云上HBase冷热分离实践

云HBase冷存储方案介绍

郭泽晖(索月) 2018.08

Content

01 典型据场景

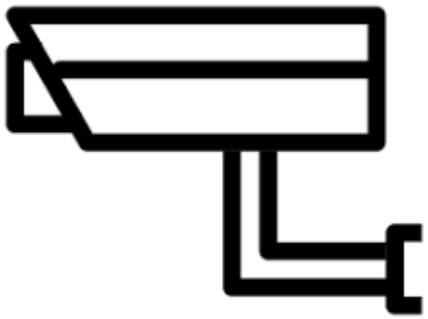
02 传统方案

03 云HBase 云端方案

01 典型据场景

典型场景

- 只有少部分数据频繁访问
- 访问频次随时间流逝而减少
- 归档备份
- ……



定义

热数据

- 频繁访问
- 数据量相对少
- 对延迟敏感

冷数据

- 极少访问(平均每GB数据月度访问不超过10W)
- 数据量大(TB级别)
- 对成本敏感

02 传统方案

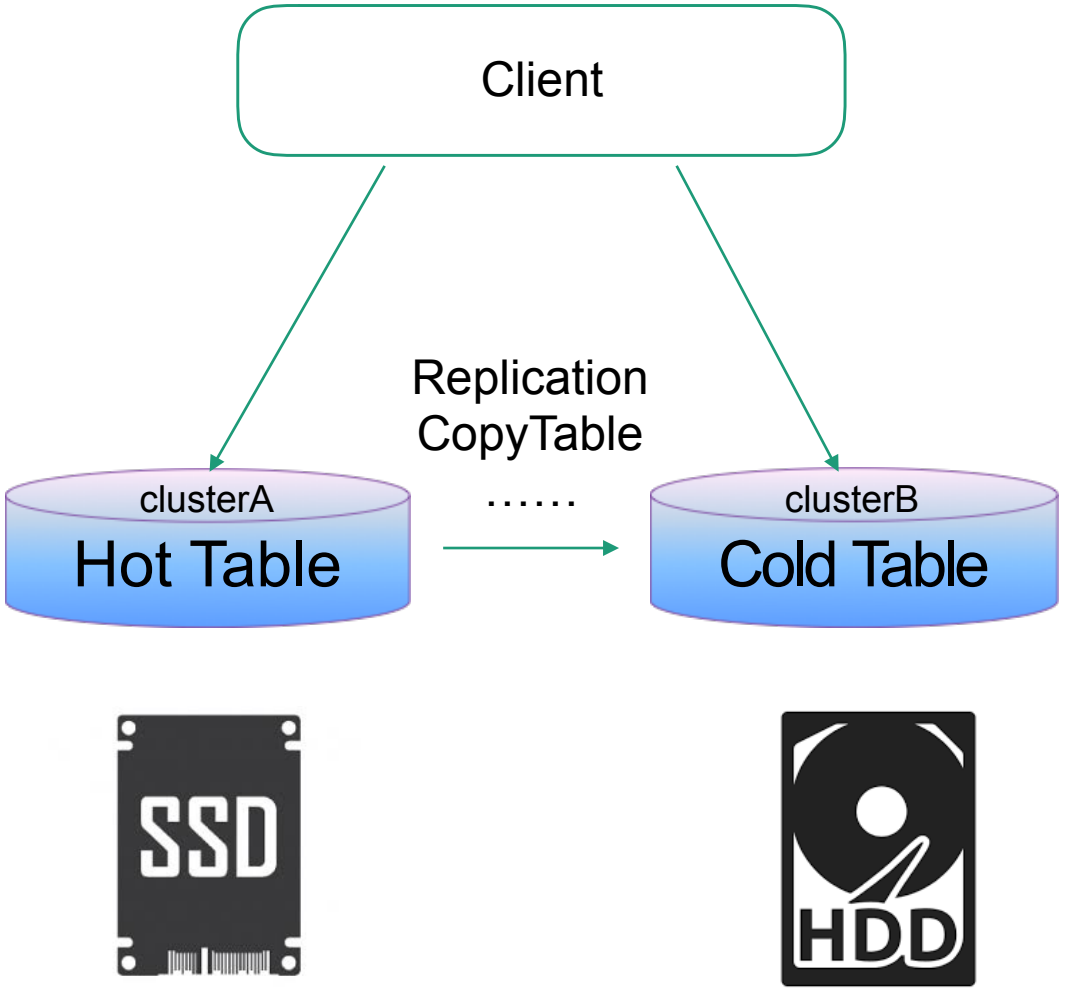
传统方案 1.X

优点

- 简单，无需改动HBase代码

缺点

- 双集群维护开销大
- 冷集群CPU可能存在浪费



2.0版本已经支持同集群冷热，参考: HBASE-14061

传统方案 2.X

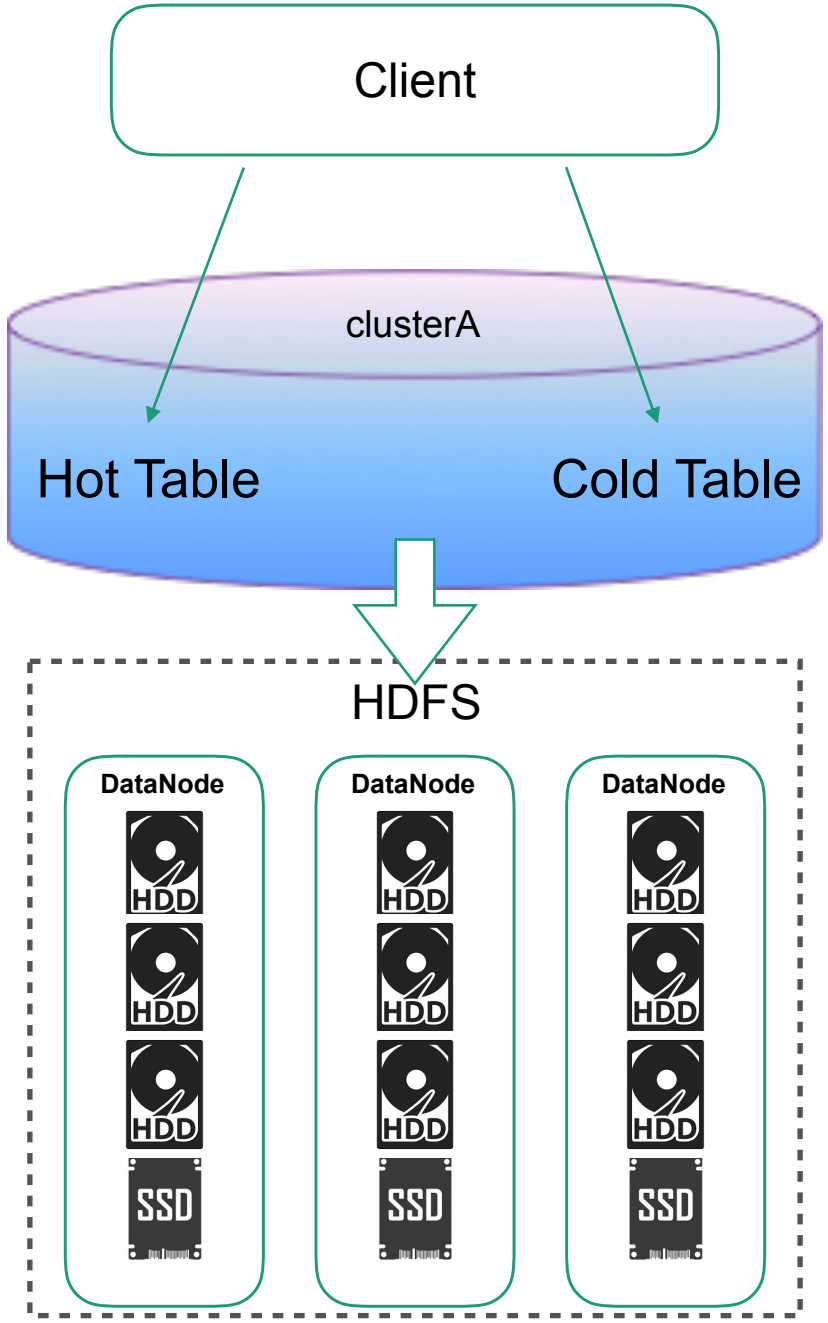
优点

- 同一集群维护开销少

缺点

- 需要根据业务考虑不同介质磁盘配比，业务变动集群配置很难跟着变动

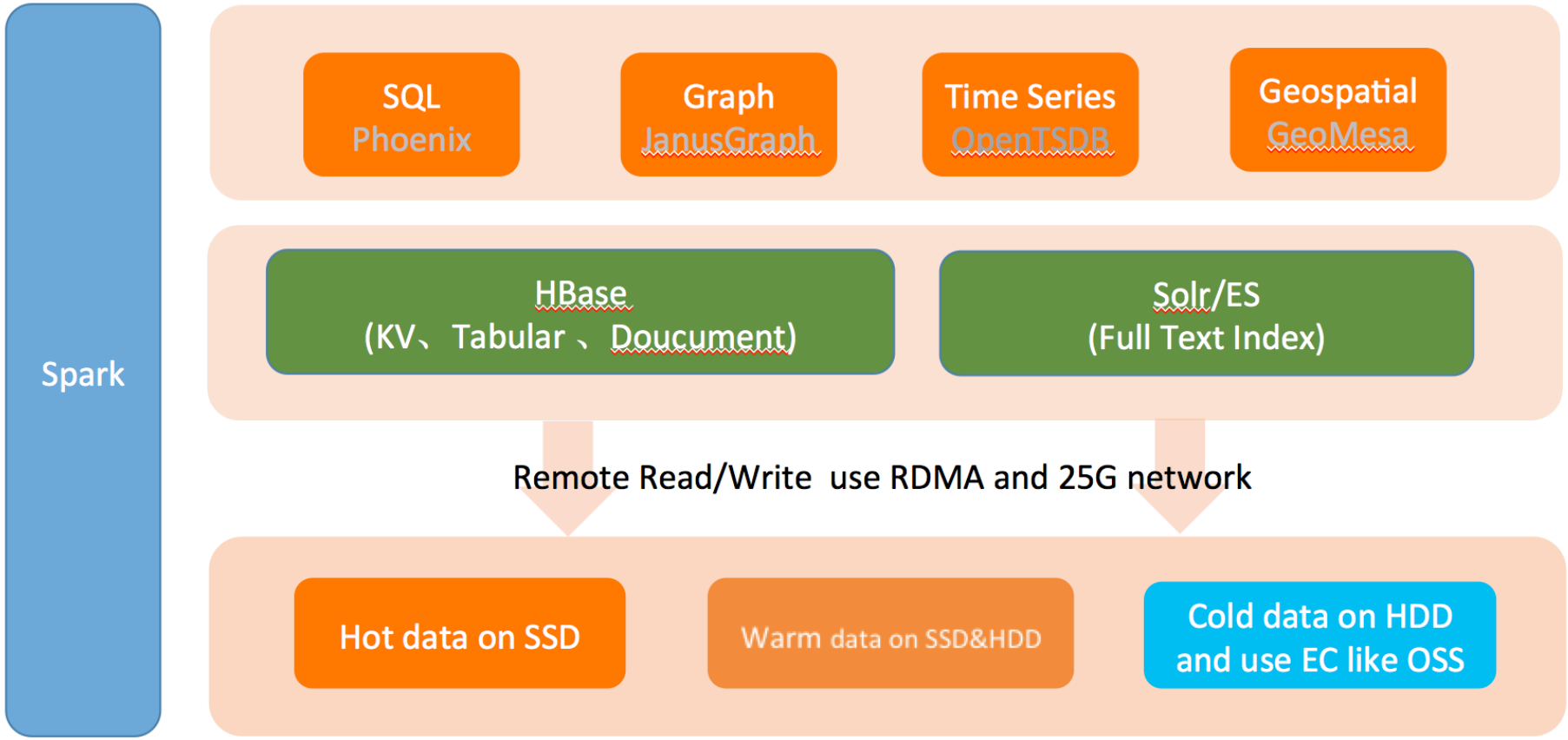
参考：[HDFS ArchivalStorage]



03 云HBase云端方案

云HBase介绍

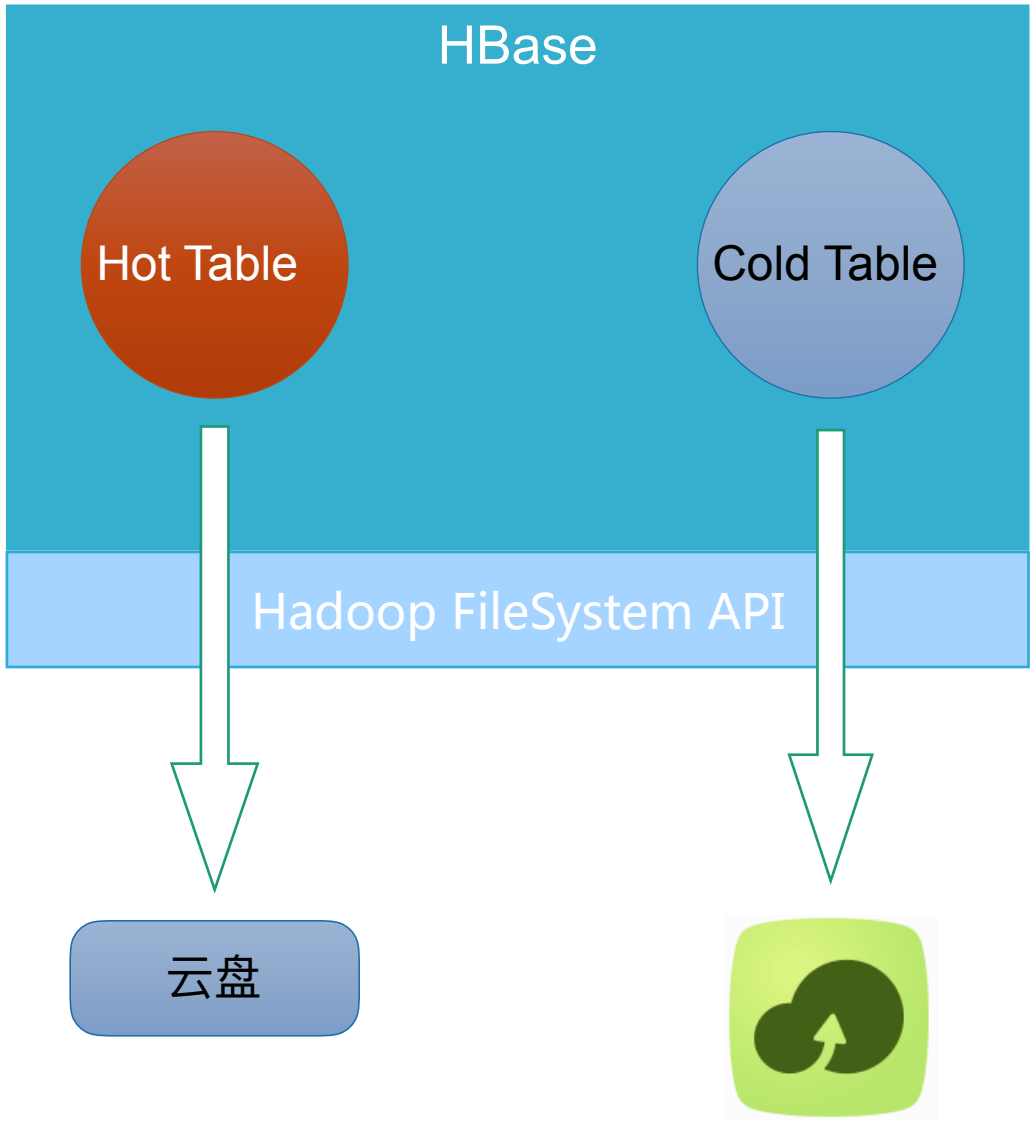
- 存储计算分离
- 完全弹性
- 多模式
- 免运维



欢迎使用云HBase https://www.aliyun.com/product/hbase

基于OSS的HBase冷存储

- 存储弹性伸缩
- 同集群管理方便
- OSS便宜可靠

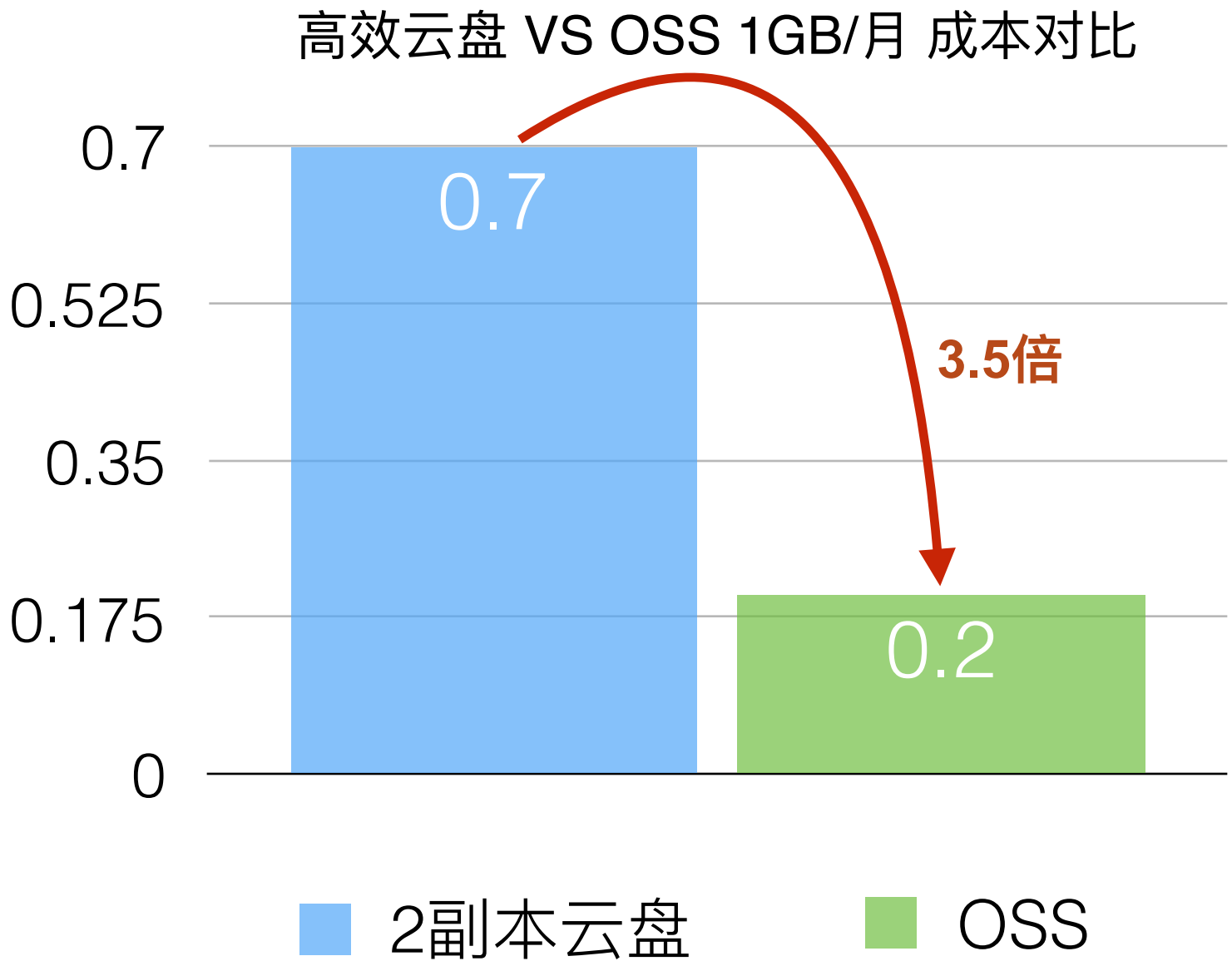


OSS对象存储

- 可以存大对象，TB级
- 数据设计持久性不低于99.999999999%
- 低成本

阿里云对象存储服务（Object Storage Service，简称 OSS），是阿里云提供的海量、安全、低成本、高可靠的云存储服务。

成本优势

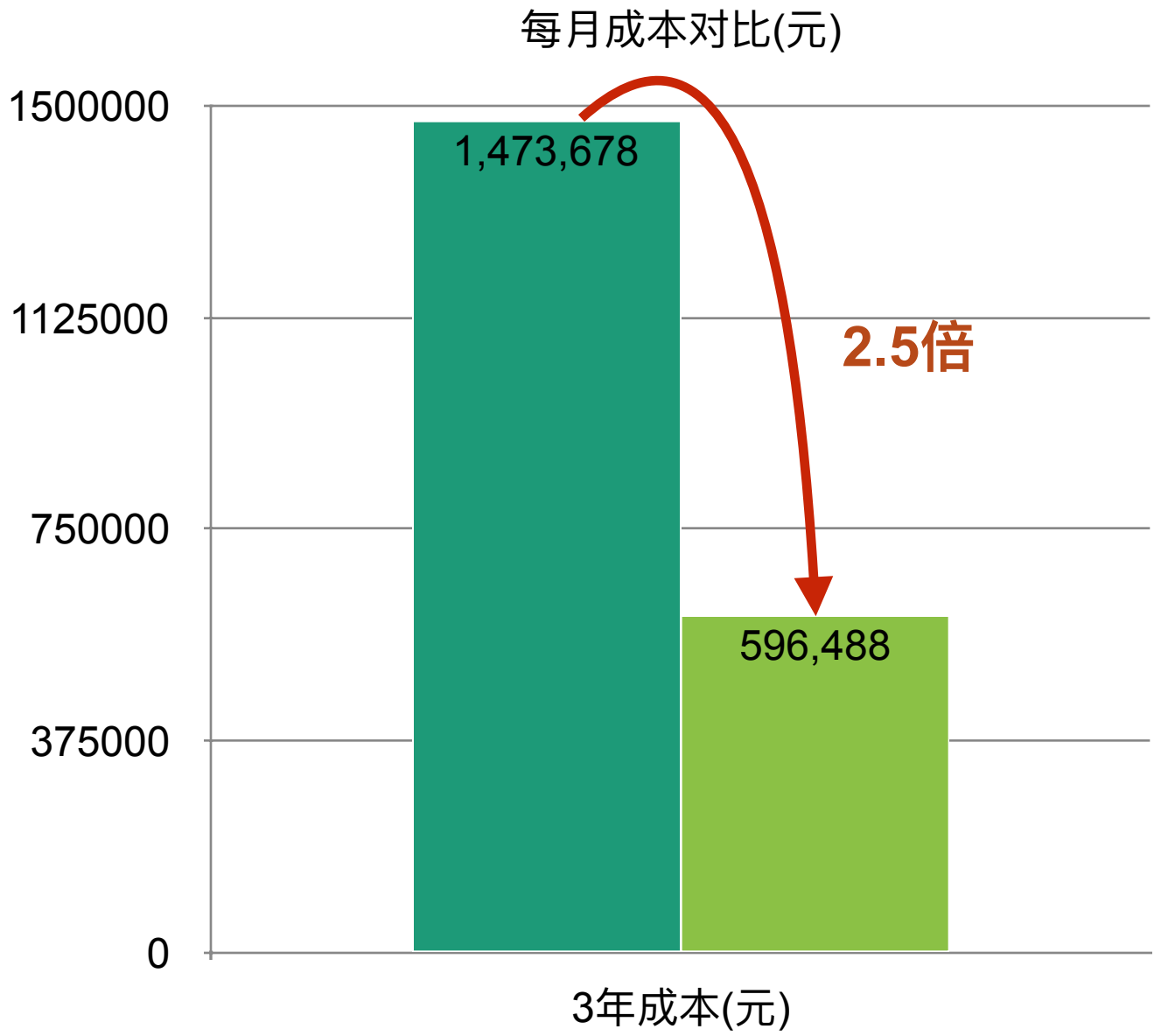


成本优势

举例

某汽车企业，拥有10万量车。每车每30秒上传7K的包，数据半年后基本不访问。

我们以3年的存储量(大约2P)来估算成本。



基于OSS架设HBase的问题

直接基于 **Hadoop社区 NativeOssFileSystem** 架设HBase存在的问题

- 只能通过多个对象来模拟目录结构
- 操作目录/文件实际上是分别操作多个对象
- 类似rename这种操作，如果中途crash，会出现不一致，无法保证原子性
-

模拟文件系统需要4个对象实际上

1个对象，/ 不是目录
是对象名的一部分

/root/parent/son/file

/root

/root/parent

/root/parent/son

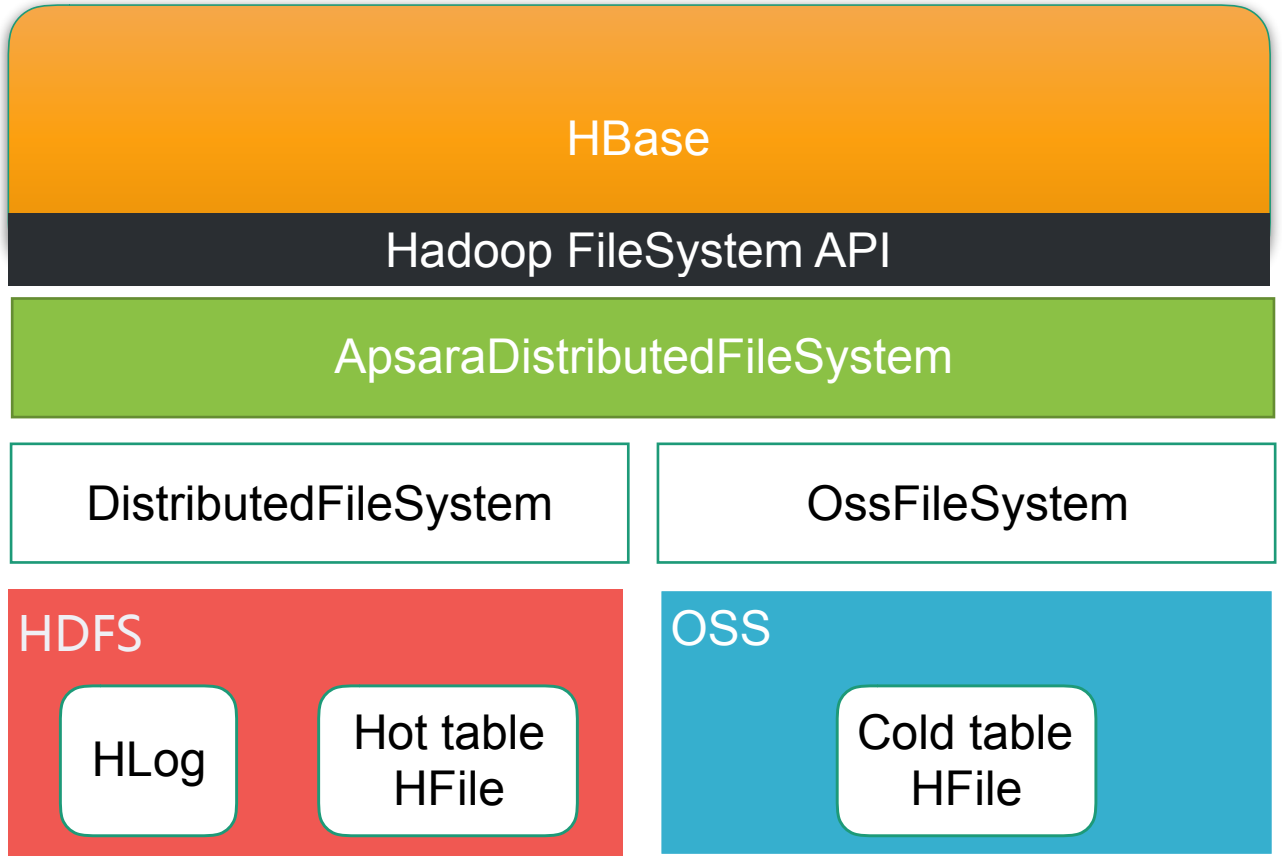
/root/parent/son/file

云HBase冷存架构

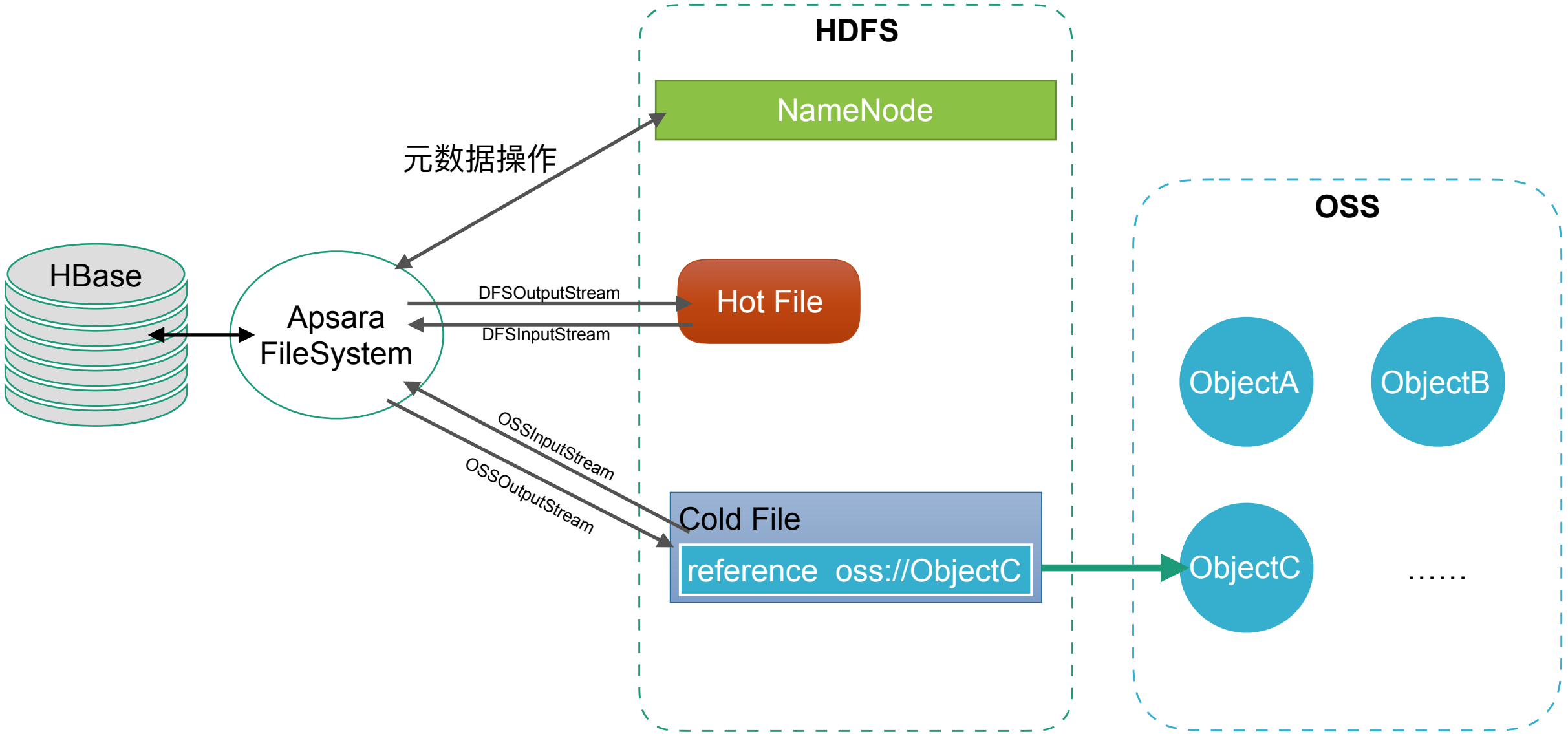
1.HLog依然放HDFS，为了写入性能考虑，适应写多读少场景

2.冷表HFile在OSS

3.热表HFile在基于云盘的HDFS



云HBase冷存架构

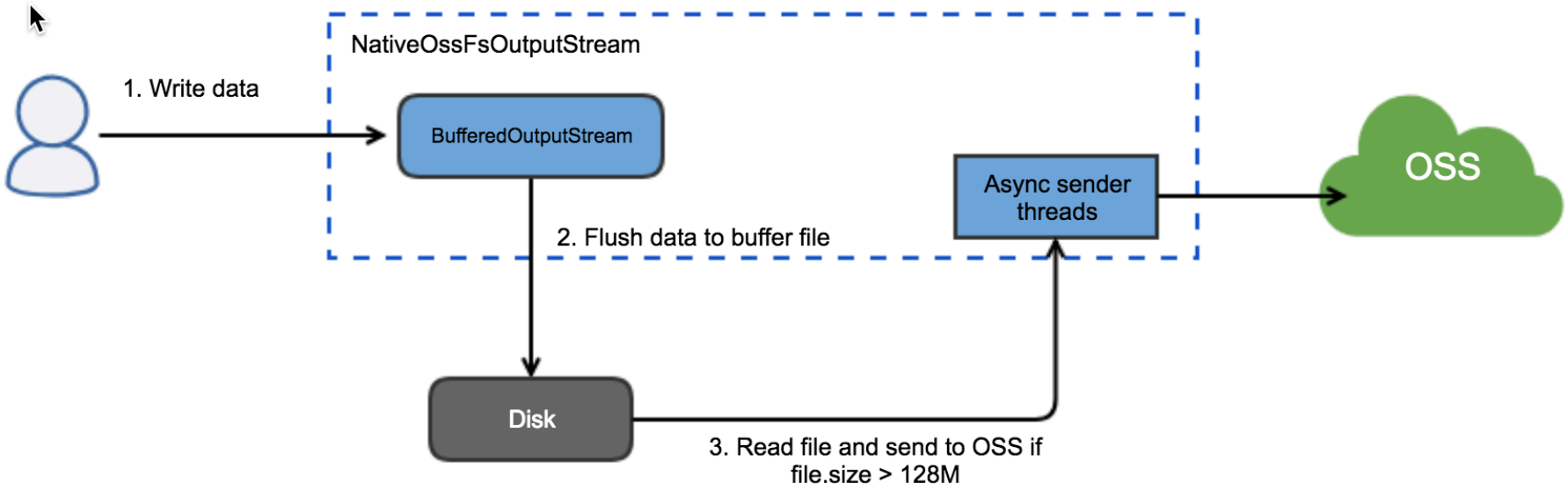

HDFS
NameNode
元数据操作
HBase
Apsara
FileSystem
DFSOutputStream
DFSInputStream
Hot File
OSSInputStream
OSSOutputStream
Cold File
reference oss://ObjectC
OSS
ObjectA
ObjectB
ObjectC
.....

社区版本对比

实现上一些限制

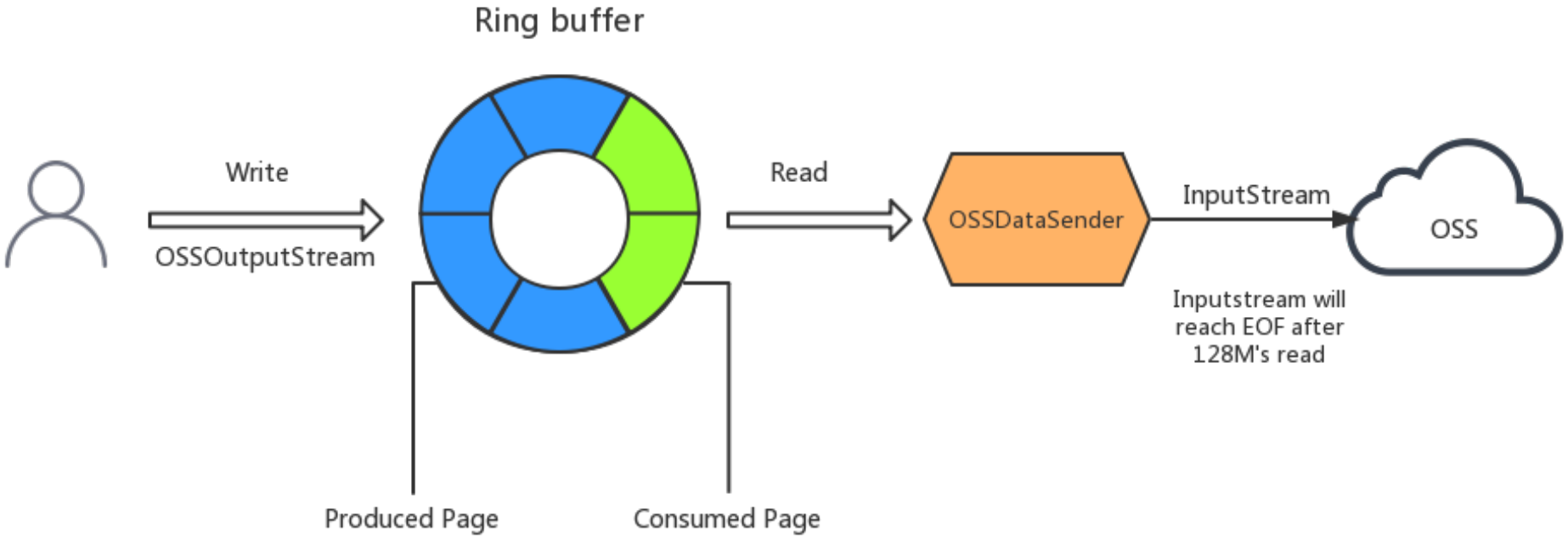
- 请求是要钱的
- Hadopo FileSystem是提供OutputStream让用户输入
- OSS是提供InputStream让用户输入

社区NativeOssFileSystem写入流程



1.数据写入需要落盘
2.依赖磁盘性能，pagecache等因素
3.写入磁盘速度不够快，异步发送线程池实际退化成单线程
4.crash时候可能需要清理磁盘上的残留文件

社区版本对比



云HBase的实现

- RingBuffer只会占据几M内存，分成page维护
- 将生产好的数据page包装成InputStream给OSS
- 当发送超过128M时候截断InputStream提交一次

社区版本对比

测试配置

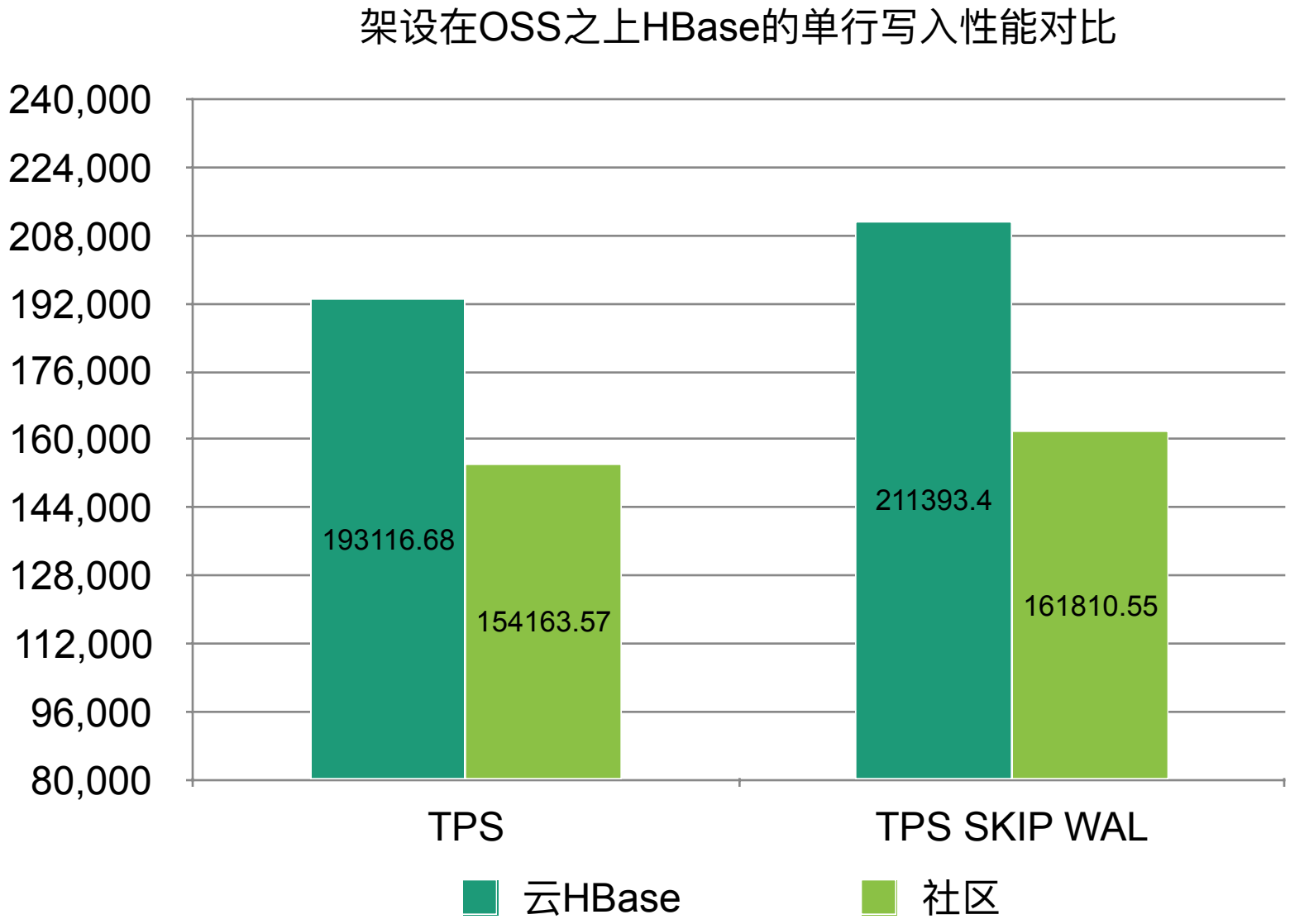
HDFS 6台8核32G DataNode

HBase 1台8核32G RegionServer

HLog 均放在HDFS上

valueSize=100B

threads=120



测试命令参考：hbase pe --nomapred --valueSize=100 --rows=1000000 --table=test --presplit=64 randomWrite 120

使用方式

非常简单，仅需一行命令！

```
create 'test', {NAME => 'info'}, CONFIGURATION => {'HFILE_STORAGE_POLICY'=>'COLD'}
```

冷存储建议使用场景

- 写多读少
- 顺序读

使用限制

1.如果持续Get，冷存储读IOPS会有限制(30左右)

2.偶尔访问，IOPS限制会适当动态放宽

2.顺序读写流量不限制

冷存与云盘性能对比

测试配置

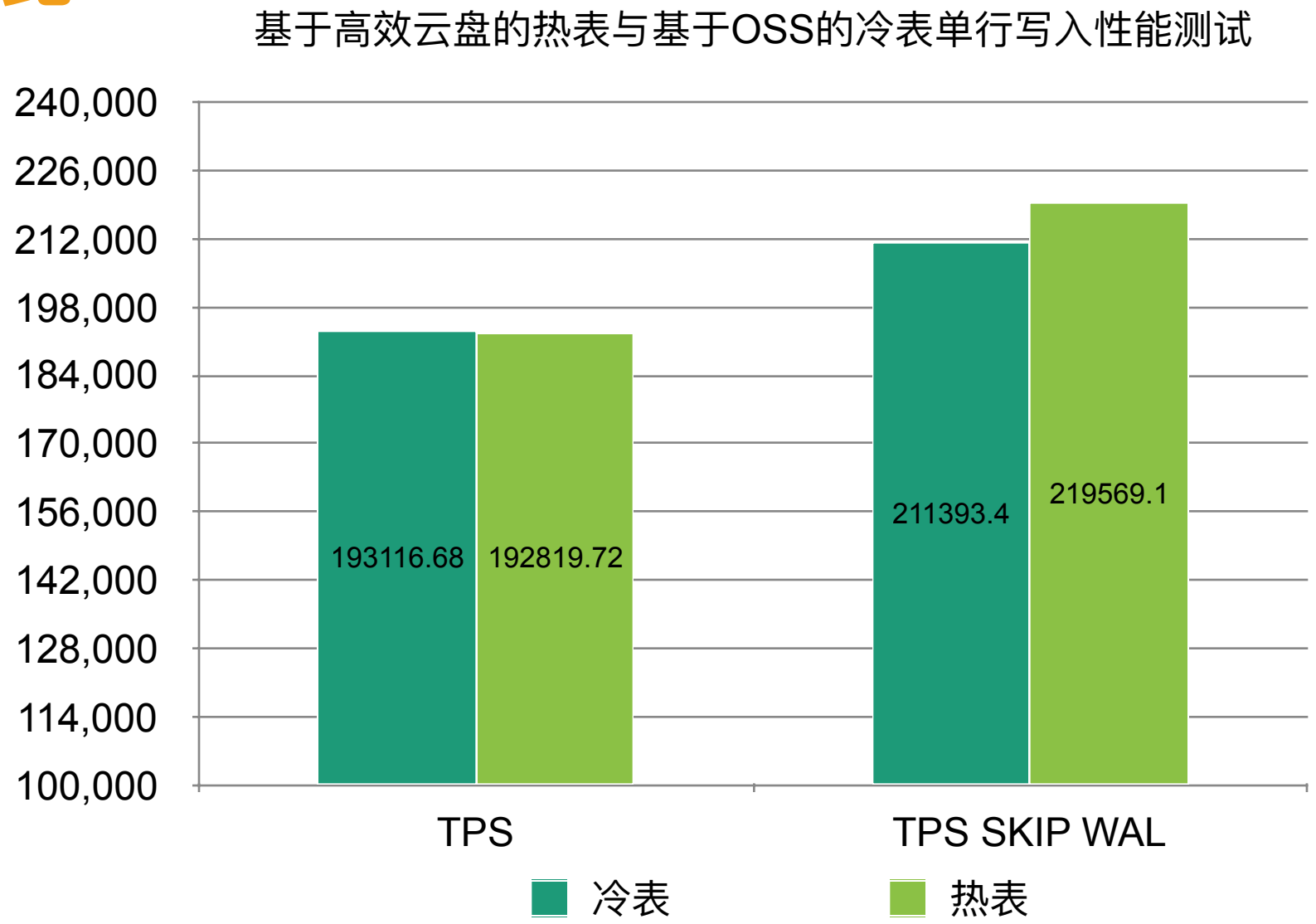
HDFS 6台8核32G DataNode

HBase 1台8核32G RegionServer

每台ECS挂载4块300G高效云盘

valueSize=100B

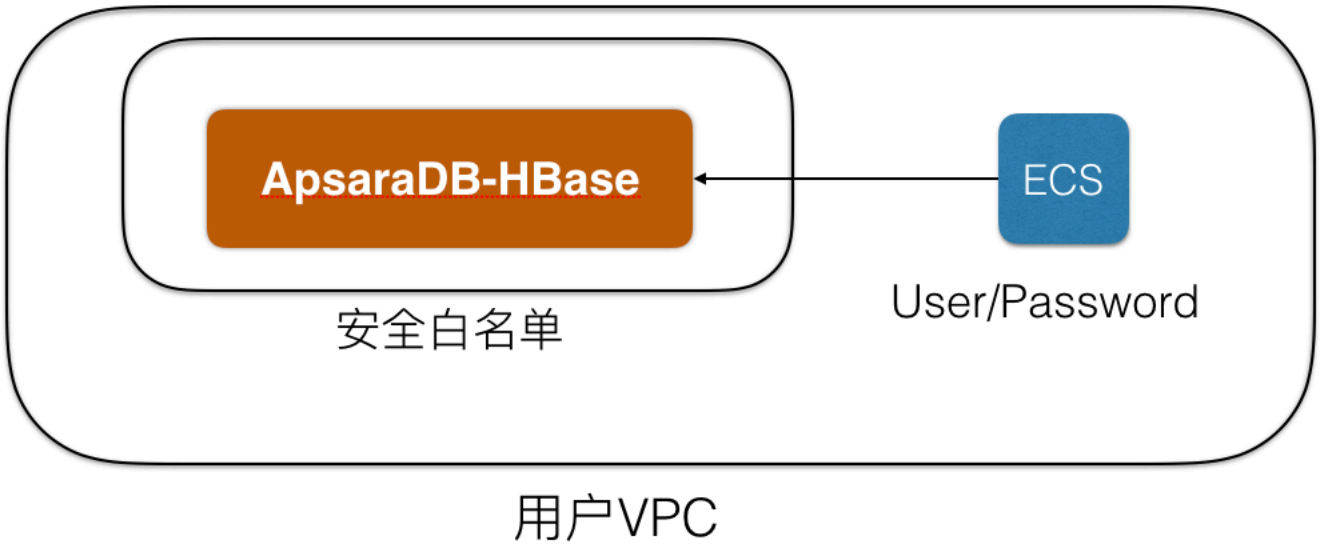
threads=120



测试命令参考：hbase pe --nomapred --valueSize=100 --rows=1000000 --table=test --presplit=64 randomWrite 120

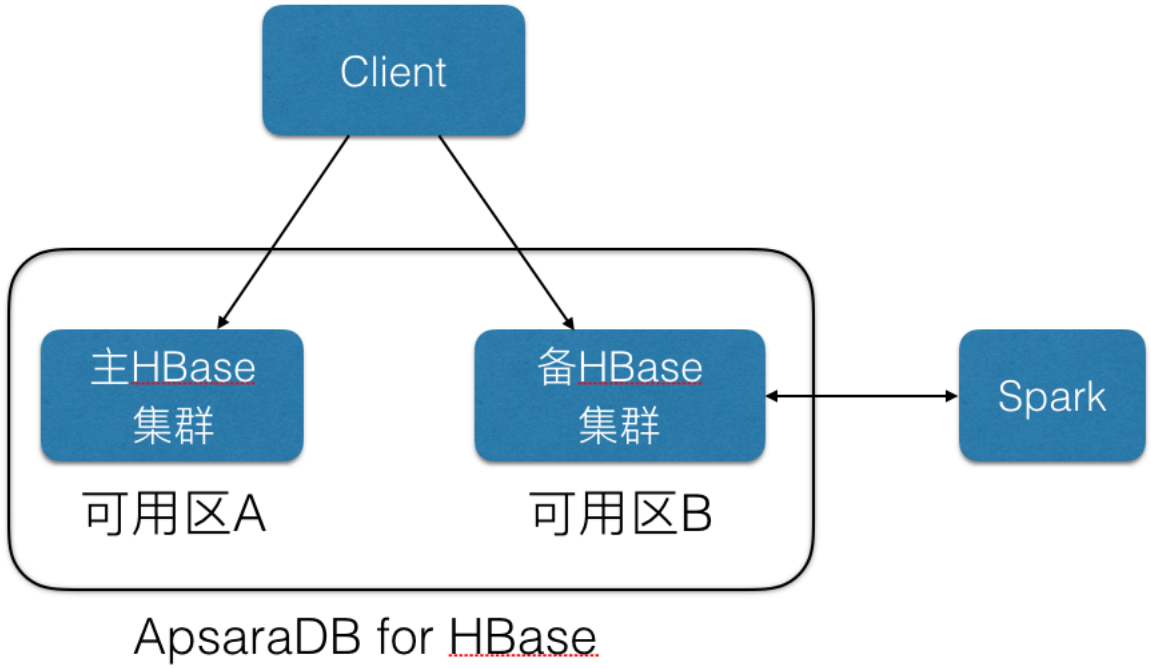
其他特性：企业级安全

- 安全白名单：机器防火墙
- VPC：网络隔离
- 认证：
  - User/Password、授权ACL
  - 跟MySQL体验一致
- 授权
  - 可以授权到表及列族



其他特性：双活

- 满足 集群级别的容灾
- 备集群 可以满足 分析的需求
- 异步同步，最终一致性
- 延迟 200ms以内



云HBase VS 自建HBase

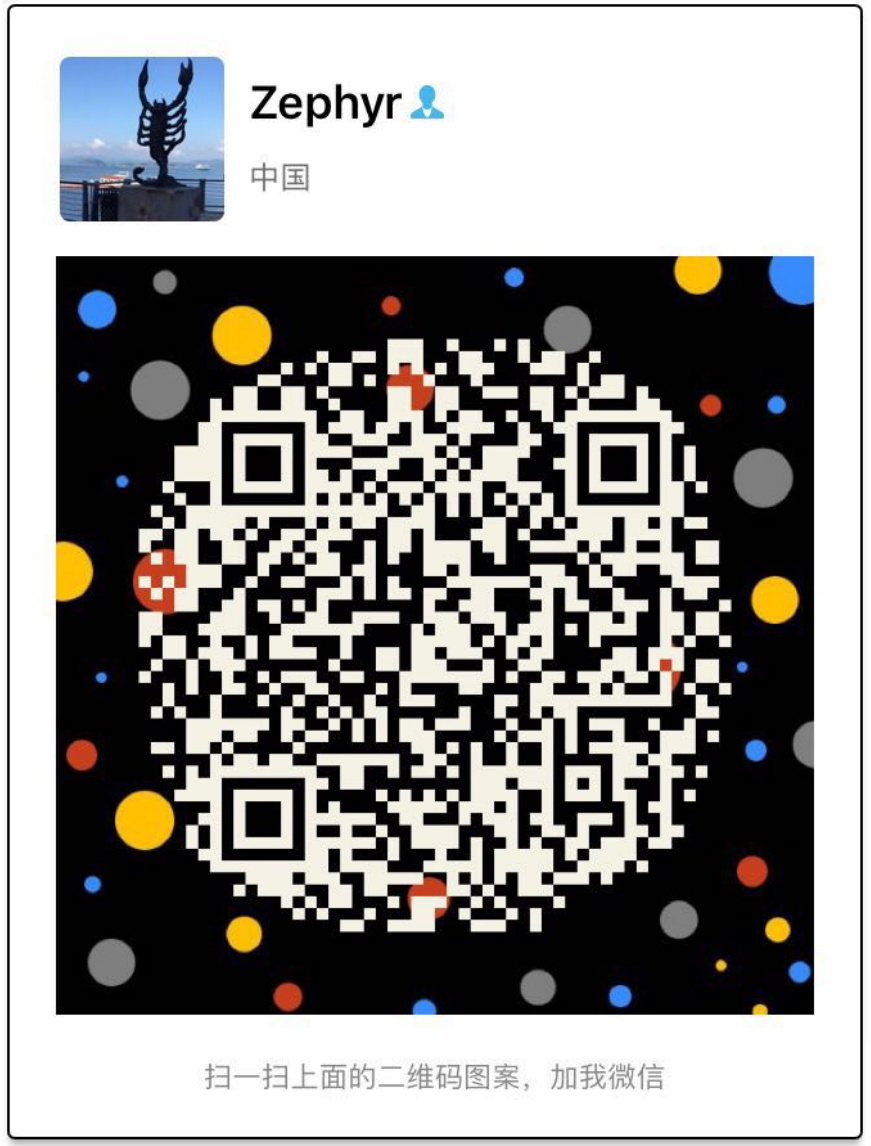
	Item	ApsaraDB HBase (ALiyun Product) https://cn.aliyun.com/product/hbase	Apache HBase (Sofeware)
Basic	High availability	99.9% ~ 99.99%	N/A
	Data reliability	99.999999999%	N/A
Online Ability	Multi-master clustering	Multi-master clustering， Multi-AZ/Regon	NO
	GC	FGC NO， YGC 5ms	GC 20s~100s， YGC 100ms+
Reduce Cost	Storage Cost	Cut by 50%+ on share cloud disk， Total 3 Copy	Maybe on Cloud Disk， Total 9 Copy
	Support Cold Storage	Support OSS， Cut by 70% at less read	NO
Multi-model DB	Multi-model DB	KV， Tabular， SQL， Graph， Time Series， Geospatial Full Text index， Search	KV， Tabular
Enterprise Characteristics	Disaster recovery	Backup and Restore	NO， maybe3.0
	Security	user/password， ACL	Kerberos， ACL
	Analytics	Spark on HBase ， More optimization	Spark on HBase
	Version upgrade	Automatic upgrade	N/A
Self-driven	Database control system	15min Create a DB/Monitor Online add storage and node/Elastic Power in future	N/A
	Diagnostic System	Big request ， Big Table merge， Hot Region ……	NO

We are hiring!

- 如果你对Hadoop生态系统，或者任何NO-SQL数据库感兴趣
- 如果你对建立云上大数据生态感兴趣
- 如果你对挑战高并发，低延迟感兴趣

HBase官方中文社区：http://hbase.group/

云HBase： https://www.aliyun.com/product/hbase



为了无法计算的价值 | 阿里云