51CTO.com
技术成就梦想

UBM

2012 / 10 / 25 ► 2012 / 10 / 26
中国 · 北京
October 25-26,2012
Beijing · China

World of Tech
WOT

2012云计算架构师峰会

Cloud Computing Architects Summit China 2012

揭示企业级IT架构转型 分享最新技术的应用落地

主办单位：51CTO.com 技术成就梦想　联合主办：UBM

分布式存储在网盘和在线备份的应用研究

@卢亿雷
Carbonite China技术总监兼高级架构师
jlu@carbonite.cn
2012-10-25

提纲

- 互联网存储应用的特点
- 网盘与在线备份的特点
- 云存储平台简介
- 总体实现方案
- 分布式数据库架构
- 分布式数据库特点
- 分布式文件系统架构

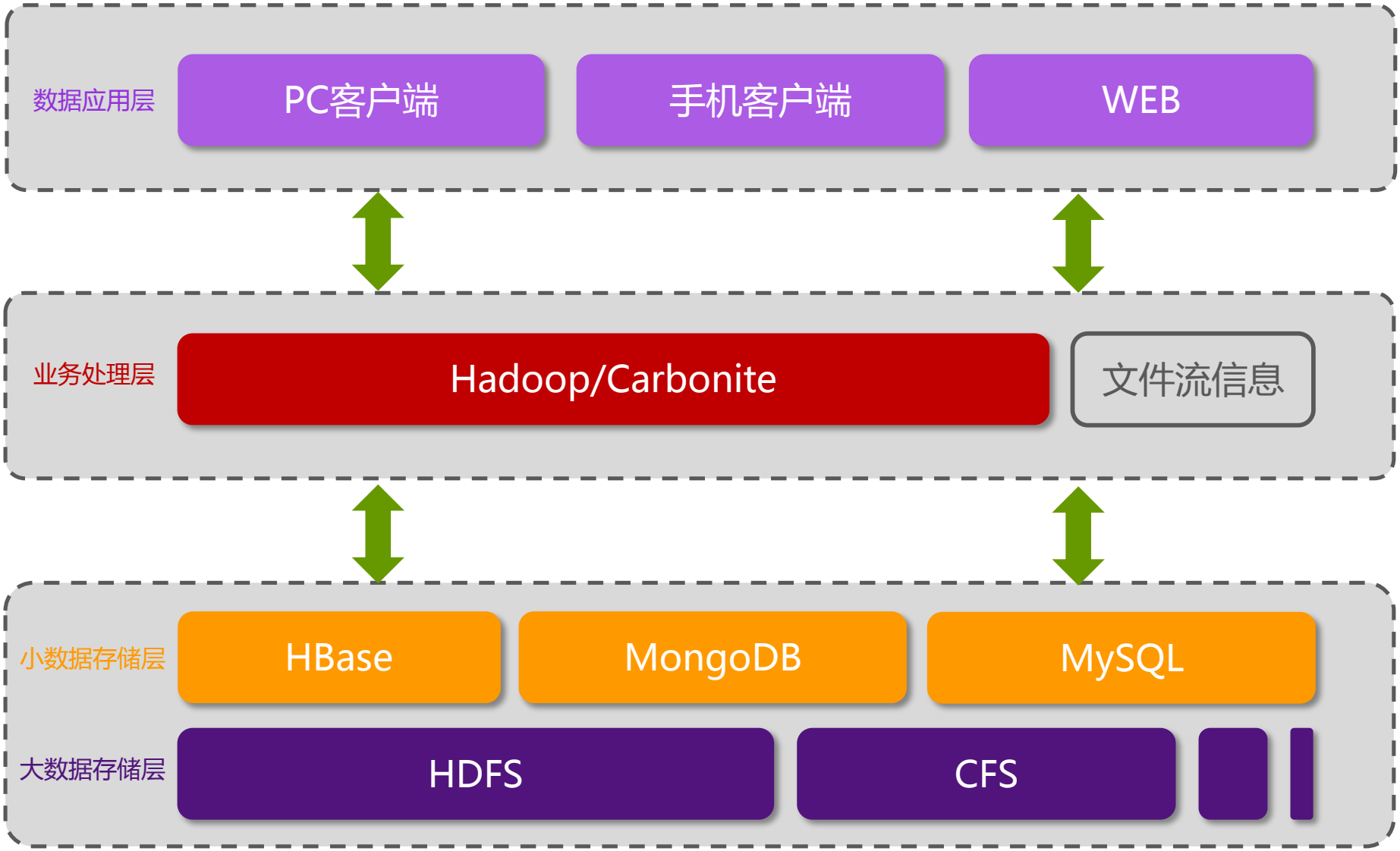
互联网存储应用的特点

- 高可靠：数据多份存储
- 高可用： 7X24小时
- 高扩展：随时扩容
- 高性能：高并发，低延迟
- 高安全：分布式密钥和严格IDC管理制度
- 高性价比：价格成本可控制
- 易监控和维护：响应及时

网盘和在线备份的特点

- 网盘与在线备份有很多相同点
  - 后台架构类似
  - 访问方式相似（上传及下载）
  - 都是云存储服务
  - 目标都是解决数据存储问题等
- 在线备份相对网盘（同步盘）的特殊性
  - 是否保持用户的原始路径
  - 是否有设备的区分
  - 是否可以备份企业级数据库
  - 是否密钥用户可以自己保存

云存储平台简介


数据应用层
PC客户端
手机客户端
WEB
业务处理层
Hadoop/Carbonite
文件流信息
小数据存储层
HBase
MongoDB
MySQL
大数据存储层
HDFS
CFS

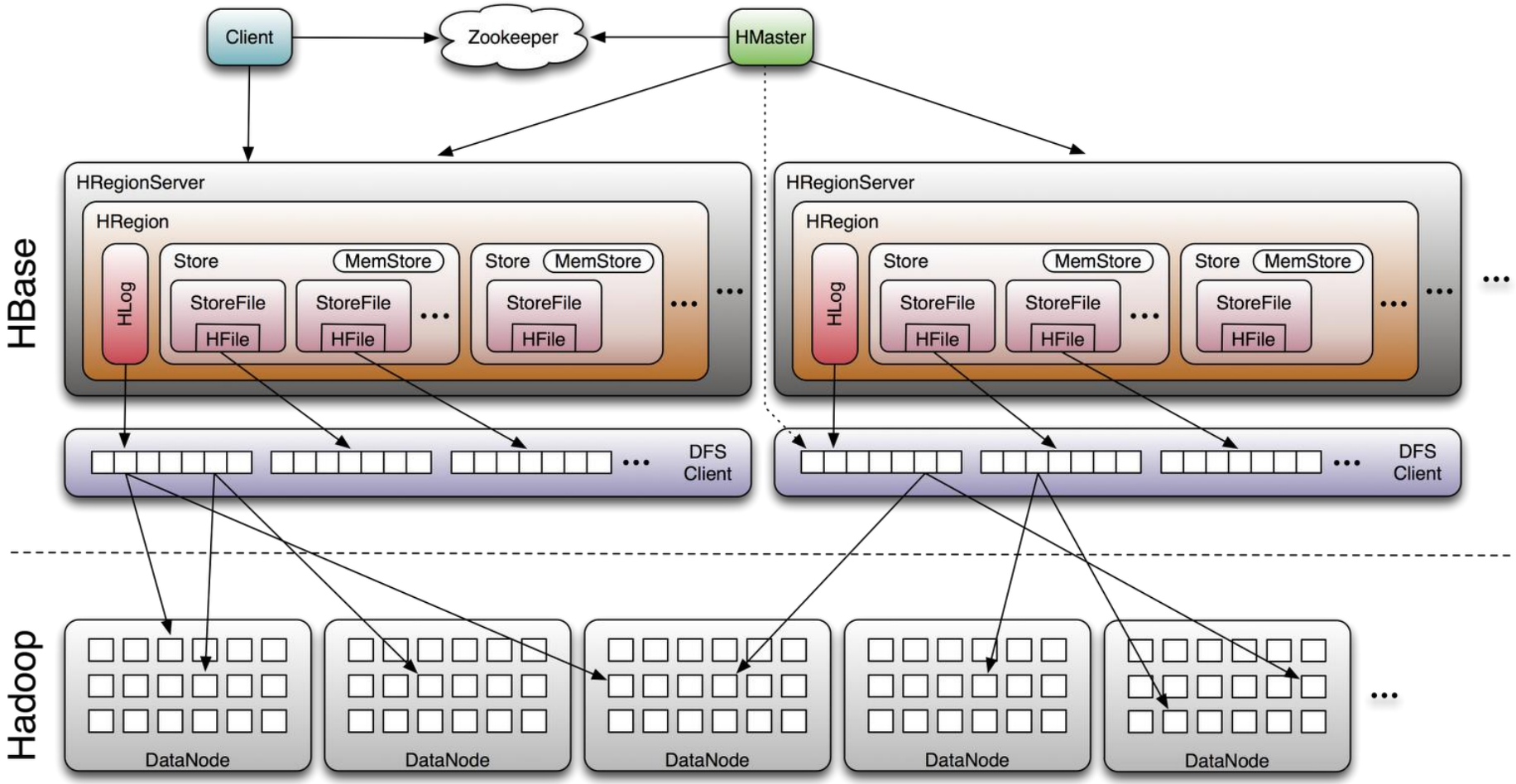
云存储平台简介(二)

- 总共有1000亿个文件
- 每天增加近1个亿新文件
- 总共服务器台数约有1万台
- 总存储量约为200P

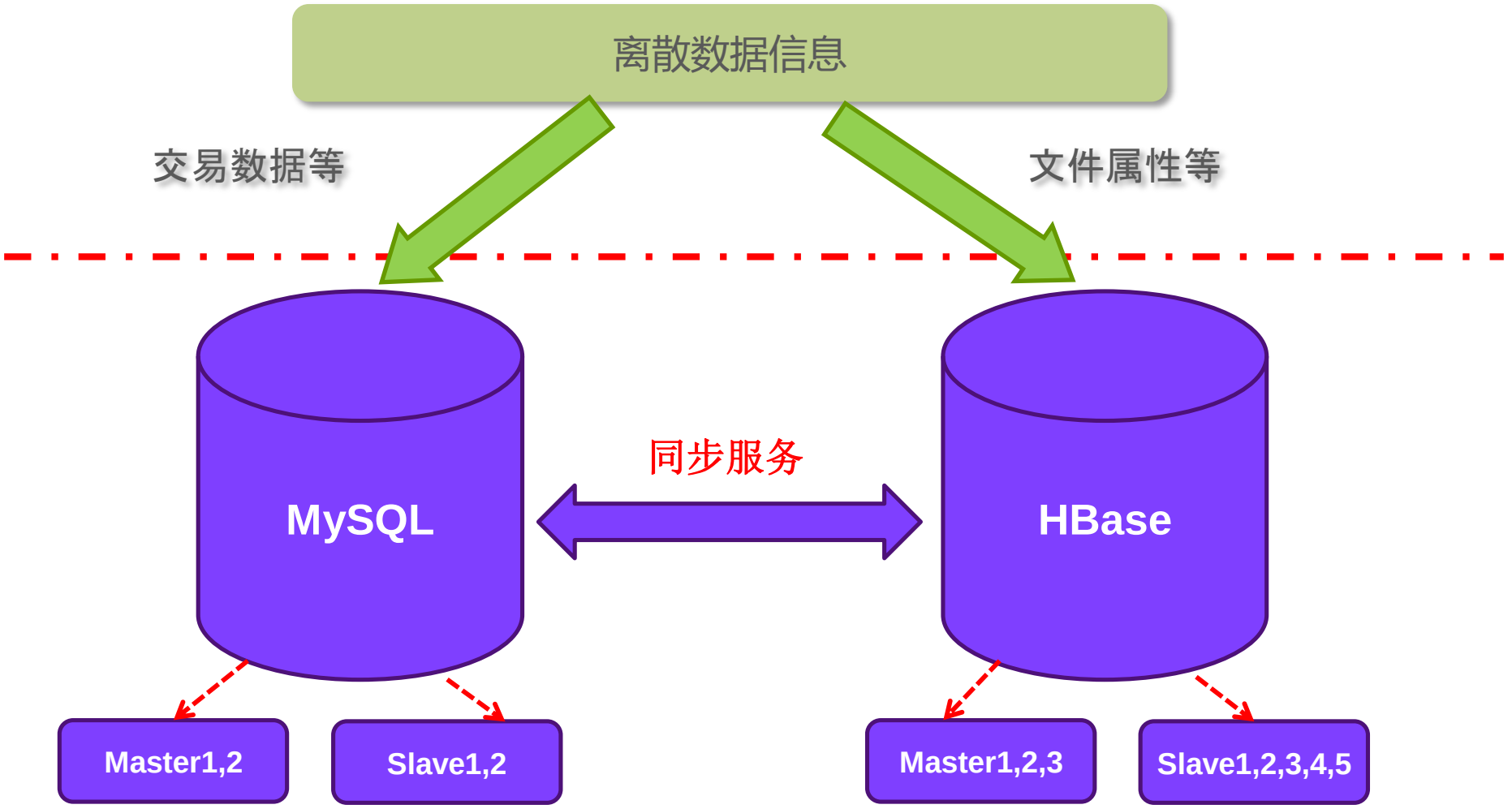
总体实现方案

- 分布式数据库：关系数据库与非关系型数据库(NoSql)集群解决结构化数据的海量存储和高效访问
  - MySQL
  - HBase
- 分布式文件系统：以文件为存储单位的非结构化数据存储
  - HDFS
  - MongoDB
- 前端应用负载均衡
  - LVS
  - Ngnix

分布式数据库架构



分布式数据库架构(二)



分布式数据库特点-HBase

- 成熟度比较高，目前已有多个著名公司在使用
  - Taobao
  - Facebook
  - Adobe
  - Twitter
  - Yahoo!
  - Trend Micro
  - 广告分析公司

分布式数据库特点-HBase（二）

凯备份™ CARBONITE

- 行强一致性
  - 同一行数据的读写只在同一台RS（RegionServer）机器上进行
  - 同一行的列的写入是原子操作
- 水平自动伸缩
  - Region的自动分裂(生产系统需要看具体情况)
  - Master的自动均衡
  - 增加RegionServer机器即增加读写吞吐量及处理能力
  - 增加DataNode机器可增加容量

分布式数据库特点-HBase（三）

- 任意增加列
- 高性能随机写
- 支持Thrift框架

- 合理设计rowKey 和 Pre-Sharding
  - 尽量避免只操作少数几台机器；
  - 根据数据量、RegionServer个数合理Pre-Sharding。
- 充分利用Filter功能
  - SingleColumnValueFilter
  - **SubstringComparator**
  - BinaryPrefixComparator
  - FamilyFilter
  - QualifierFilter
  - ColumnPrefixFilter
  - ColumnPaginationFilter

- 可根据应用需求重写某些方法
  - **SubstringComparator**

```
@Override
public int compareTo(byte[] value) {
  String laststr = Bytes.toString(value).toLowerCase();
  return laststr.contains(substr) ? 0 : 1;
}
```

- 充分FilterList的addFilter
  - BinaryPrefixComparator
  - QualifierFilter
  - ColumnPrefixFilter
  - PageFilter

HBase经验分享（四）

- 考虑容量开启压缩
  - 目前主要是lzo方式
- 提高随机读性能
  - 前端增加一个分布式缓存Radis系统
- 系统参数优化
  - GC策略：-XX:+UseConcMarkSweepGC -XX:+UseParNewGC -XX:CMSInitiatingOccupancyFraction=70
  - 读写策略优化

系统参数优化

- 读优化：
  - hbase.regionserver.handler.count
  - hbase.regionserver.global.memstore.upperLimit/lowerLimit
  - hbase.hregion.memstore.block.multiplier
  - hbase.hstore.blockingStoreFiles
  - hbase.hregion.max.filesize

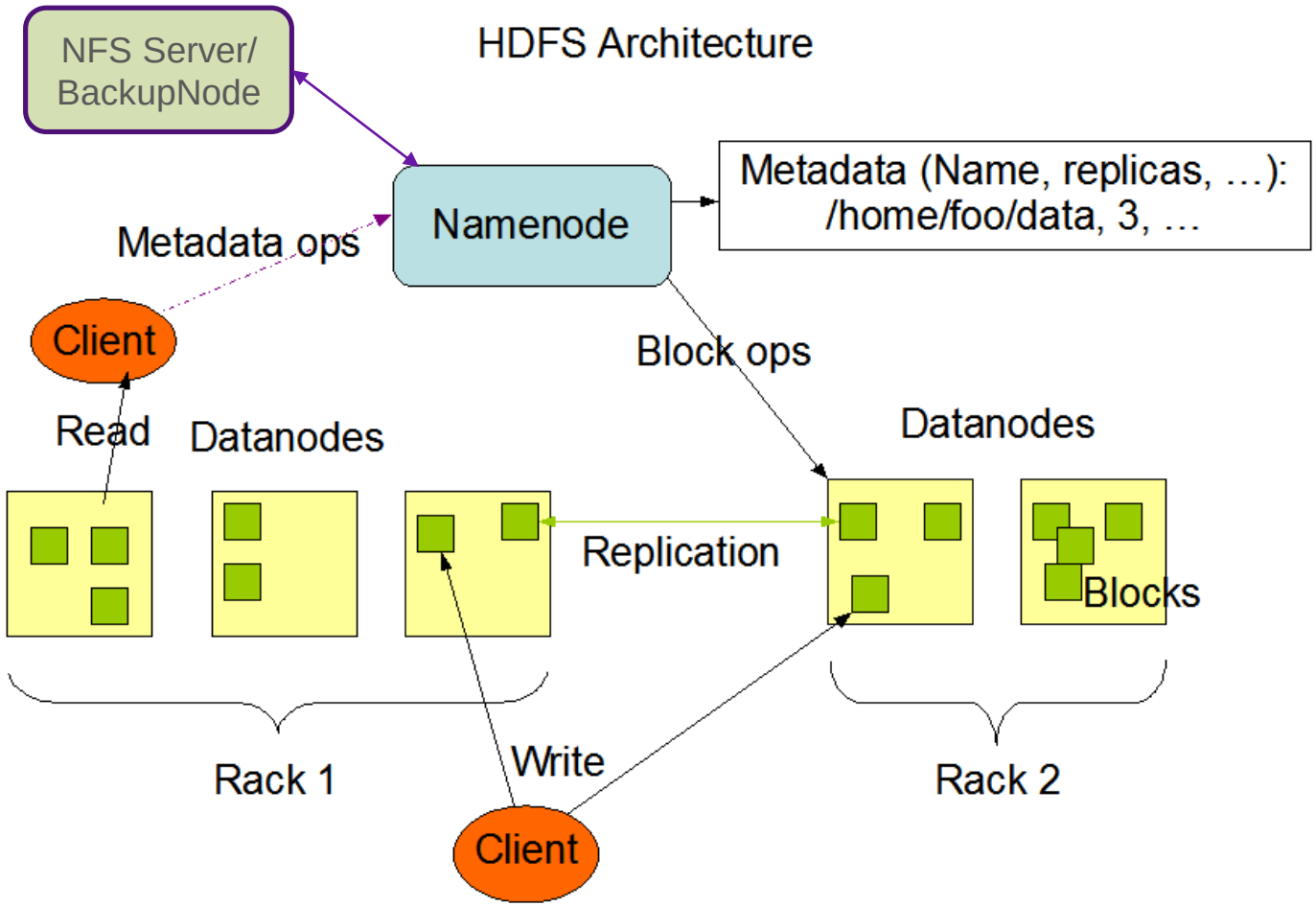
- 写优化：
  - Bloomfilter
  - in-memory
  - Blockcache
  - hfile.block.cache.size

分布式文件系统架构

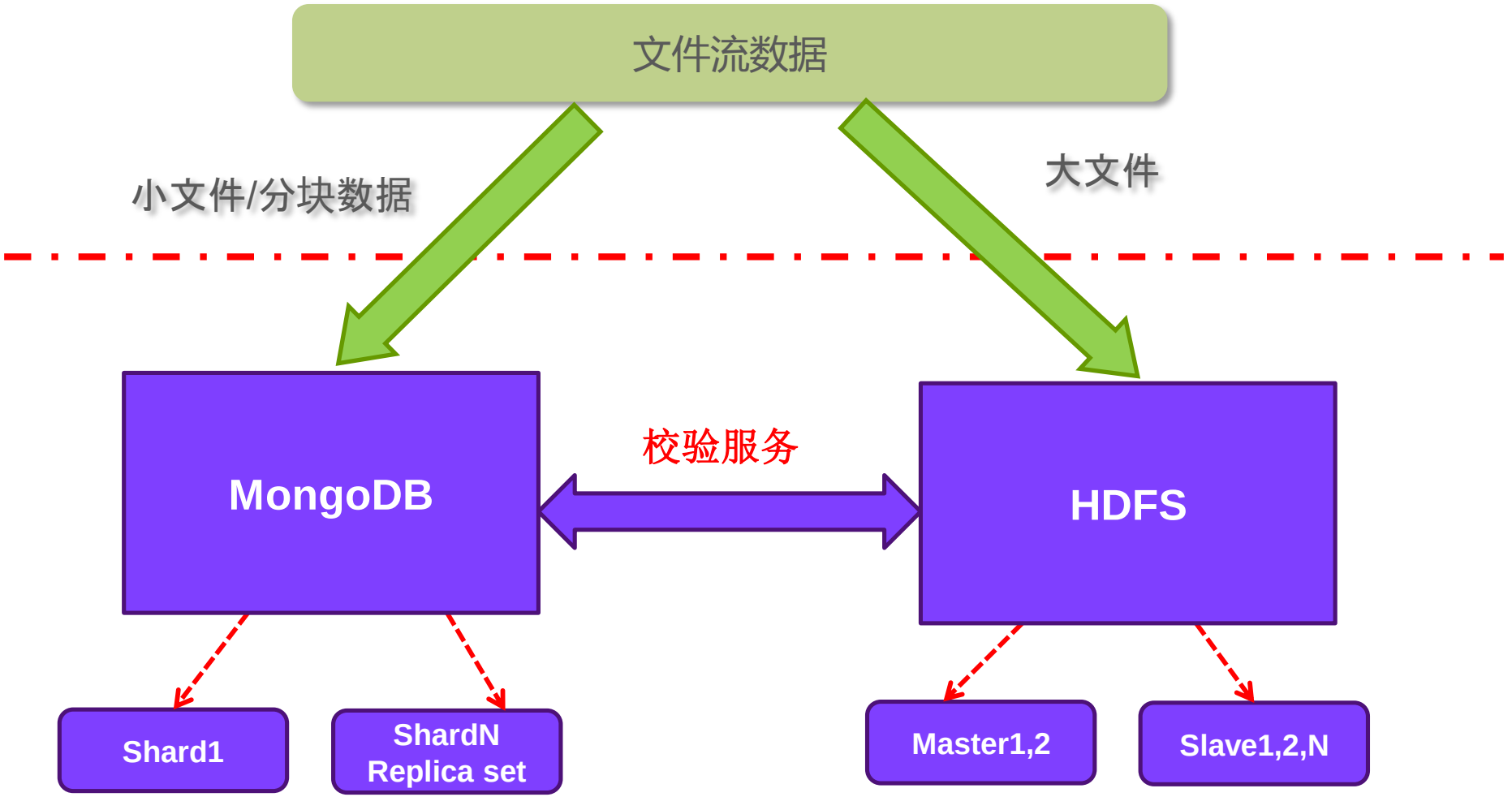
凯备份™ CARBONITE

- 大文件数据
  - HDFS
- 小文件数据及分块数据
  - MongoDB
- 文件“垃圾”数据回收

分布式文件系统架构-HDFS（二）



分布式文件系统架构(三)



- 尽量创建索引
- 限定返回结果条数
- Filter只返回需要的数据
- 优化主键，尽量自己控制主健ID
- UUID主键使用BinaryData数据类型存储

- 系统参数优化
  - GC策略
- 带宽策略优化
  - 带内与带外心跳的区分
  - NameNode的备份至NFS时尽量不影响正常带宽使用
- 同步锁机制尽量少用
  - 所有文件IO操作的地方尽可能不要加同步锁
- 文件复本数设置
  - 尽量根据应用的访问频率设置不同份数

NameNode数据损坏原因分析及修复

- 查看日志，确定大概问题位置
- 备份fsimage，edits等
- 使用OfflineImageViewer方法查看（可能不行）
- 编译相应Hadoop版本
- 在相关读取文件信息的地方适当加一些TRY/CATCH
- 多次查找分析元数据信息
- Replay重新生成fsimage

THANKS

www.carbonite.cn