TRANSWARP
星 环 科 技

星环信息科技（上海）有限公司

领航大数据与人工智能基础软件新纪元

星环信息科技（上海）有限公司
www.transwarp.io
2019/5/01

TRANSWARP
星 环 科 技

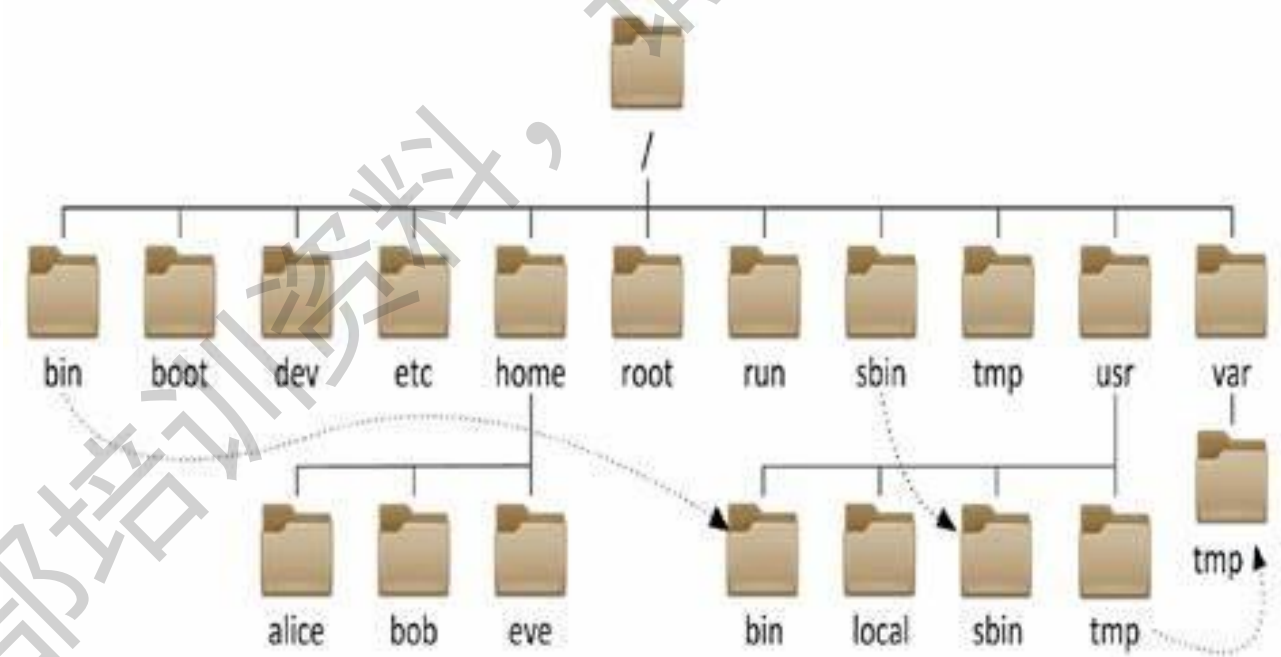
content
目 录

/01 简介

简介 - 文件系统

文件系统 - 操作系统提供的存储介质访问接口

- 名字空间
- 元数据
- API接口
- 安全模型
- ext4 / xfs /ntfs



简介 - 分布式文件系统

分布式文件系统 - 在分布式系统环境中提供类似的访问接口

- 透明访问
- 存储量大
- 高吞吐量
- 提高数据的可靠性
- 高可用
- 高效运维
- 低成本
- 安全性

- 磁盘错误
- 网络故障
- 电源故障
- 数据错误，出现在磁盘，网络，内存
- 系统与软件错误
- 运维的误操作

简介 - 基本概念

- 文件模型
  结构化与非结构化
  文件访问/操作方式
  文件存储/传输模式

- 元数据与namespace
  集群节点信息
  文件自身的元数据，文件名、大小、创建/修改时间、用户与组信息等
  文件内容存储位置关系

/02 常见文件系统

常见分布式文件系统

- NFS

  C/S架构的文件共享协议，用户可以透明访问远程共享服务器上的文件。

- GFS / HDFS

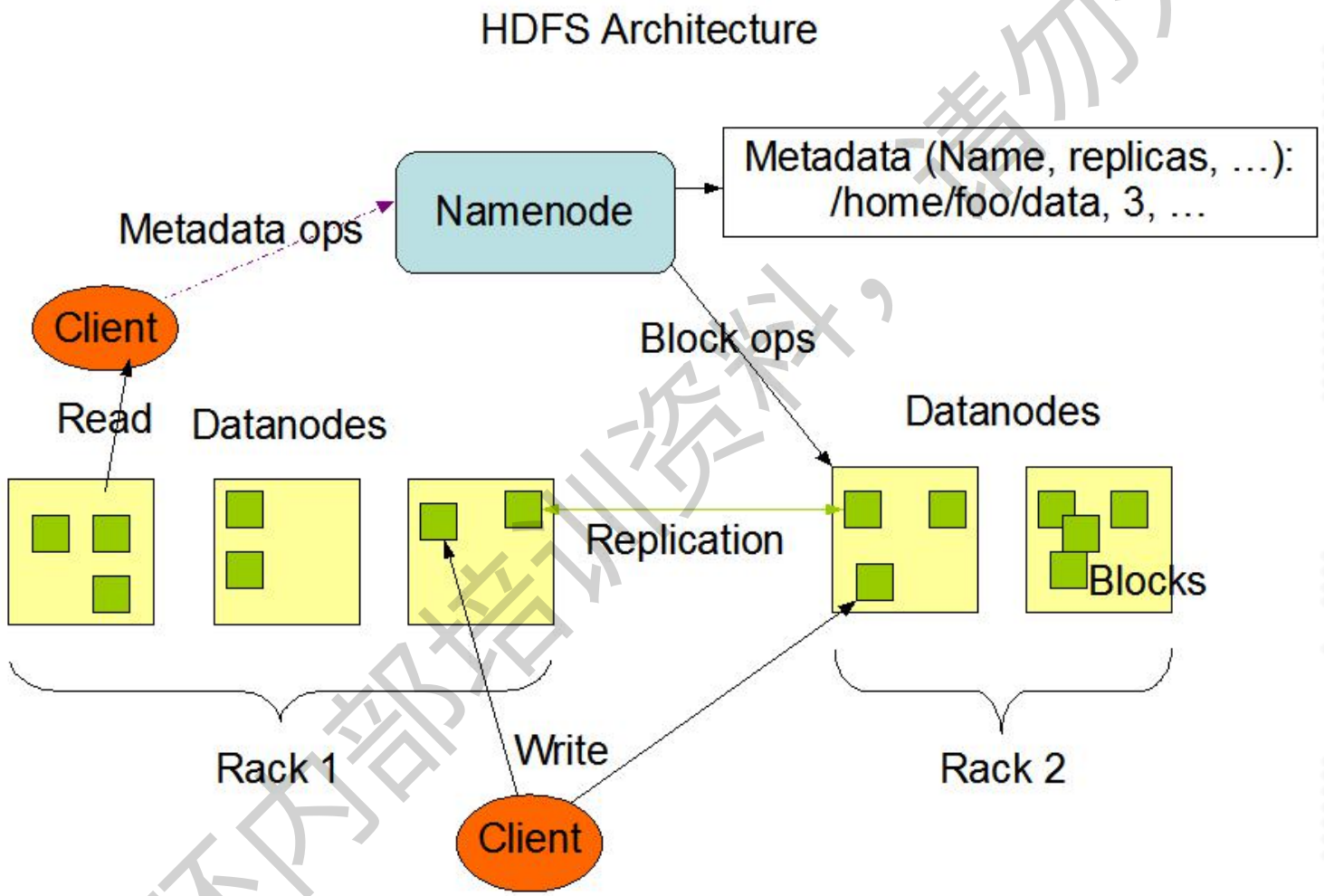
  大数据领域存储引擎，适合大文件，一次写入多次读取，流失数据访问的场景。

- Ceph

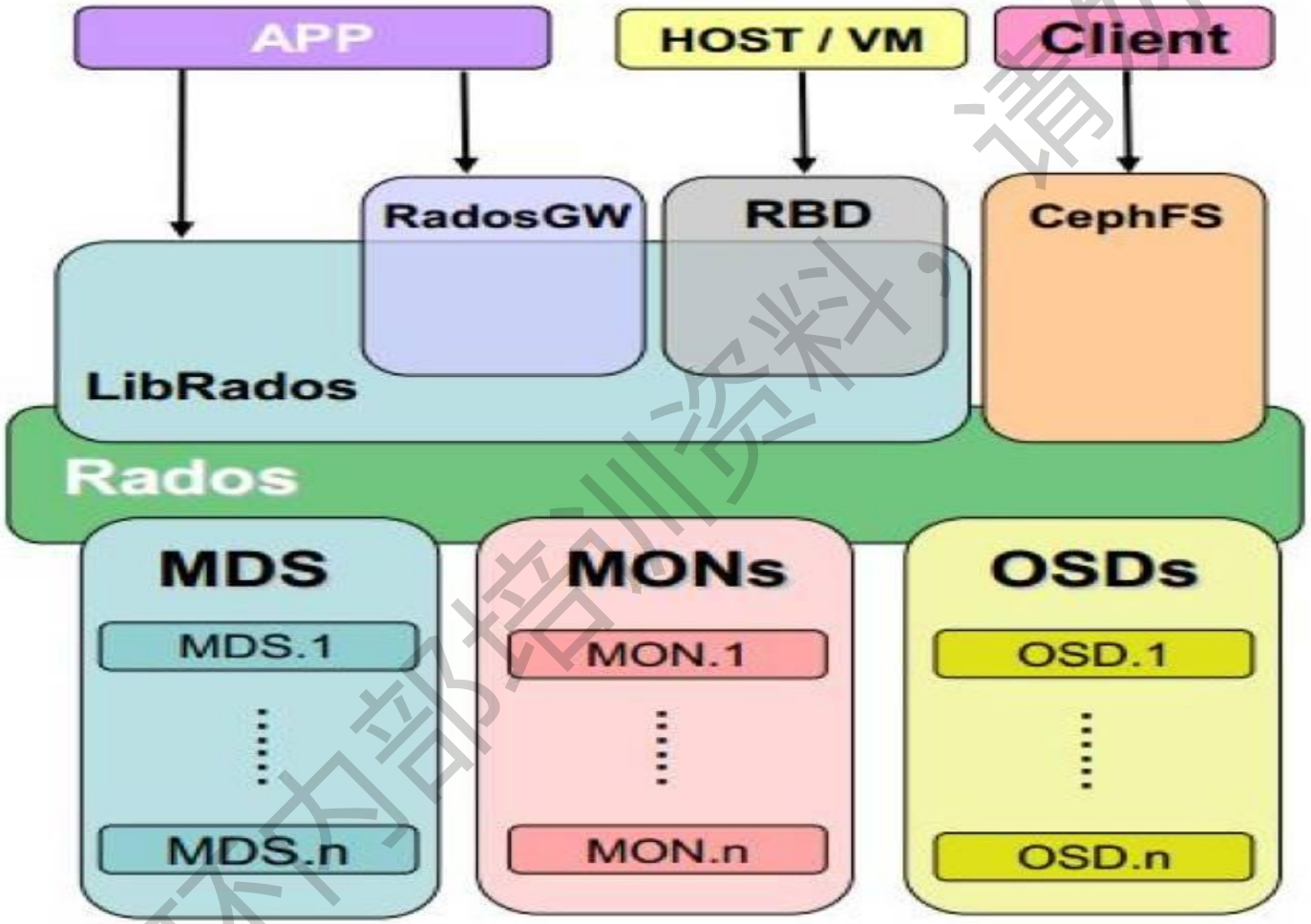
  通用分布式存储系统，通过cephfs提供文件存储，兼容POSIX协议。

- GlusterFS / MooseFS / MogileFS / FastDFS

常见分布式文件系统 - HDFS



常见分布式文件系统 - CEPH



/03 设计要点

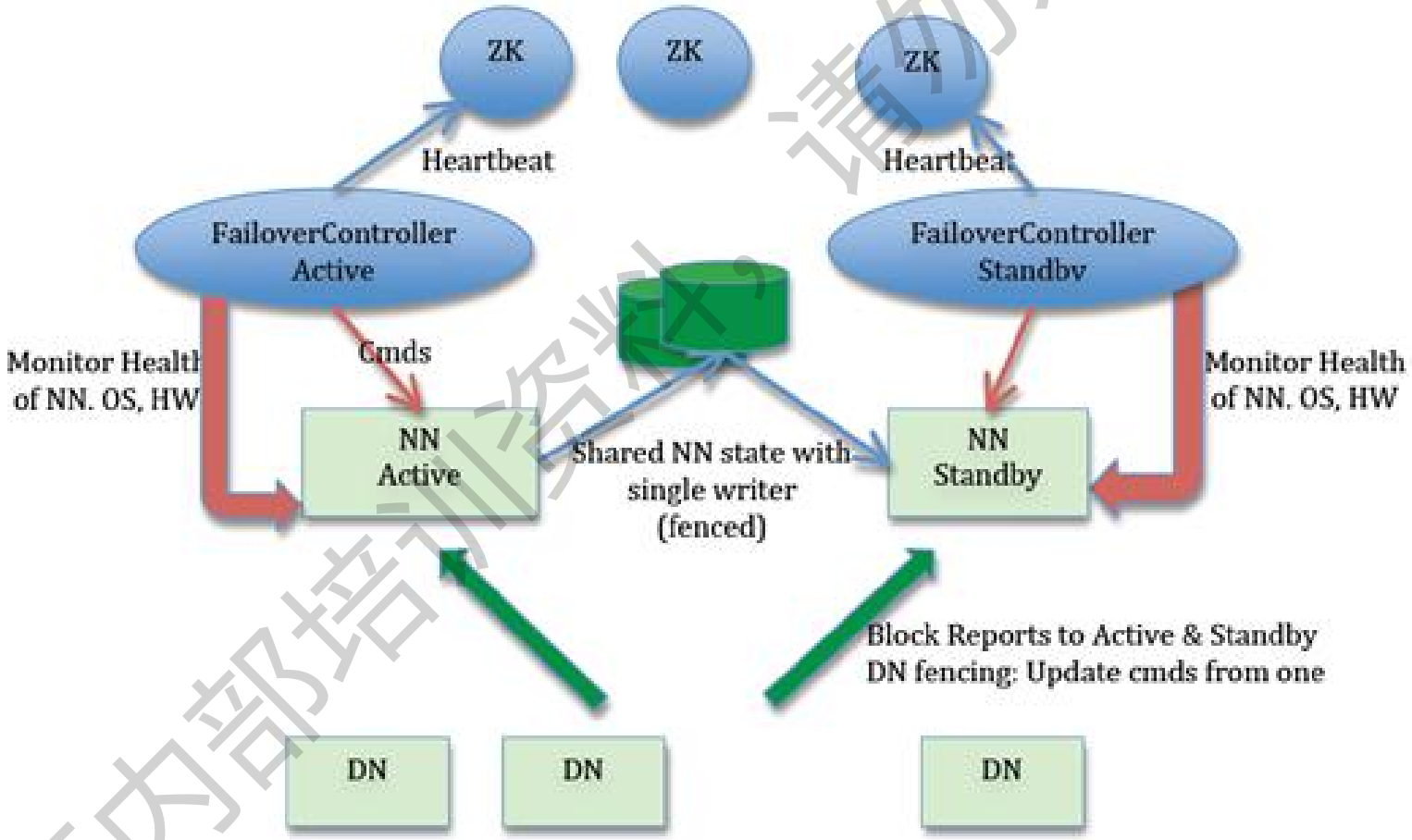
设计要点

- 如何保证数据不丢失
- 如何保证数据正确
- 如何保证数据高可用
- 如何保证可扩展性
- 读写IO路径

- 性能
- 数据复制与平衡
- 安全
- 系统状态监控与告警

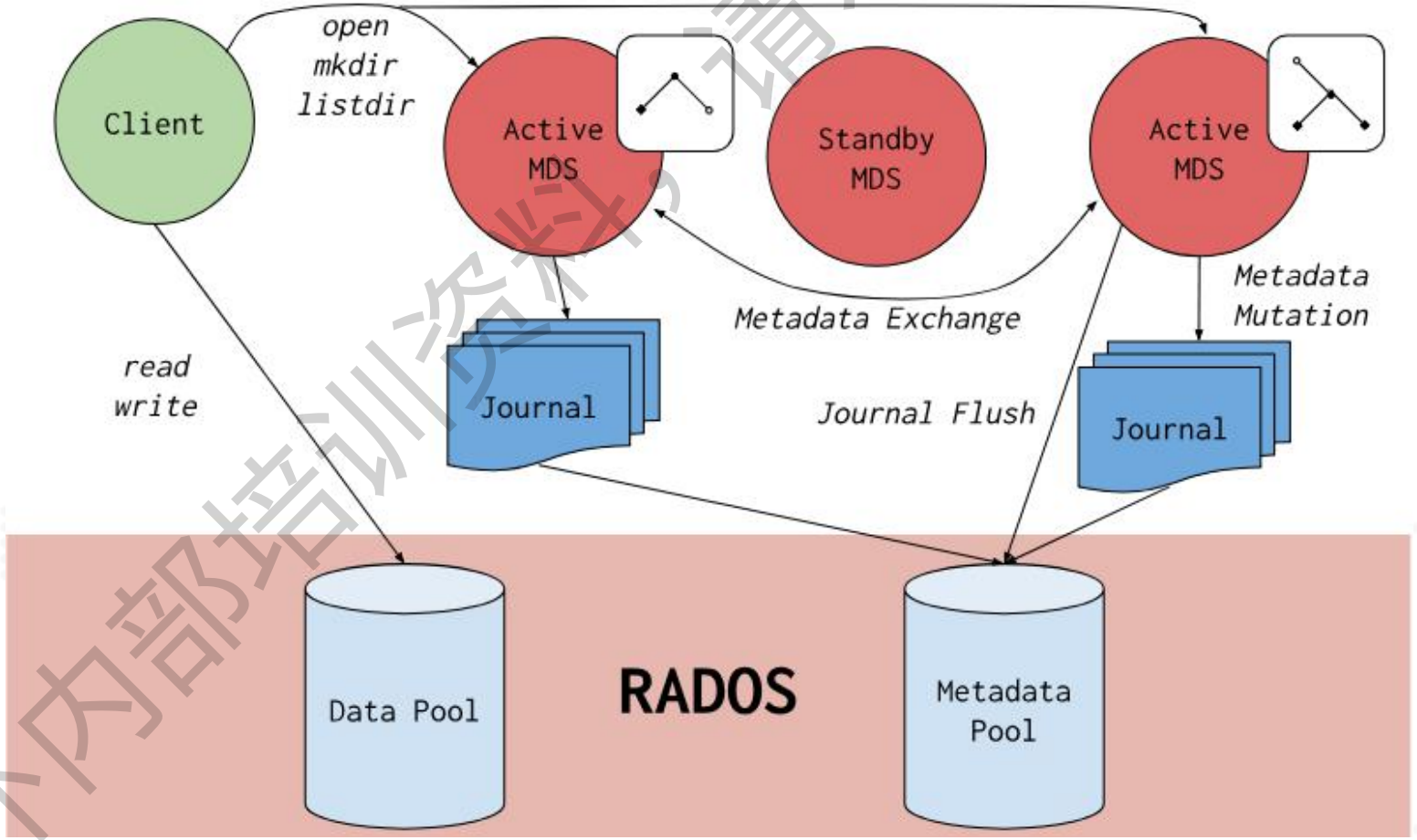
设计要点-元数据服务的高可用

- HDFS HA with QJM



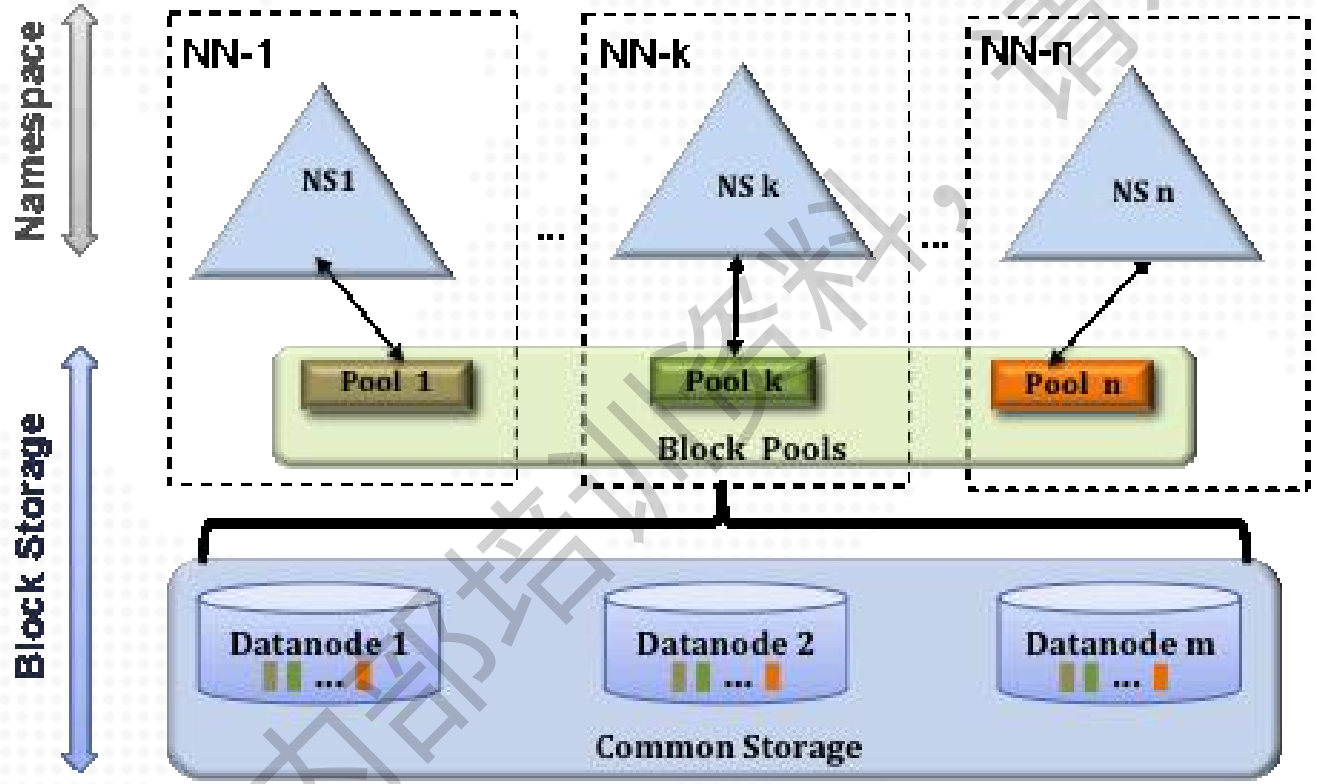
设计要点-元数据服务的高可用

- CEPH 一致性协议（Paxos）



设计要点-元数据服务的可扩展性

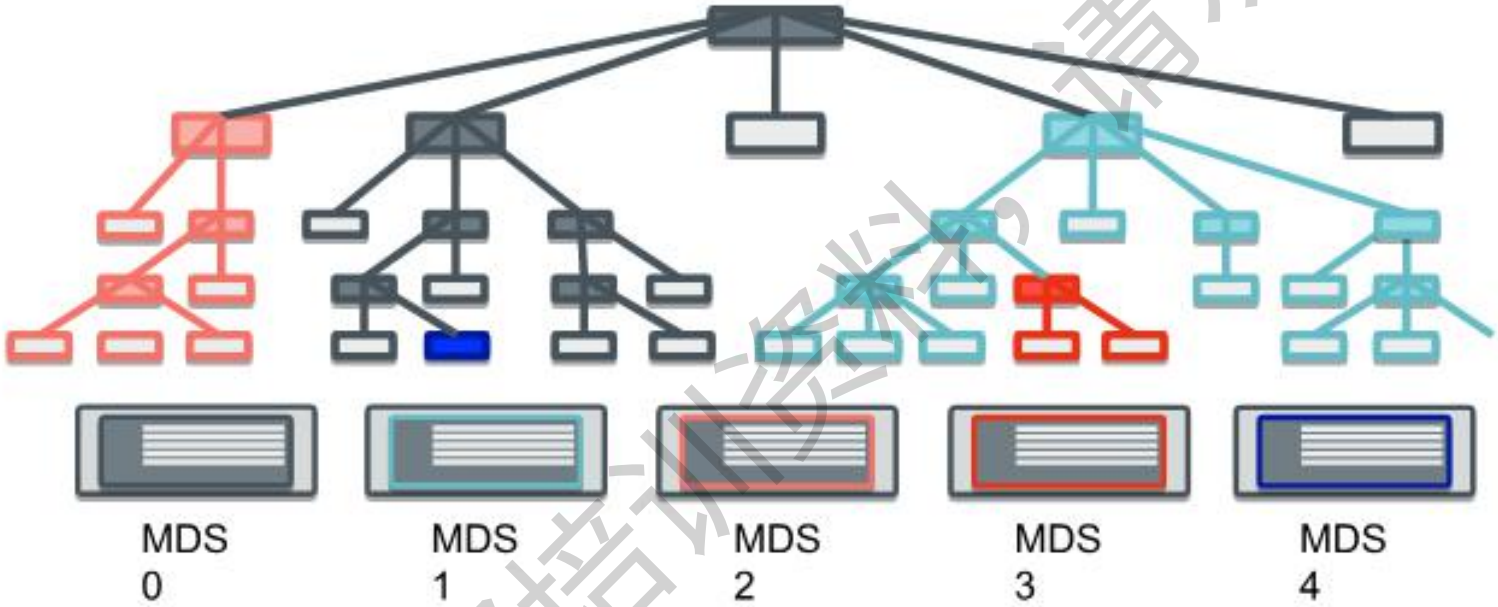
- HDFS Federation



设计要点-元数据服务的可扩展性

TRANSWARP 星 环 科 技

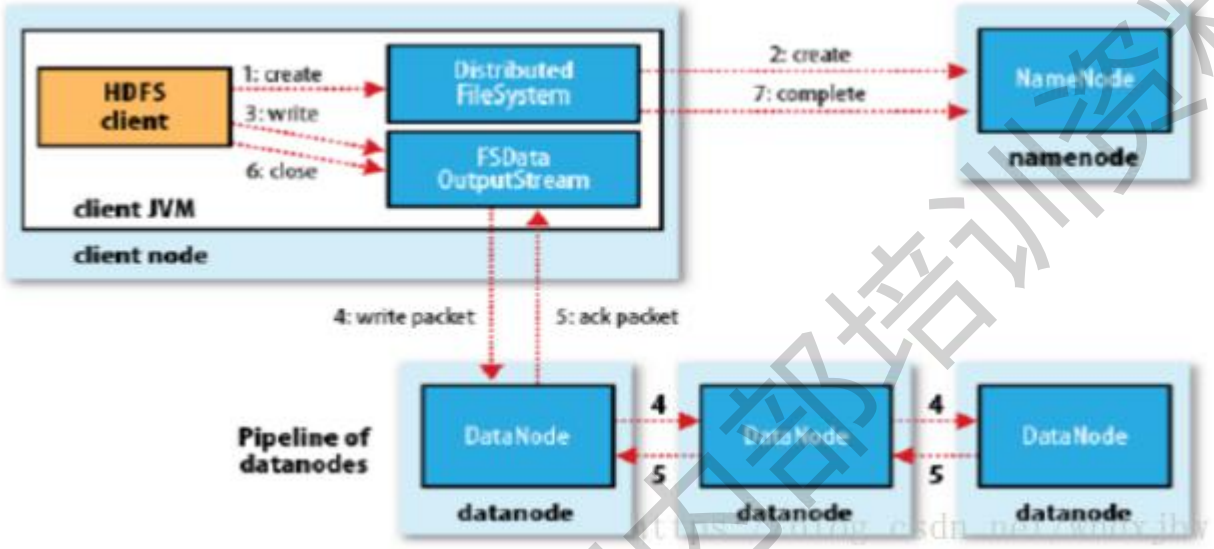
- CEPH 动态子树划分

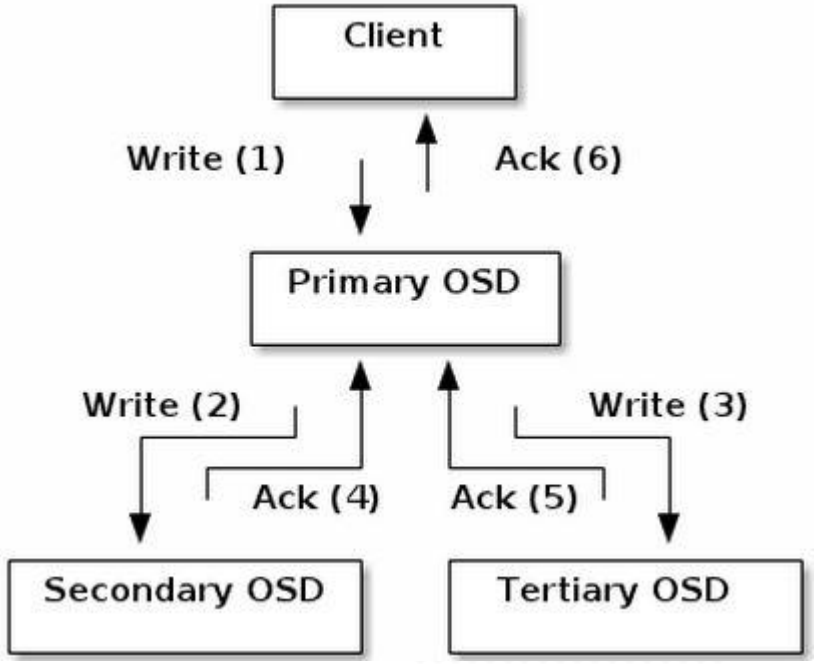


In this file system hierarchy tree, the shaded nodes represent directories and unshaded nodes files. The subtrees are partitioned in such a way that the metadata for the grey subtree is handled by MDS 0, the light blue subtree by MDS 1, the orange subtree by MDS 2, the red subtree by MDS 3 and the dark blue subtree(which is just a single directory) by MDS 4.

设计要点-写流程

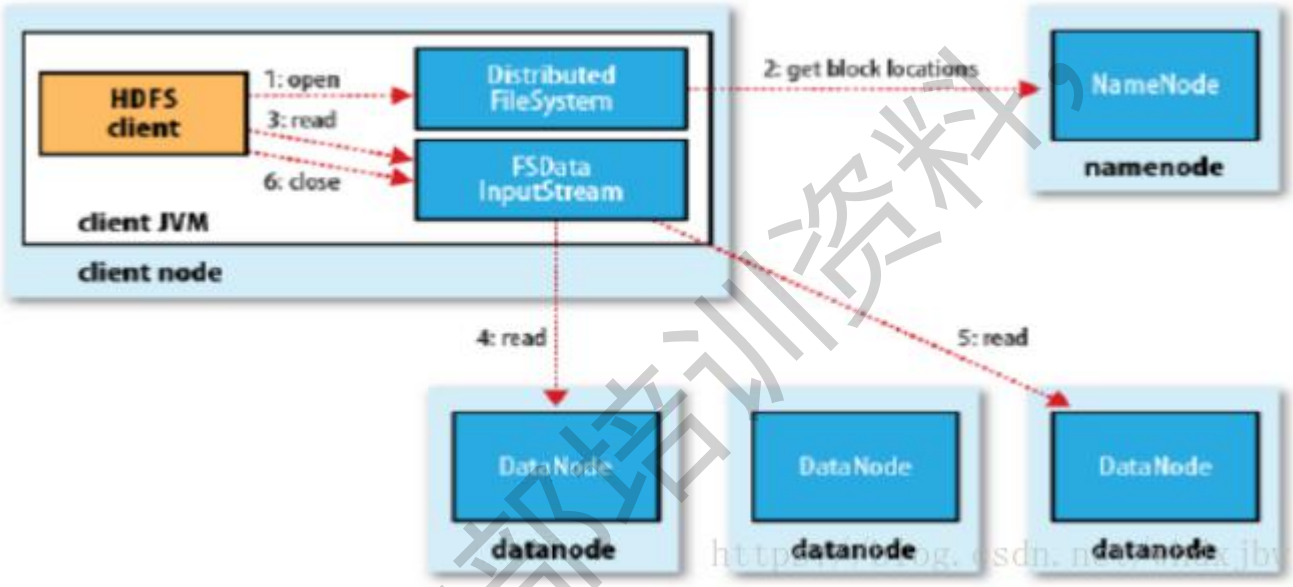
- Hadoop API 与 POSIX
- 元数据中心节点 有与无
- Pipeline写入 与 主从写入
- Journal分区 与 持久化存储





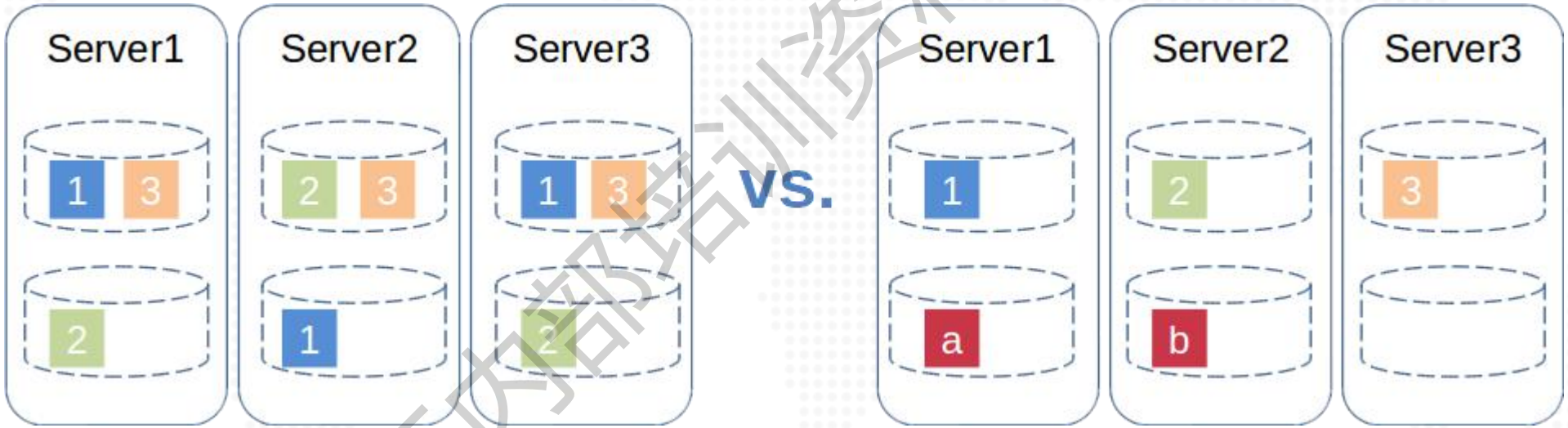
设计要点-读流程

- 强一致性/弱一致性/最终一致性
- 读策略，轮询/最近/延时最小/读主



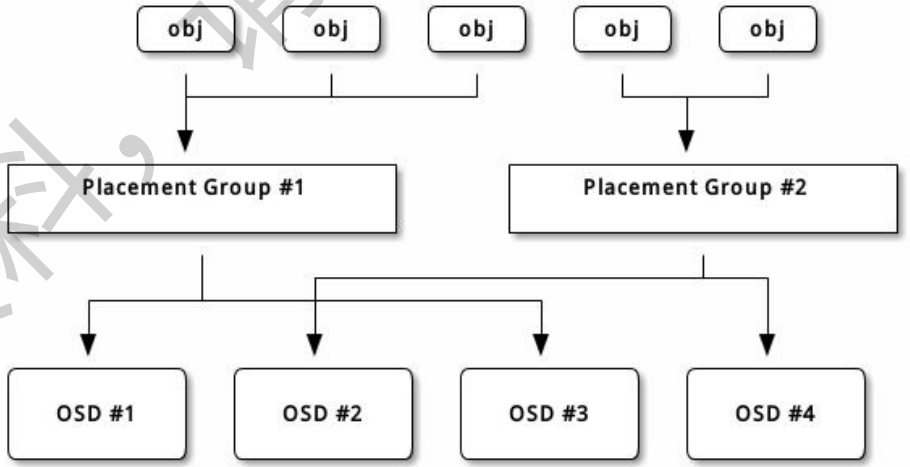
设计要点-数据冗余

- 防止丢数据。在出现各种异常/意外导致某个节点上的数据丢失后，数据还可以从其他节点读取/恢复。
- 实际数据存储。
  本地文件系统（ext4/xfs），单机存储引擎（LevelDB/RocksDB），专用文件系统（BlueStore）。
- 损坏副本的检测与自动复制，复制的QoS/优先级控制
- 副本策略和纠删码（EC）



设计要点-数据均衡

- 为什么需要数据均衡
- 如何保证存储节点数据均衡。
  有中心节点的主动分配 与 CRUSH算法的保证
- 扩容后的节点间数据的再平衡
  一致性哈希算法
  Balancer 与 自动平衡
- 对业务透明，较低的优先级，更严格的带宽控制



设计要点-数据删除/清理

- 数据删除策略
  实时删除 与 延时删除
  回收站机制

- 脏数据的检测与清理
  定期进行数据扫描，计算校验码

- 多余副本的清理

设计要点-数据安全

- 数据传输策略
  认证/MD5校验/加密
- 访问认证　　Kerberos/Cephx
- 访问权限控制
- 空间配额
- 数据加密
- 数据压缩
- 快照

Thanks

www.transwarp.io

公司地址 / Our Office

上海：徐汇区虹漕路88号B座11F&12F&15F，A座9F
北京：海淀区西直门北大街甲43号金运大厦B座1101室
广州：天河区体育东路140-148号南方证券大厦1015-1016室
郑州：郑东新区龙子湖湖心岛卫华研究院科研楼13层
南京：雨花台区宁双路19号云密城J栋10楼
联系电话：4007-676-098