
SACC
2021 中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2021
数字转型 架构重塑
IT168.com
ChinaUnix.net
ITPUB
云上会议 网络直播 | 2021.5.20-2021.5.22

贝壳Hadoop集群演进

- 关于贝壳
- Hadoop集群概况
- 集群演进
- 未来规划

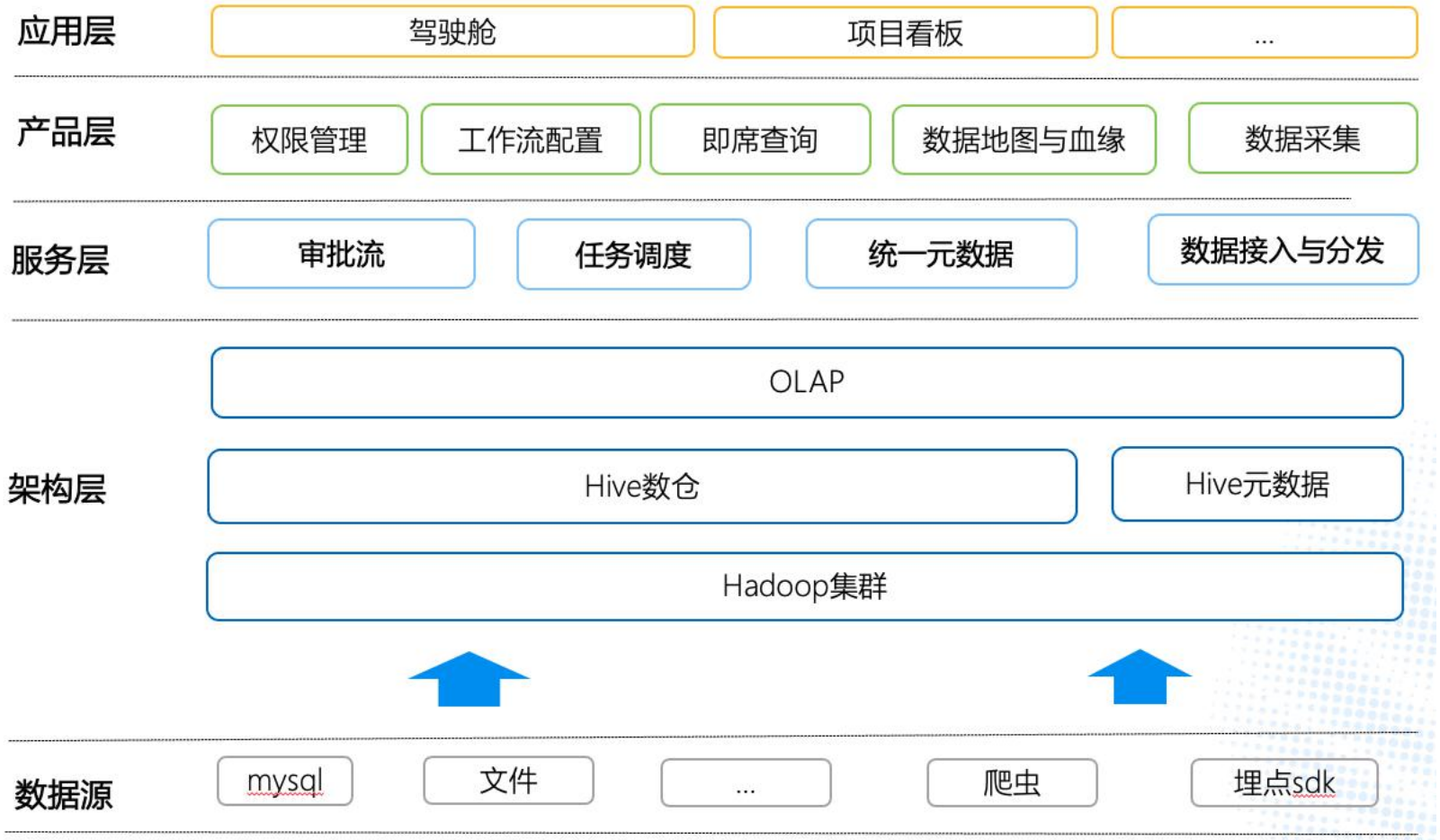
关于贝壳


贝壳找房
是科技驱动的新居住服务平台，我们致力于为3亿家庭提供全面、可靠的品质居住服务。
二手房
新房
租赁
装修
其他

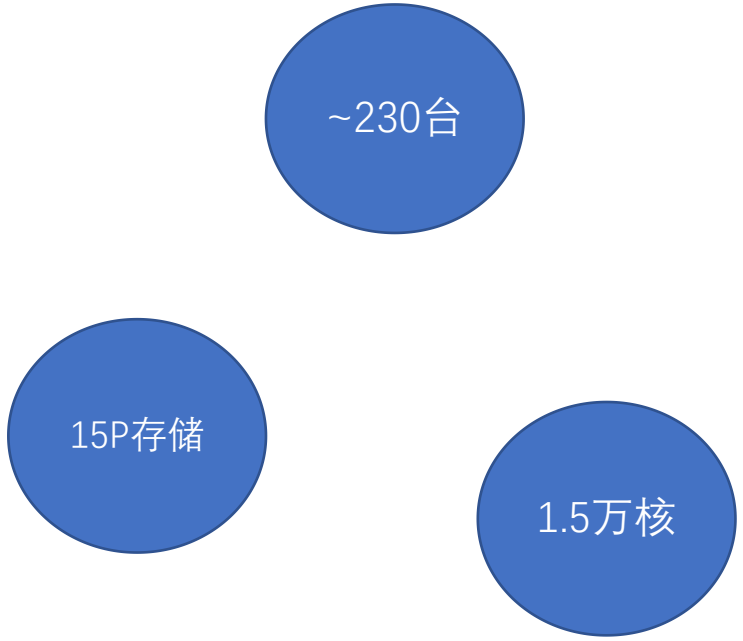
我们的发展历程

传承使命 自我进化

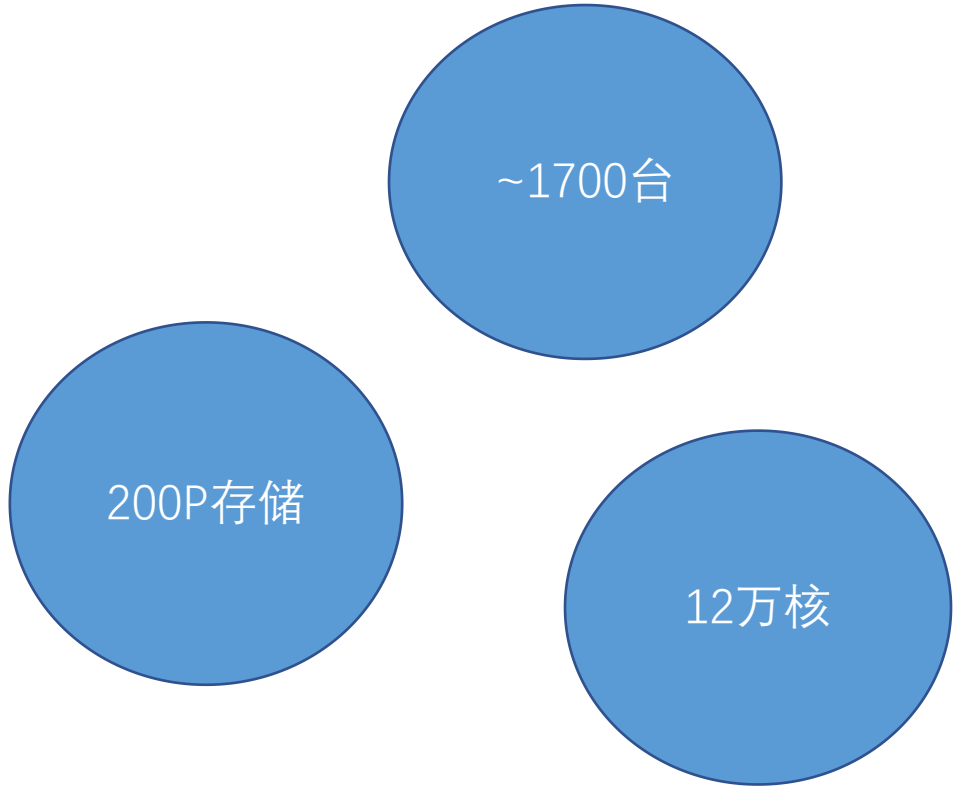
Hadoop集群概况


应用层
驾驶舱
项目看板
...
产品层
权限管理
工作流配置
即席查询
数据地图与血缘
数据采集
服务层
审批流
任务调度
统一元数据
数据接入与分发
架构层
OLAP
Hive数仓
Hive元数据
Hadoop集群
数据源
mysql
文件
...
爬虫
埋点sdk

Hadoop集群概况



2018年5月



2021年5月

Hadoop集群概况

- Hadoop集群服务于业务，业务与企业定位、规模相关
- 链家时代，集群更多用于存储数据，规模预期小
- 贝壳时代，集群更多用于挖掘数据价值，规模预期大

集群演进-存储治理

- 透明压缩
  - 1.5 副本
  - HDFS分层存储 + ZFS文件系统
  - 2017年10月上线
  - 版本 2.7.3

集群演进-存储治理

- ZFS是什么？
  - OpenSolaris开源计划的一部分，ZFS于2005年11月发布
  - 支持压缩

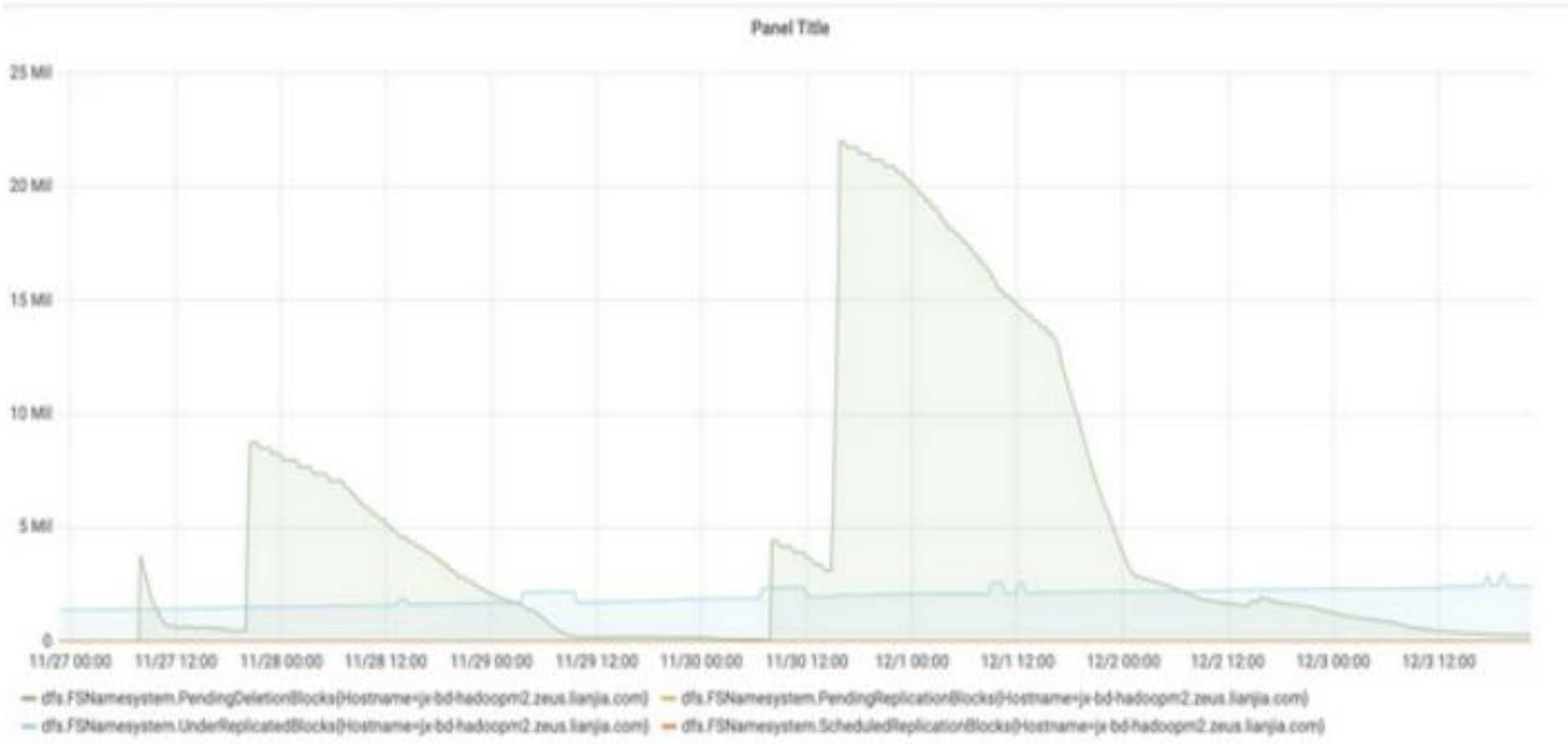
集群演进-存储治理

- 透明压缩的问题
  - ZFS不可控
  - Datanode节点稳定性下降
  - Namenode性能下降

集群演进-存储治理

- Namenode性能下降
  - 单台机器下线需要数天
  - UnderReplicatedBlocks 缓慢增加
  - PendingDeletionBlocks 下降缓慢

集群演进-存储治理


Panel Title
25 Mil
20 Mil
15 Mil
10 Mil
5 Mil
0
11/27 00:00
11/27 12:00
11/28 00:00
11/28 12:00
11/29 00:00
11/29 12:00
11/30 00:00
11/30 12:00
12/1 00:00
12/1 12:00
12/2 00:00
12/2 12:00
12/3 00:00
12/3 12:00
dfs.FSNamesystem.PendingDeletionBlocks{Hostname=jx-bd-hadoopm2.zeus.lianjia.com}
dfs.FSNamesystem.PendingReplicationBlocks{Hostname=jx-bd-hadoopm2.zeus.lianjia.com}
dfs.FSNamesystem.UnderReplicatedBlocks{Hostname=jx-bd-hadoopm2.zeus.lianjia.com}
dfs.FSNamesystem.ScheduledReplicationBlocks{Hostname=jx-bd-hadoopm2.zeus.lianjia.com}

集群演进-存储治理

- 调用栈信息

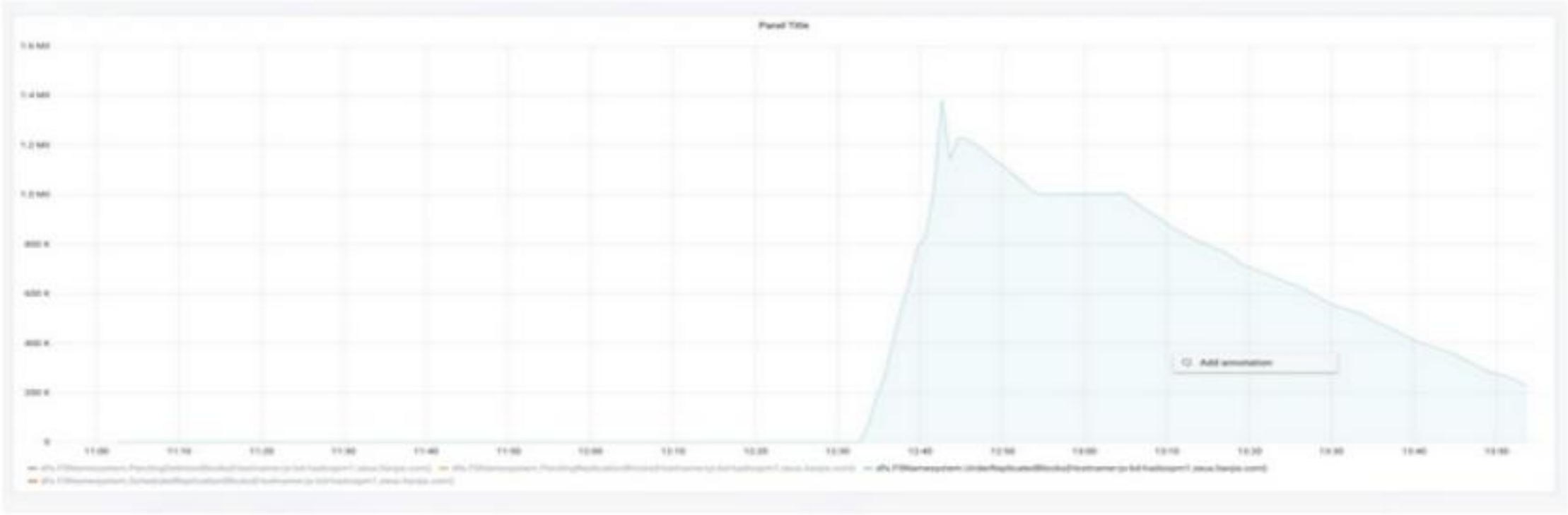
```
"org.apache.hadoop.hdfs.server.blockmanagement.BlockManager$ReplicationMonitor@64a8c844" #34 daemon prio=5 os_prio=0 tid=0x00007f772e03a800 nid=0x6288f runnable [0x00007f4507c0f000]
   java.lang.Thread.State: RUNNABLE
        at org.apache.hadoop.net.NetworkTopology$InnerNode.getLoc(NetworkTopology.java:296)
        at org.apache.hadoop.net.NetworkTopology$InnerNode.getLoc(NetworkTopology.java:296)
        at org.apache.hadoop.net.NetworkTopology.getNode(NetworkTopology.java:556)
        at org.apache.hadoop.net.NetworkTopology.countNumOfAvailableNodes(NetworkTopology.java:808)
        at org.apache.hadoop.net.NetworkTopologyWithMultiDC.countNumOfAvailableNodes(NetworkTopologyWithMultiDC.java:259)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyDefaultWithMultiDC.chooseRandom(BlockPlacementPolicyDefaultWithMultiDC.java:803)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyDefaultWithMultiDC.chooseTarget(BlockPlacementPolicyDefaultWithMultiDC.java:473)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyDefaultWithMultiDC.chooseTarget(BlockPlacementPolicyDefaultWithMultiDC.java:300)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyDefaultWithMultiDC.chooseTarget(BlockPlacementPolicyDefaultWithMultiDC.java:177)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager$ReplicationWorkWithMultiDC.chooseTargets(BlockManager.java:4448)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.computeReplicationWorkForBlocksWithMultiDC(BlockManager.java:1740)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.computeReplicationWork(BlockManager.java:1419)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.computeDatanodeWork(BlockManager.java:4341)
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager$ReplicationMonitor.run(BlockManager.java:4293)
        at java.lang.Thread.run(Thread.java:748)
~
~
```

集群演进-存储治理

- 根因分析
  - 存储异构，周期性将历史数据设置为COLD，启动Mover进程归档数据
  - Mover更新StorageType远慢于数据置为COLD的动作
  - chooseStorageTypes未考虑副本的StorageType与该文件设置存储策略的匹配与否，导致chooseStorageTypes返回副本需求与result表示的已经存在副本存在逻辑上的冲突（ chooseTarget 方法）
  - 修改chooseStorageTypes代码，将result传入，如果类型与当前已有块不匹配，将该副本从result删除

集群演进-存储治理

- 效果– UderReplicatedBlock快速下降， 详情参考 HDFS-15715



集群演进-存储治理

- 搭建 Hadoop 3.2.1集群，启用EC
- 高密度存储机型（24 * 12T），进一步控制存储成本
- 依托元数据与生命周期管理，进行存储编排、流转

集群演进-存储治理

- 解决一个问题带来更多问题
- 要优先选用社区、业界主流方案

集群演进-搬迁

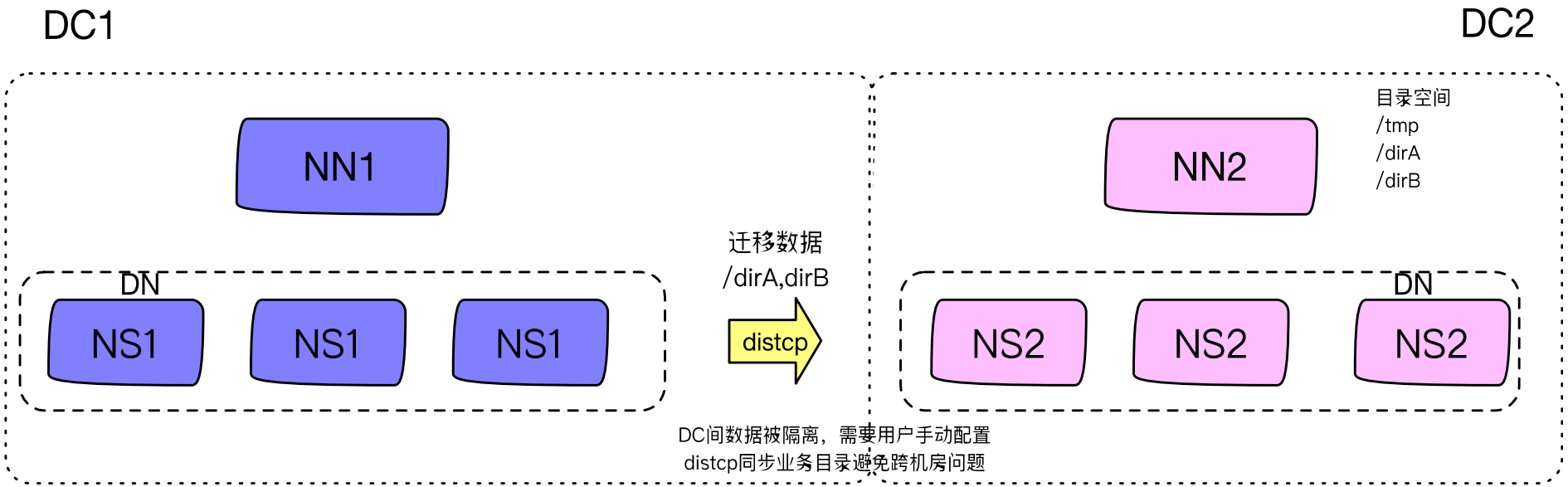
- 搬迁的原因与动机
  - 机房容量规划有问题，无法扩容
  - 现有集群问题解决成本高
  - 获得3.2.1 版本红利，EC，全局调度等
  - 拆分Namespace

集群演进-搬迁

- 搬迁升级的内容
  - 搬迁，从亦庄机房搬迁到通州机房
  - 升级，从2.7.3升级到3.2.1
  - 拆分，将单ns的对象数控制在4亿以下

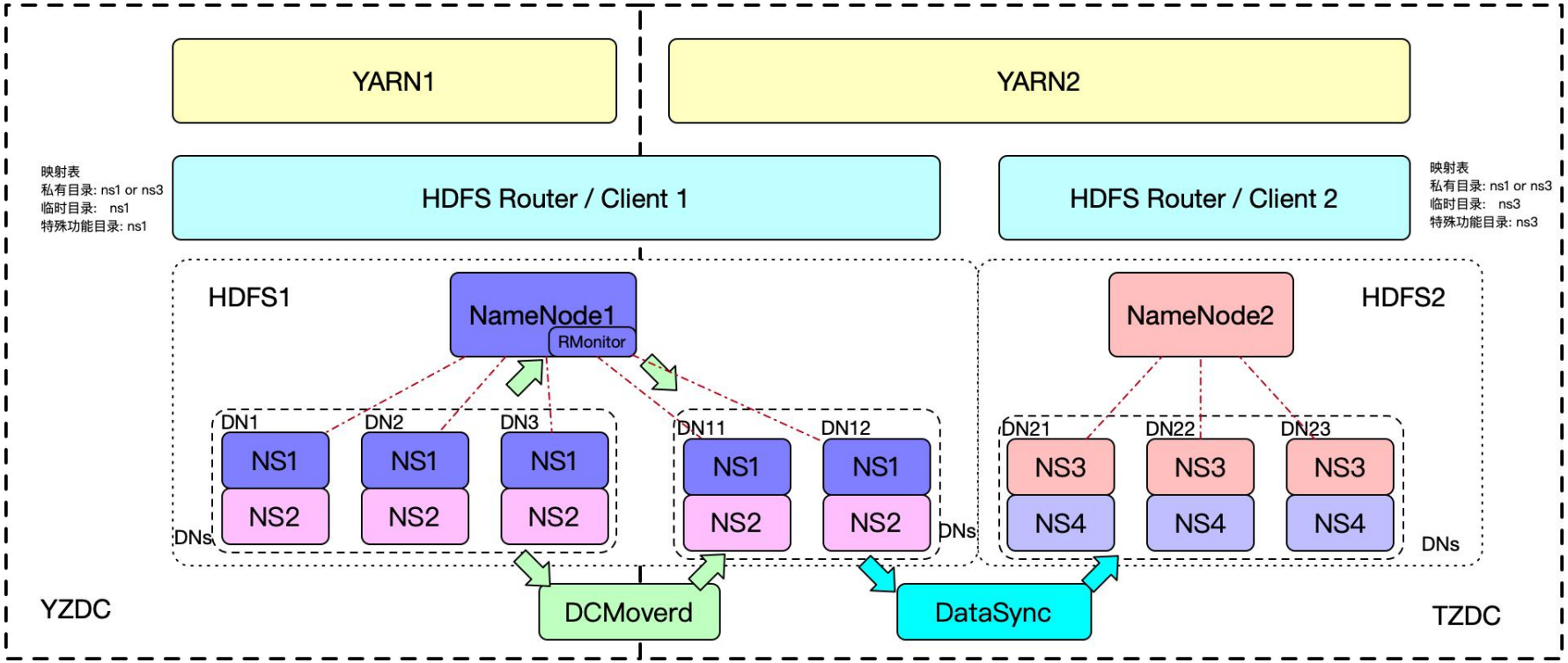
集群演进-搬迁

- 普通方案


DC1
DC2
NN1
NN2
目录空间
/tmp
/dirA
/dirB
迁移数据
/dirA,dirB
distcp
DN
NS1
NS1
NS1
DN
NS2
NS2
NS2
DC间数据被隔离，需要用户手动配置
distcp同步业务目录避免跨机房问题

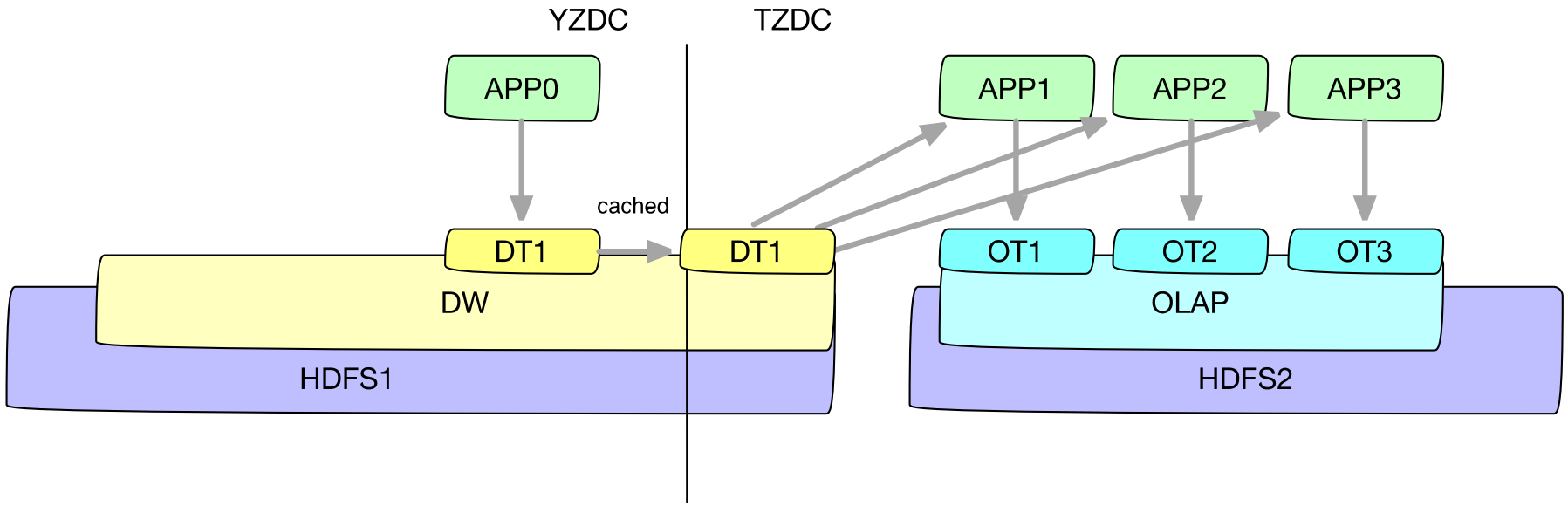
集群演进-搬迁

- 贝壳方案


YARN1
YARN2
映射表
私有目录: ns1 or ns3
临时目录: ns1
特殊功能目录: ns1
HDFS Router / Client 1
HDFS Router / Client 2
映射表
私有目录: ns1 or ns3
临时目录: ns3
特殊功能目录: ns3
HDFS1
NameNode1
RMonitor
DN1
DN2
DN3
NS1
NS2
DNs
DN11
DN12
DNs
NameNode2
HDFS2
DN21
DN22
DN23
NS3
NS4
DNs
YZDC
DCMoverd
DataSync
TZDC

集群演进-搬迁

- 贝壳方案-多机房读写策略



集群演进-搬迁

	普通方案	贝壳方案
业务感知	强	弱
人力投入	多	少
技术难度	易	难

集群演进-搬迁

- 历程
  - 2019年12月技术方案讨论
  - 2020年03月立项启动
  - 2020年08月联调测试与准备
  - 2020年11月正式启动搬迁
  - 2021年04月完成核心数仓搬迁

未来规划

- Hadoop 3.2.1 改进优化
- 在离线混合部署
- 多机房
- Ozone
- Hadoop与k8s架构融合


SACC
THANKS