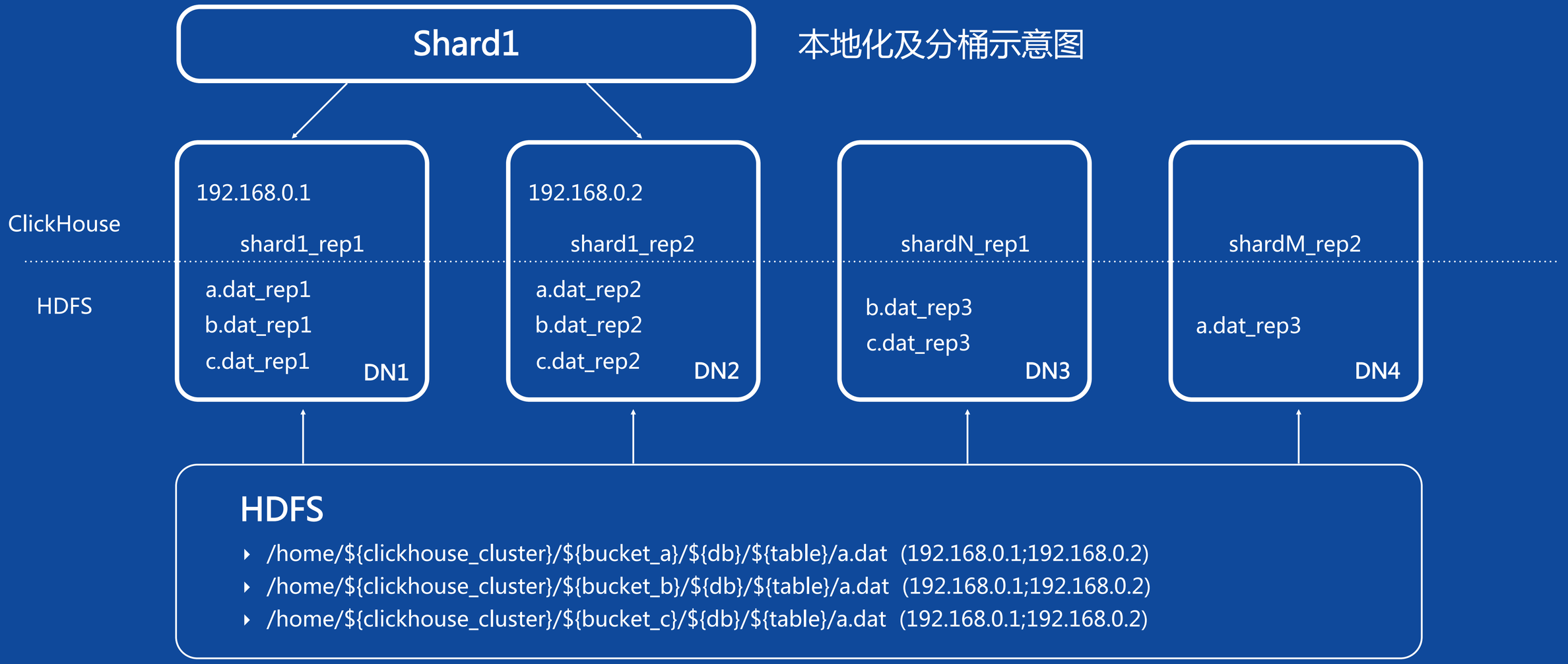
ClickHouse on HDFS

快手

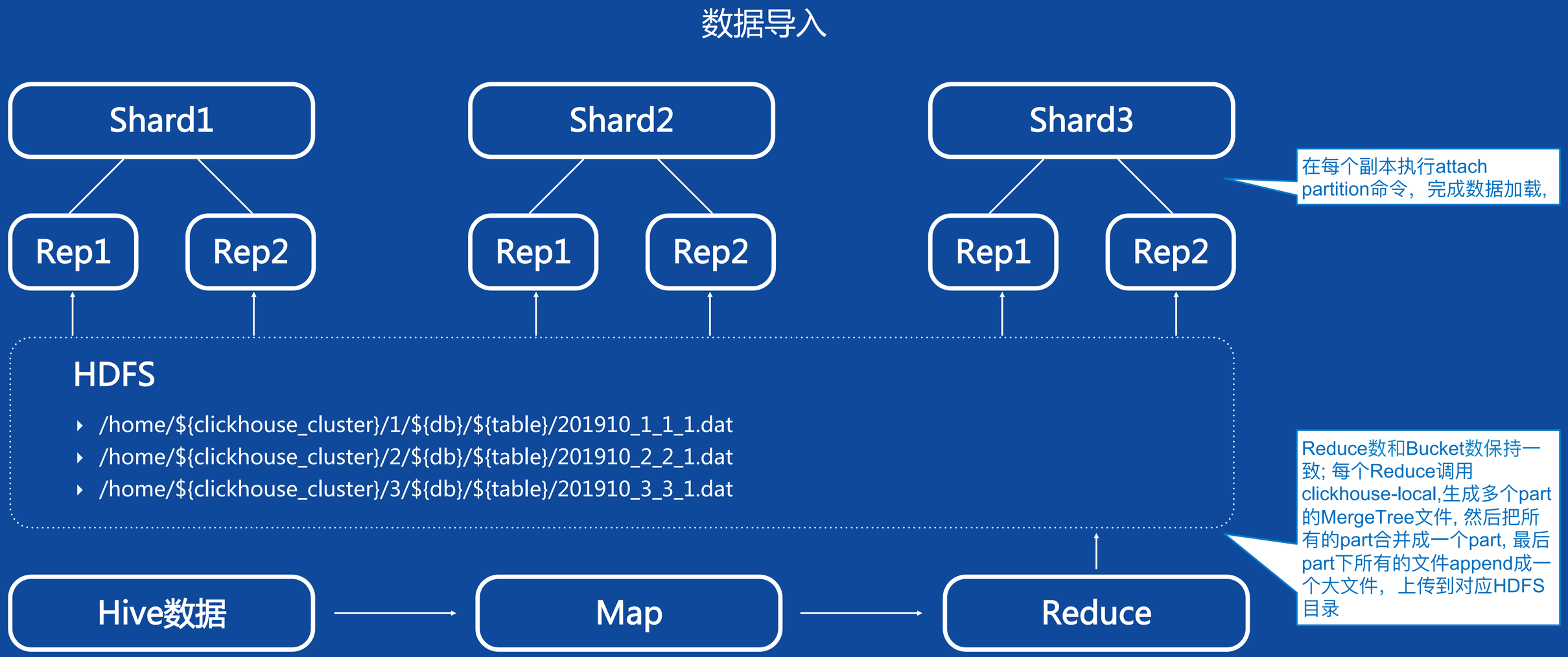
扩容问题

- 数据在HDFS上分bucket存储，每个shard对应一个或多个bucket
- HDFS目录组织形式：/home/${clickhouse_cluster}/${bucket_num}/${db}/${table}/
- IP扩展属性设置到${bucket_num}目录这一级
- 每个shard配置shard_num, cluster_size, bucket_total三个参数
  shard_num%cluster_size == [1, bucket_total]%cluster_size

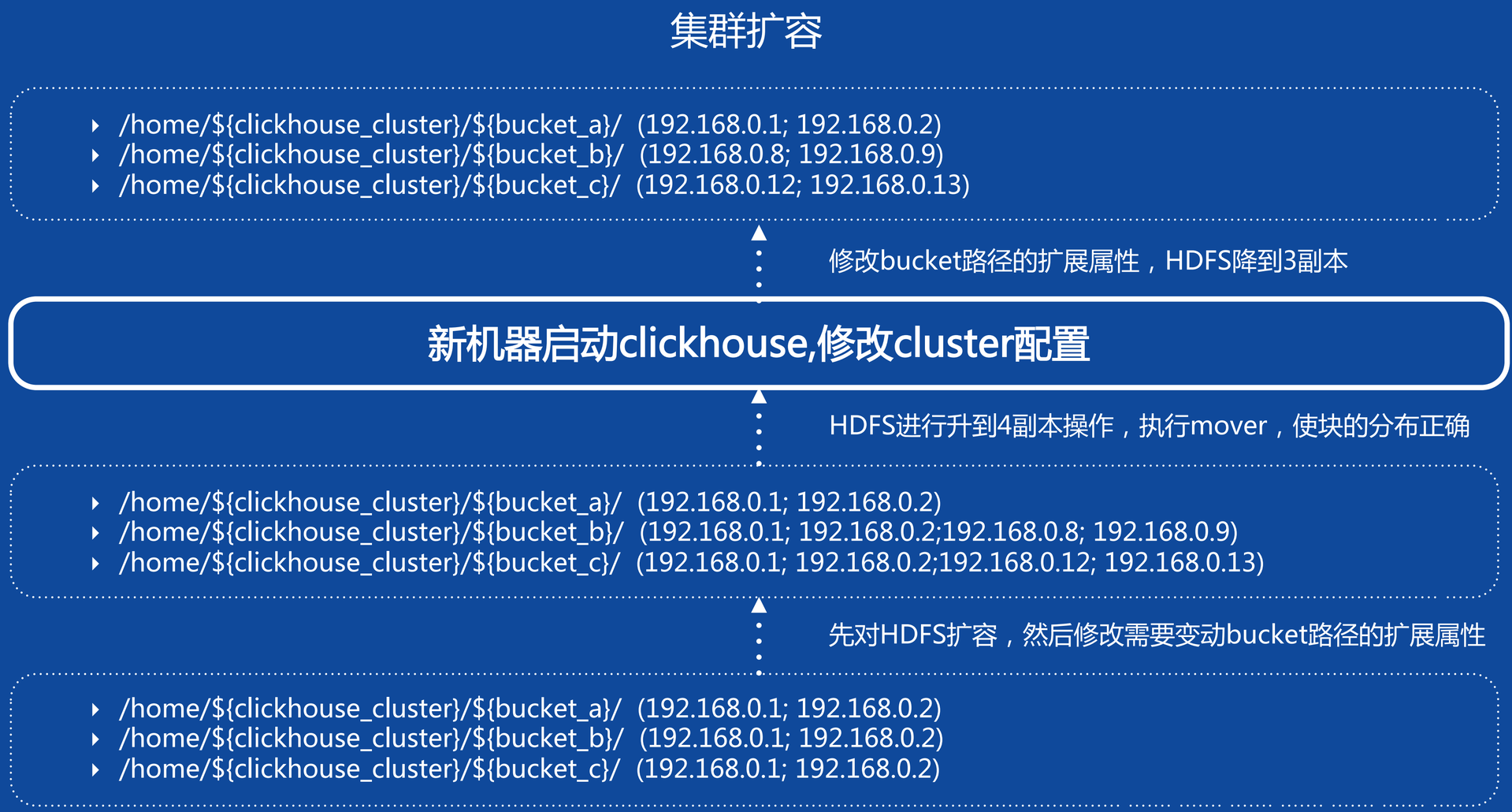
ClickHouse on HDFS



ClickHouse on HDFS



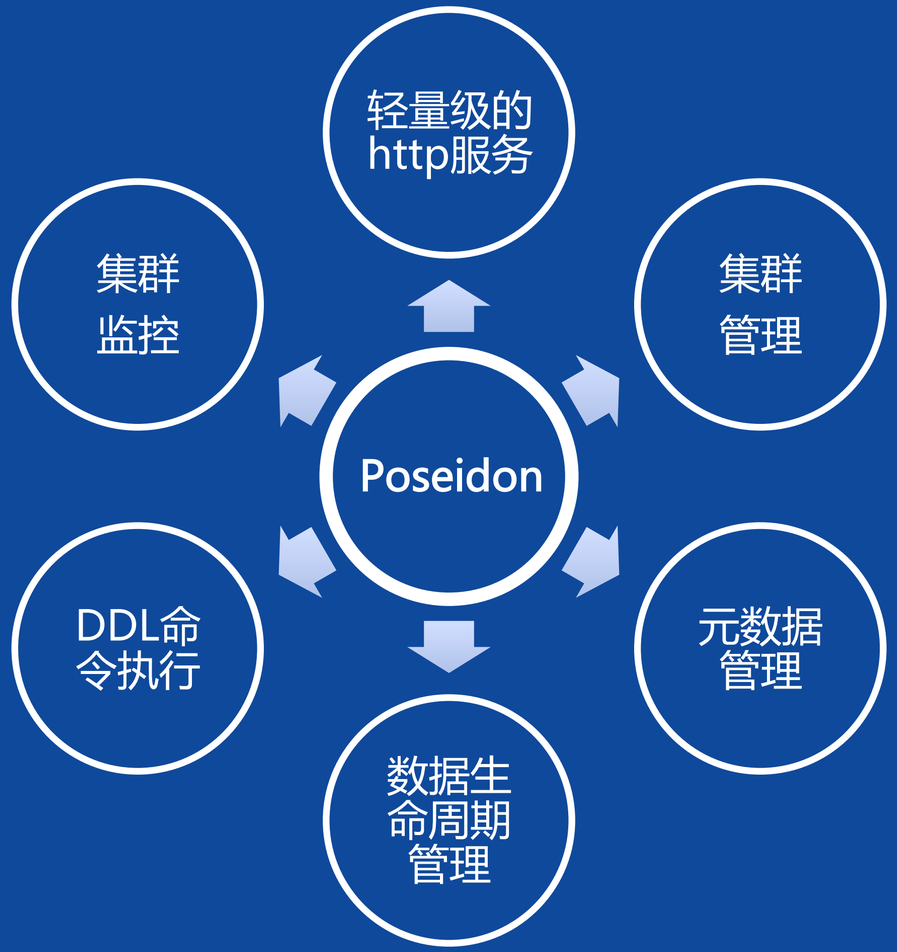
ClickHouse on HDFS



ClickHouse on HDFS

Poseidon

ClickHouse on HDFS



ClickHouse on HDFS

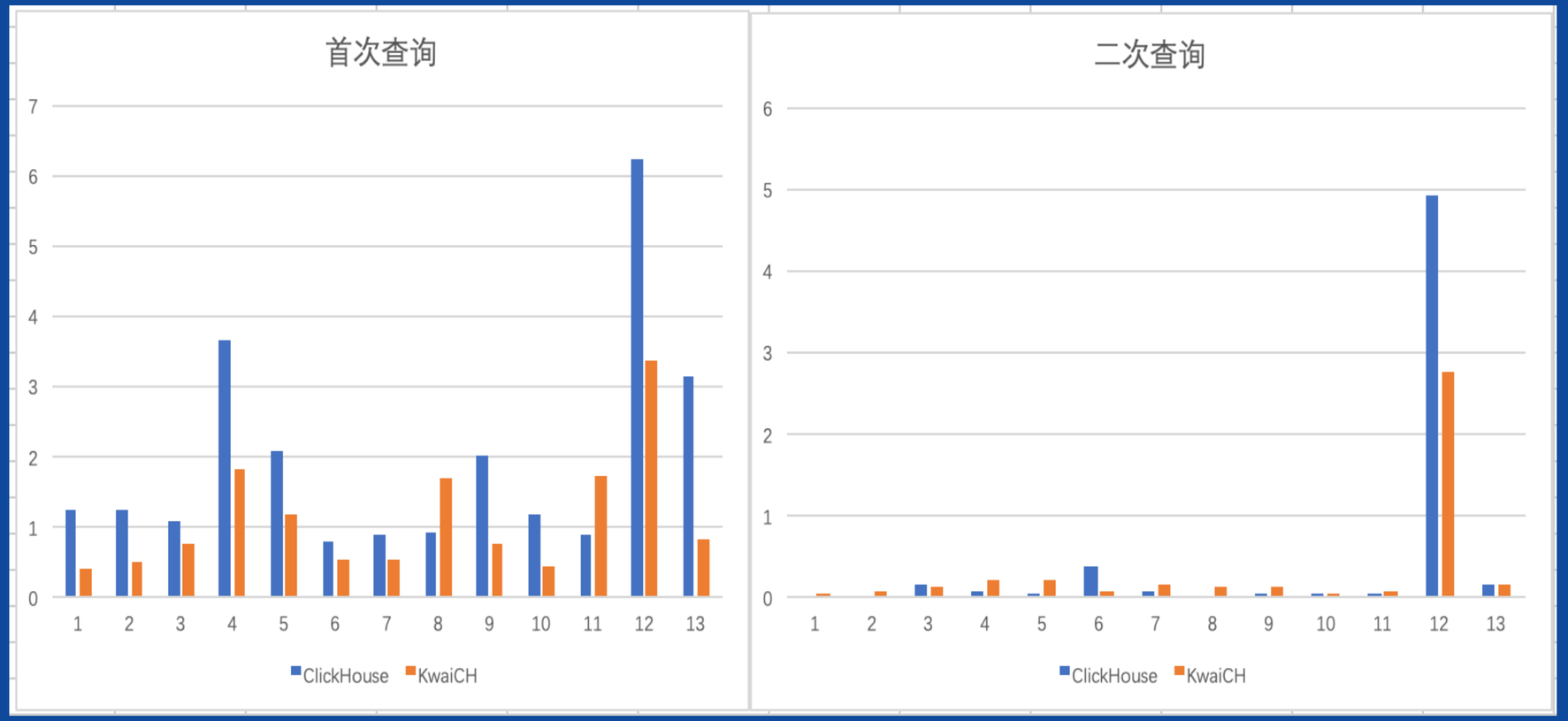
快手

其他优化点

- 缓存块的location信息，减少RPC请求耗时
- 缓存文件句柄
- 检测块是否分布在指定IP的datanode
- Mover工具定时修正块的分布

ClickHouse on HDFS

性能对比(机器环境，数据量，查询sql完全对等)



查询在100ms以上的SQL，性能提升20%~70%，查询在100ms以下的SQL，性能降低50%~80%

ClickHouse on HDFS

性能分析

HDFSMergeTree分bucket, part的数量是多于原来，ClickHouse的索引是粗粒度的，最少读取索引到一个块，造成HDFSMergeTree 会多读一些数据。

HDFS上的数据块会均匀分布在每台机器多块SATA盘中，MergeTree的raid盘, 多块并发读取的速度大于raid的速度

ClickHouse on HDFS

总结

性能	数据	成本	运维
利用HDFS短路读特性	剥离数据管理功能	相同磁盘，比raid能存储更多数据	扩容方便
缓存必要的信息	提高数据可靠性		支撑超大规模
充分利用多块盘提高读取性能			减少运维压力
保留原生ClickHouse高性能执行部分			

后续计划

- 实时导入实现计算存储分离
- 增强ClickHouse的SQL优化器
- 智能视图、聚合索引增强
- 运维自动化

CHINA · BEIJING

THANKS

Global
Architect Summit