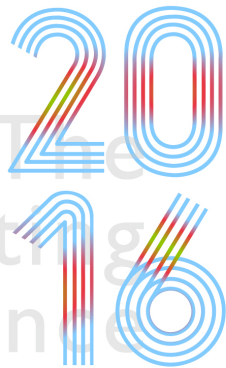





2016 杭州·云栖大会
THE COMPUTING CONFERENCE

云栖社区
yq.aliyun.com

海量数据分布式存储 ——Apache HDFS之最新进展



主办单位:  杭州

 Alibaba Group
阿里巴巴集团

战略合作伙伴: 

郑锴

Intel 研发经理, Hadoop committer



扫码观看大会视频

目录

content

- 大数据发展趋势
- HDFS 存储演化
 - HDFS 缓存 (Cache)
 - HDFS 分层存储 (HSM)
 - HDFS 纠删码 (EC)
- HDFS在未来
 - 智能存储管理 (SSM)
 - 对象存储
 - 存储在云端





大数据发展趋势



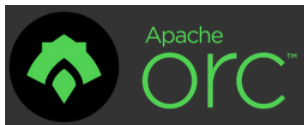
要存储和处理的数据量越来越庞大

- 物联网的发展使得接入设备越来越多
- 实时流处理技术的发展使数据导入速度越来越快
- 数据分析（OLAP）日趋成熟
- 人工智能（AI）新时代，人们希望聚集更多的历史数据进行深度学习



对处理数据速度的期望越来越高

- 能处理大量数据只是基本要求，还要处理的快
- 新数据从产生到及时被处理，催生各种实时流处理框架
- 顺序读取已不能满足要求，各种存储格式改进和跨越式读取



存储的场景更复杂，更丰富

- 一个集群，同时支持好：
 - 大文件、小文件
 - 热数据、冷数据
 - 在线处理、离线分析
- 对象存储



存储设备的两极：越来越廉价和越来越快

- 要么更廉价，更多更老的数据促进更廉价的设备，tape死而复生？
- 要么更快，SSD步伐越来越快
- 3XD Point技术和NVM设备，存储和内存统一起来



网络带宽也越来越高

- 10Gb的网络已经是标配
- 40Gb乃至100Gb也即将到来



存储和计算相分离，大数据加速向云端迁移

- 云计算，大势所趋
- 弹性计算，更灵活，可伸缩
- 跨集群、跨数据中心，远程读取不可避免

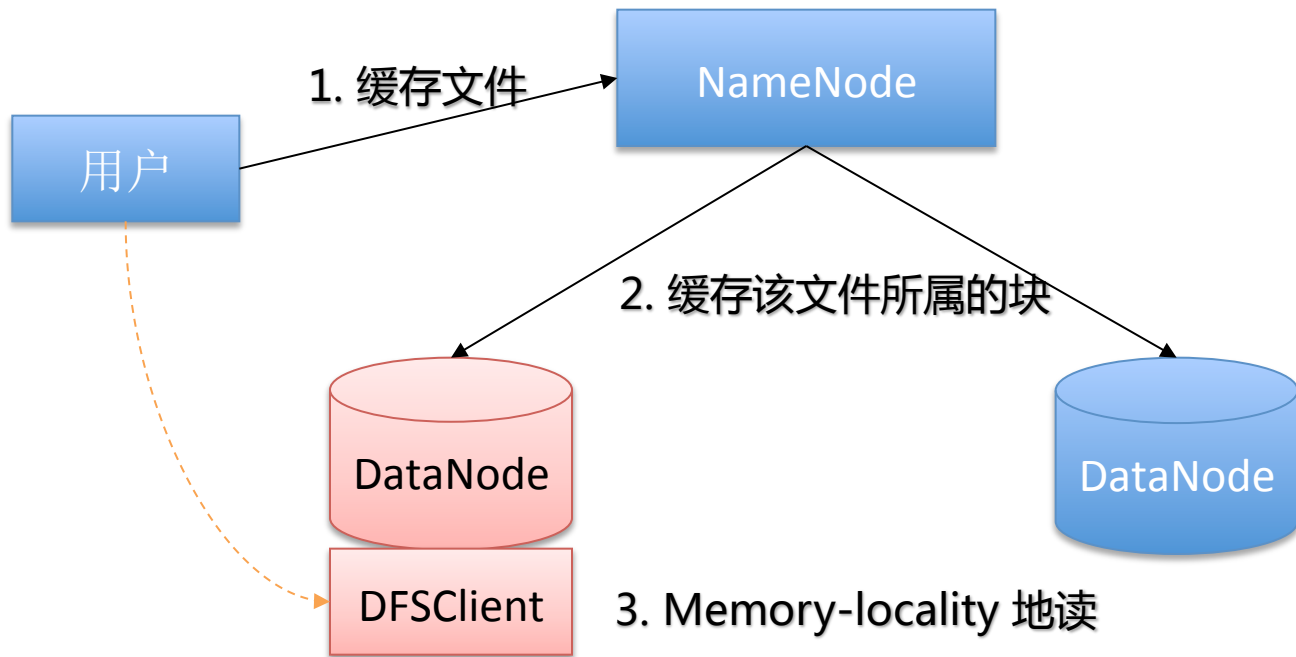




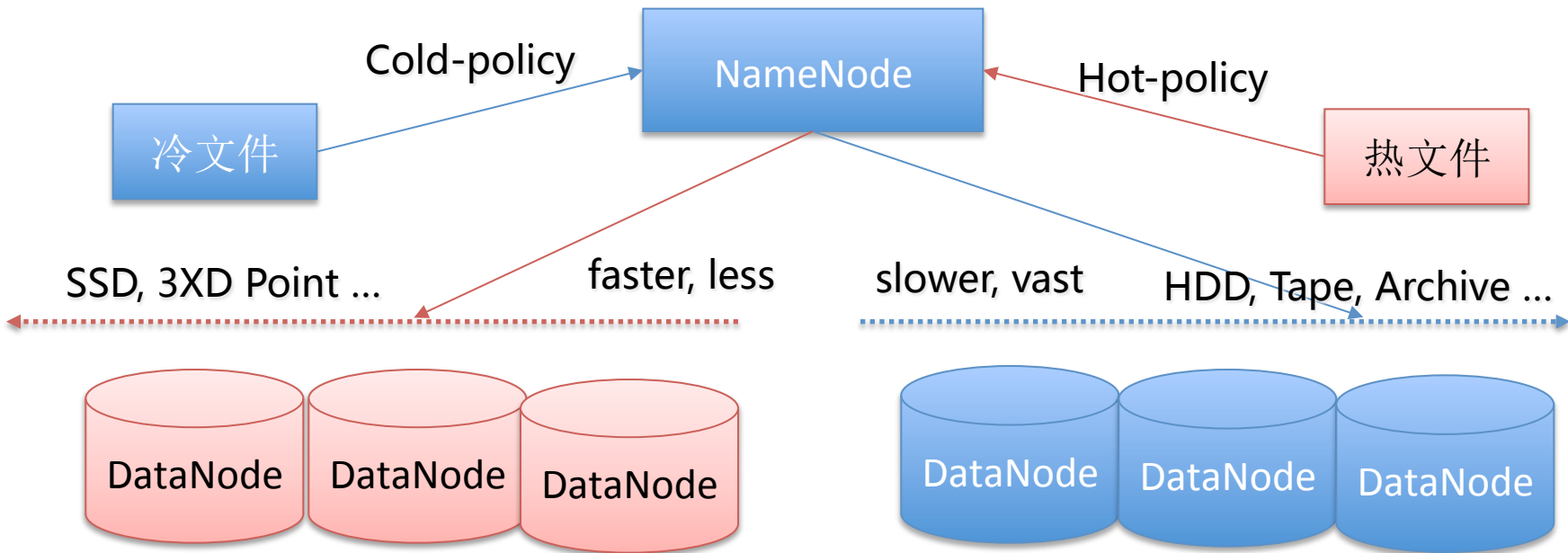
HDFS 存储演化



HDFS Cache 缓存支持



HDFS HSM 多层次存储体系

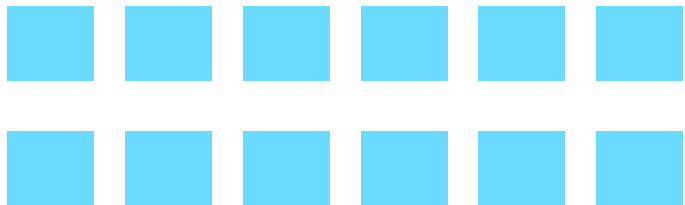


HDFS EC 纠删码支持 (1)

一个例子



额外存储



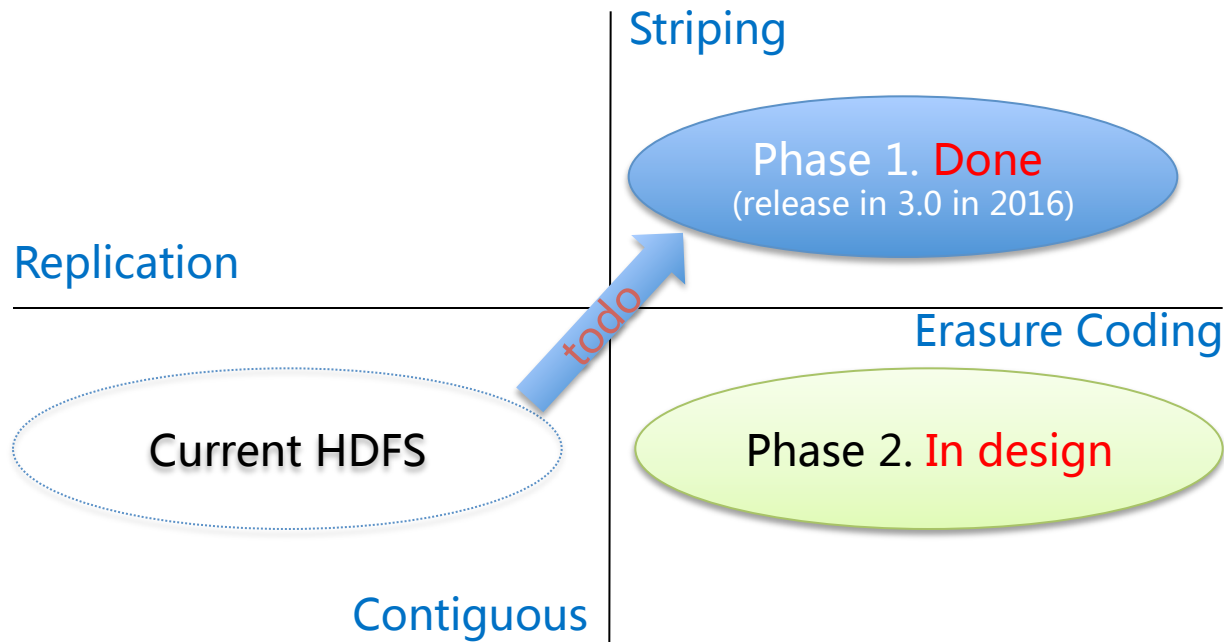
HDFS 3 备份模式保存2个副本
需要12个冗余块

额外存储开销
200% vs 50%



HDFS EC默认使用RS(6,3)
仅需3个冗余块

HDFS EC 纠删码支持 (2)

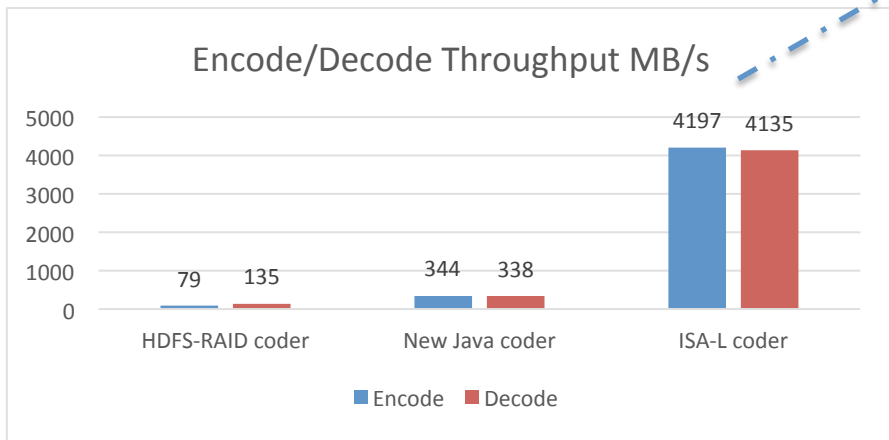


- Hadoop 3.0 with EC:
 1. Alpha1 released
 2. Alpha2 on going
 3. GA in the year

HDFS EC 纠删码支持 (3)

硬件加速的ISA-L 编解码器，性能>10X

- 开源 (<https://github.com/01org/isa-l>)
- 通过利用硬件的高级指令集 (如SSE , AVX , AVX2) 来实现EC编解码的优化
- 同时支持Linux和Windows平台

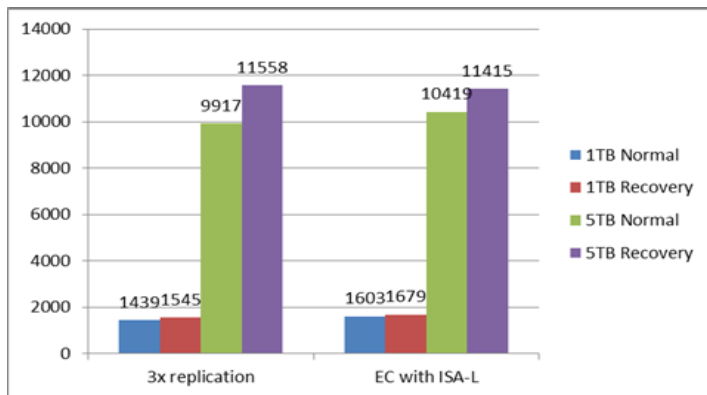


ISA-L Native Coder:

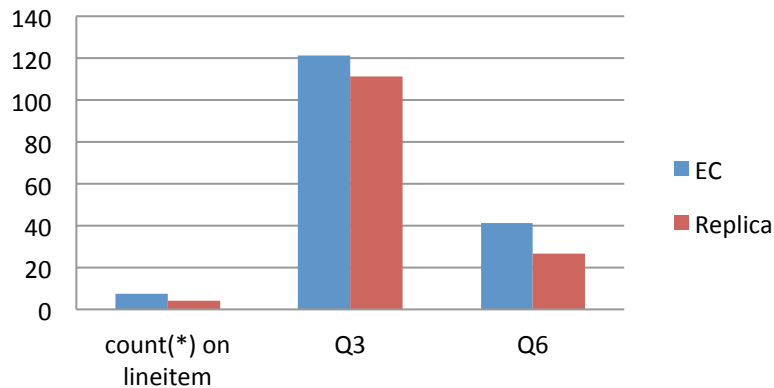
- ✓ Direct ByteBuffer, 数据零拷贝
- ✓ 核心数据一次初始化，高速缓存

HDFS EC 纠删码支持 (4)

TeraSort Execution Time (s)



TPC-H 500 GB Query Time (s)



Note: 2 DataNodes Killed

HDFS 在未来

HDFS 在未来

智能存储管理 (SSM) (Smart Storage Management)

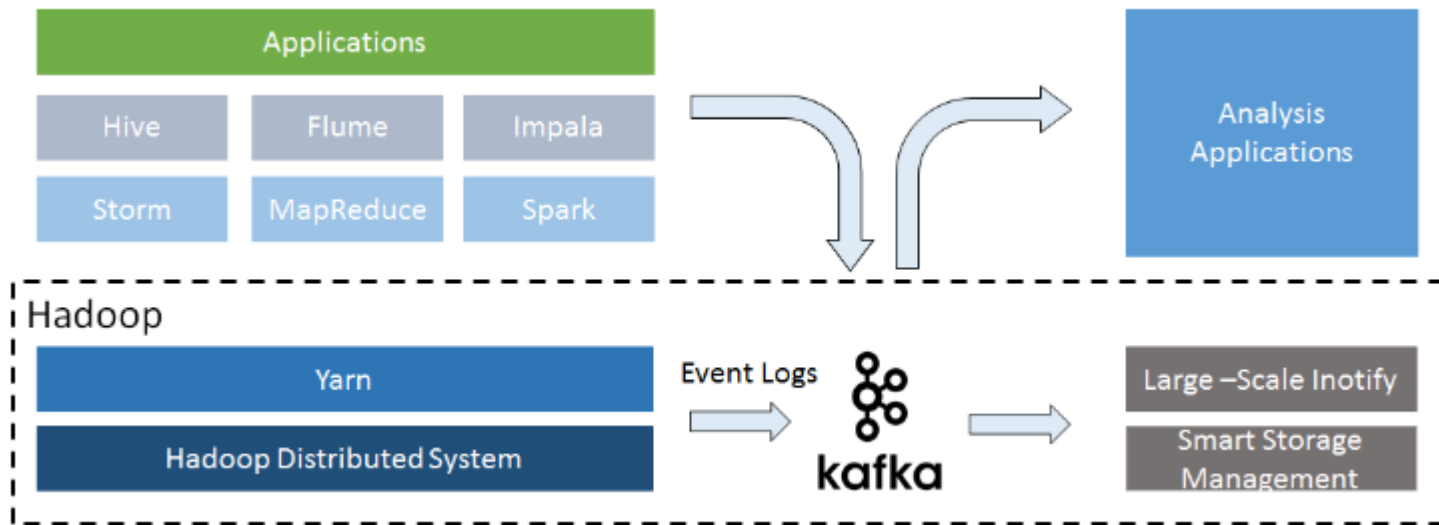
目前面临的问题和挑战，在支持了那么多非常好的特性之后

- 如何选择合适的文件形态？
 - 对于Replica，如何及时调整合适的备份数？
 - 对于EC，如何选择合适的schema？
- 如何预测数据的读取趋势，提前将非常热的数据cache起来
- 如何及时感知数据的温度
 - 将经常读的数据转入到SSD
 - 将变冷的数据移入到廉价设备
- 如何评估存储设备的存储和读取效率？

HDFS Smart Storage Management (SSM)

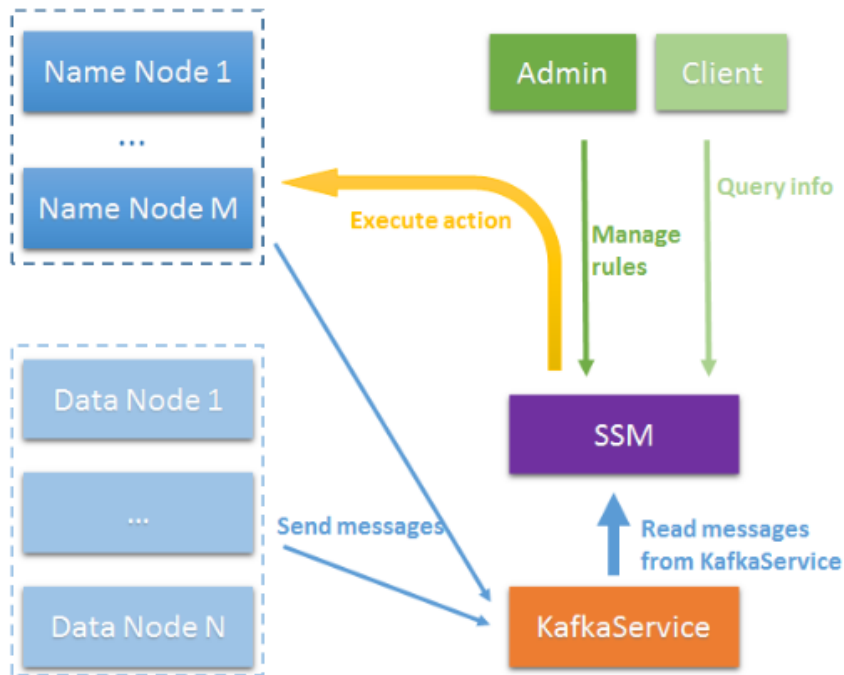
- 端到端的全面的智能存储解决方案
- 完整地收集集群的存储和数据访问统计
- 简化地、智能地和全面地及时感知集群存储状态变化并作出存储策略调整
- [HDFS-7343](#) : 正在开发当中, 欢迎反馈和参与!

将Kafka引入到 Hadoop , 作为基础服务 (KafkaService)

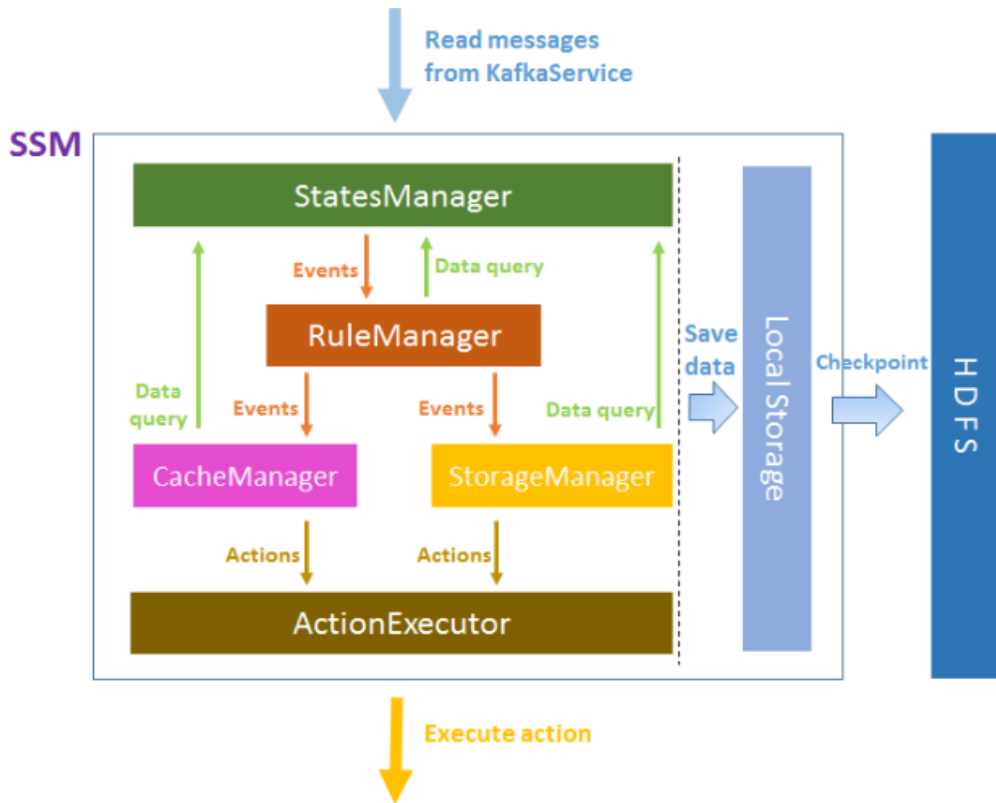


[HADOOP-13633](#), 欢迎反馈和参与 !

系统架构



系统设计



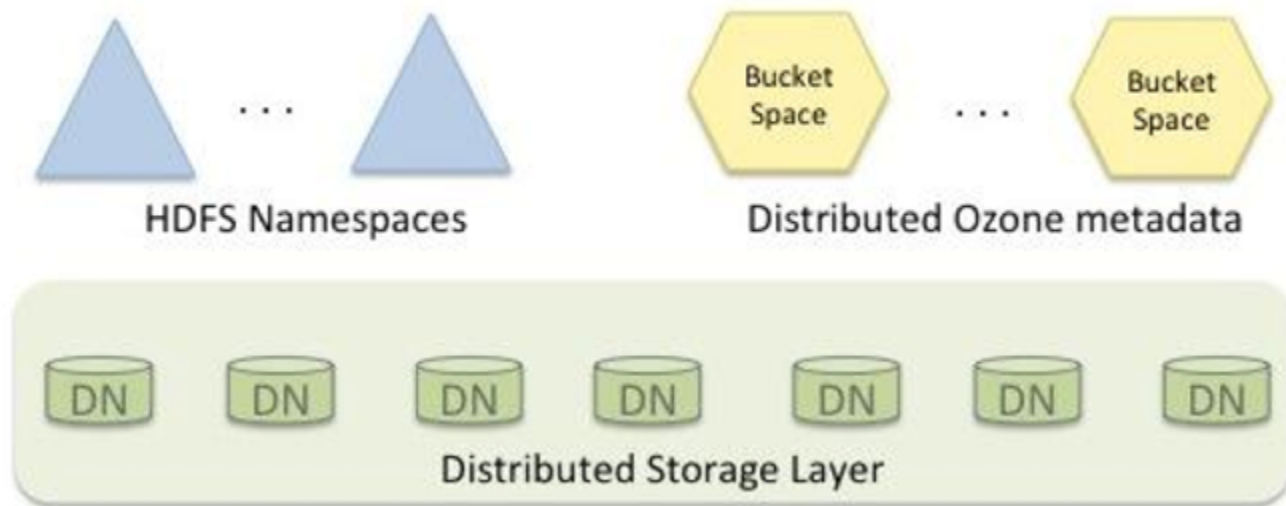
HDFS 在未来

对象存储 (Object Store)

Object store in HDFS (1) [HDFS-7240](#)

- Hadoop正在演化成为一个更为通用的平台，甚至支持传统的服务和应用
- 对象 (Object) 更为轻量，没有file metadata 和 ACL，基于K/V的API
在一些场景下更为友好
- 支持对象存储是一个流行的趋势：S3，Azure，Aliyun，.....
- 目标：
 - ❑ 支持数以亿计的数据对象
 - ❑ 支持任意大小的对象，从几K到几十MB
 - ❑ 保证一致性、可靠性和可用性

Object store in HDFS (2)



HDFS 在未来

存储在云端

统一的Hadoop文件系统和API



Hadoop兼容文件系统抽象层: 统一的存储API接口 `hadoop fs -ls s3a://job/`



快速弹性的HDFS缓存层 [HDFS-9806](#)

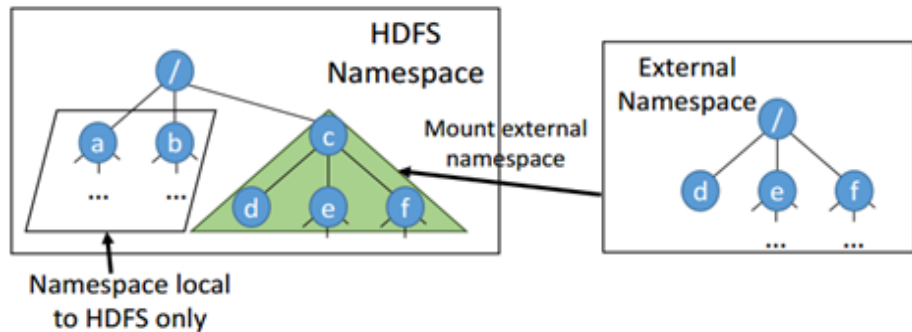


Figure 1: Loading an external namespace as part of HDFS namespace.

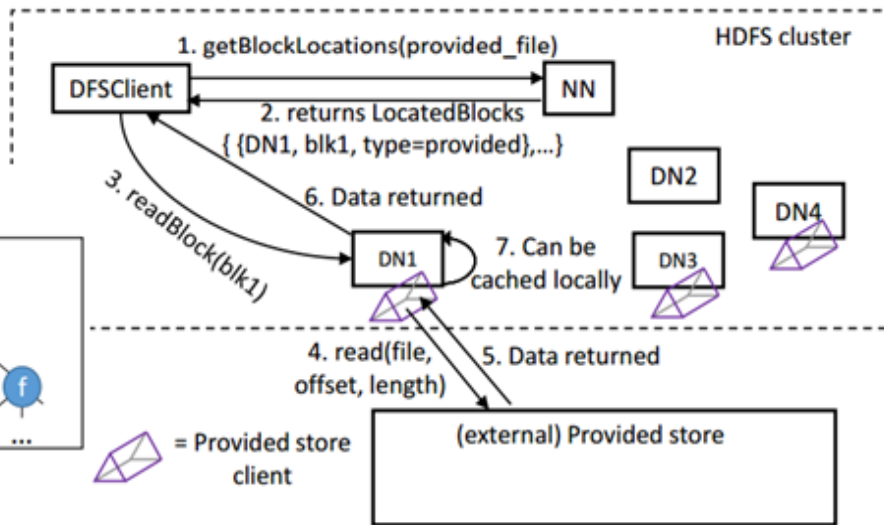


Figure 2: DFSClnt reading a PROVIDED HDFS block.

Reference

- HDFS-EC perf blog: [Progress Report: Bringing Erasure Coding to Apache Hadoop](#)
- Zhang Zhe, Chen Xiao: [Apache Hadoop十周岁：展望前方 讲话](#)
- OZONE: [AN OBJECT STORE IN HDFS](#)

2016 The
Computing
Conference
THANKS

欢迎交流和协作
kai.zheng@intel.com