HDFS技术原理

www.huawei.com

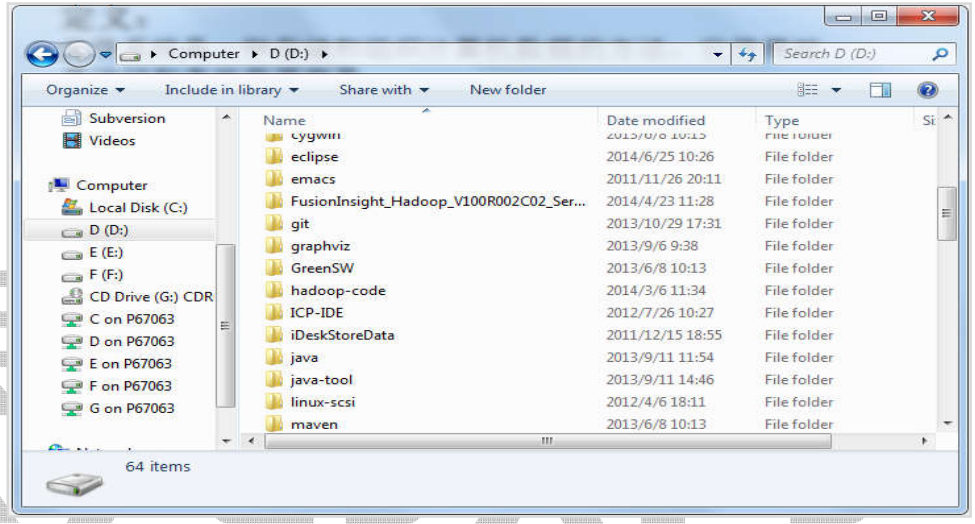
HUAWEI

目标

- 学完本课程后，您将能够:
  - 了解HDFS使用的场景
  - 了解HDFS系统架构
  - 了解HDFS关键特性

目录

字典与文件系统



字典	文件系统
部首检字表 （一）部首目录 （二）检字表 （三）难检字笔画索引	文件名 元数据（Metadata）
字典正文	数据块（Block）

HDFS概述

- HDFS(Hadoop Distributed File System)基于Google发布的GFS论文设计开发。
- 其除具备其它分布式文件系统相同特性外，还有自己特有的特性：
  - 高容错性：认为硬件总是不可靠的。
  - 高吞吐量：为大量数据访问的应用提供高吞吐量支持。
  - 大文件存储：支持存储TB-PB级别的数据。

HDFS适合做什么？
大文件存储与流式数据访问

HDFS不适合做什么？
- 大量小文件存储
- 随机写入
- 低延迟读取

常见的分布式文件系统

GFS	•Google公司为了满足本公司需求而开发的基于Linux的专有分布式文件系统。。尽管Google公布了该系统的一些技术细节，但Google并没有将该系统的软件部分作为开源软件发布。
HDFS	•Hadoop 实现了一个分布式文件系统（Hadoop Distributed File System），简称HDFS。
Ceph	•是加州大学圣克鲁兹分校的Sage weil攻读博士时开发的分布式文件系统.没有生产应用。
Lustre	•Lustre是一个大规模的、安全可靠的，具备高可用性的集群文件系统，它是由SUN公司开发和维护的。可以支持超过10000个节点，数以PB的数据量存储系统。是目前主流网盘的解决方案。
mogileFS	•由memcahed的开发公司danga一款perl开发的产品，目前国内使用mogielFS的有图片托管网站yupoo等。
FastDFS	•一个开源的轻量级分布式文件系统，它对文件进行管理，功能包括：文件存储、文件同步、文件访问（文件上传、文件下载）等，解决了大容量存储和负载均衡的问题。特别适合以文件为载体的在线服务，如相册网站、视频网站等等。
TFS	•TFS（Taobao !FileSystem）是一个高可扩展、高可用、高性能、面向互联网服务的分布式文件系统，主要针对海量的非结构化数据，它构筑在普通的Linux机器 集群上，可为外部提供高可靠和高并发的存储访问。
GridFS	•MongoDB是一种知名的NoSql数据库，GridFS是MongoDB的一个内置功能，它提供一组文件操作的API以利用MongoDB存储文件。

HUAWEI

HDFS特性说明

HDFS适合做什么？

检测硬件失效

多副本存储

多并发请求

大文件存储

数据一致性

多硬件平台

移动计算能力

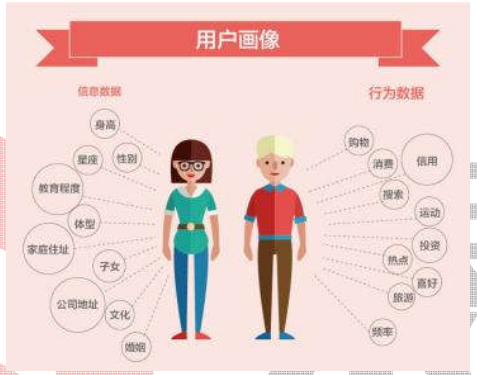
HDFS不适合做什么？

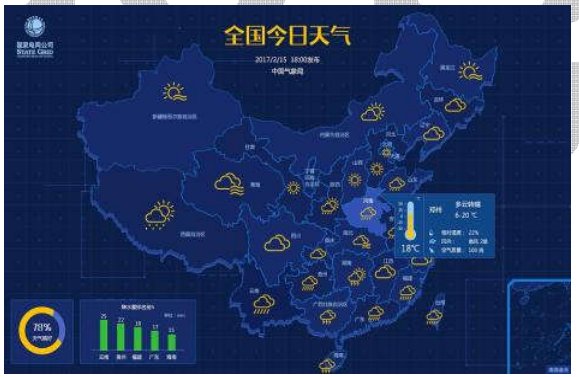
低延迟读取、

随机写入、

大量小文件

HDFS应用场景举例

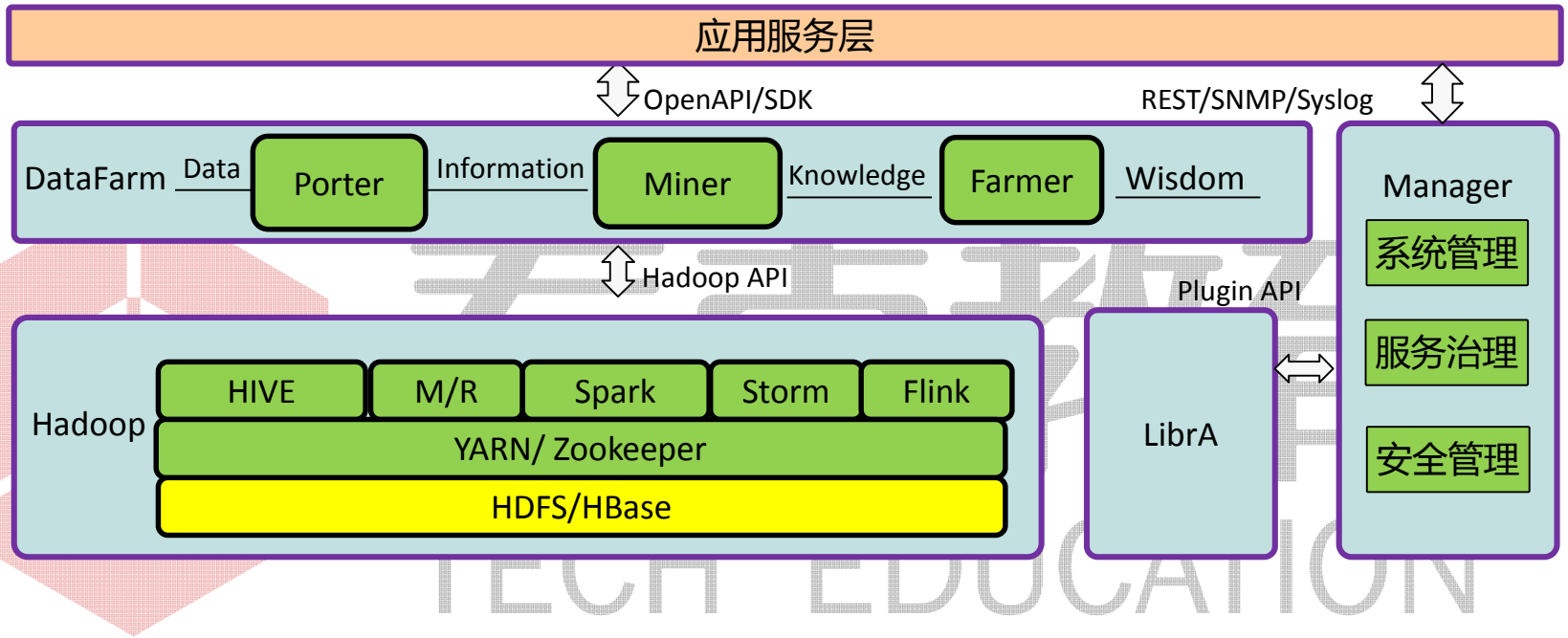




HUAWEI

目录

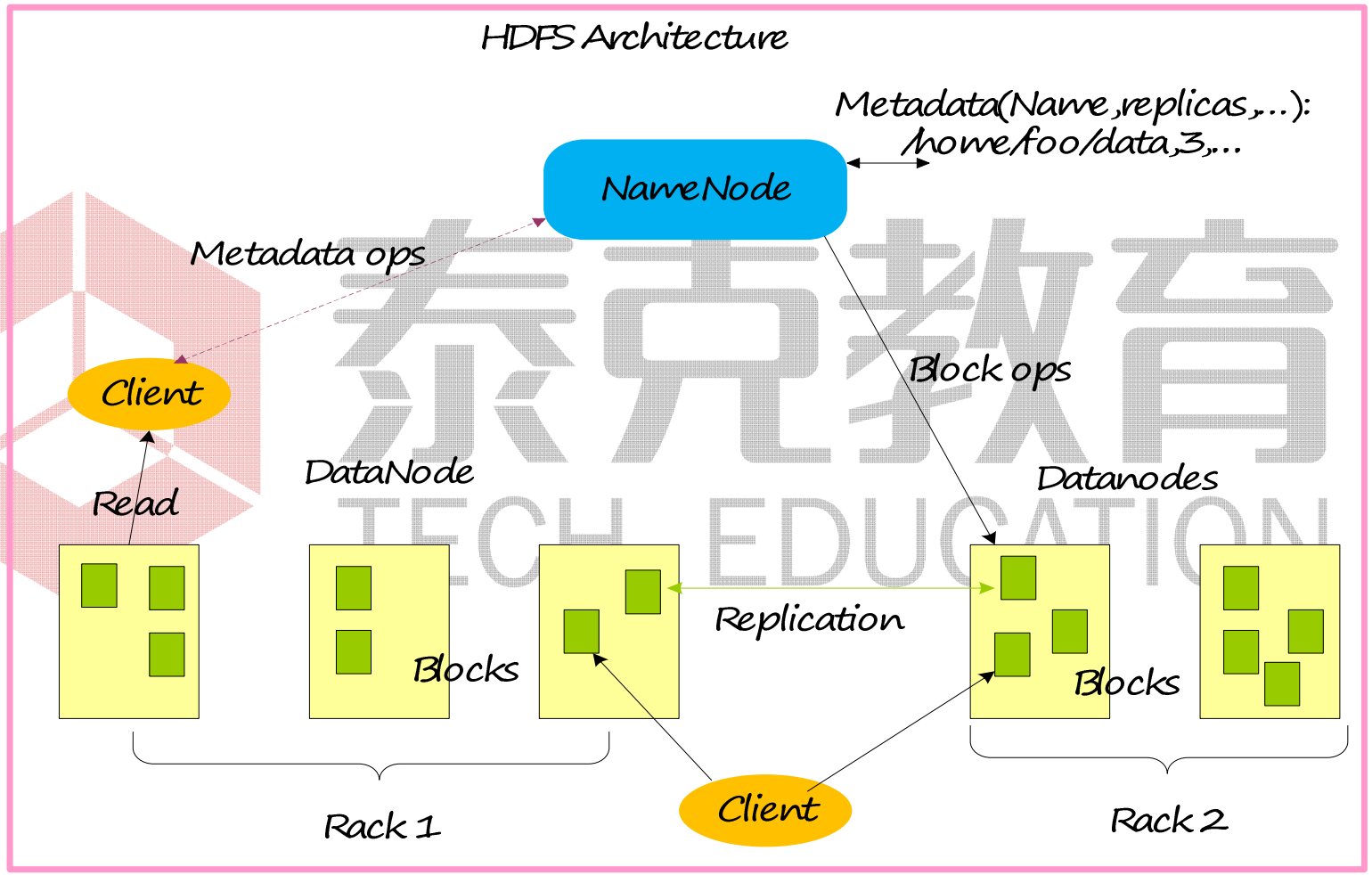
HDFS在FusionInsight产品的位置



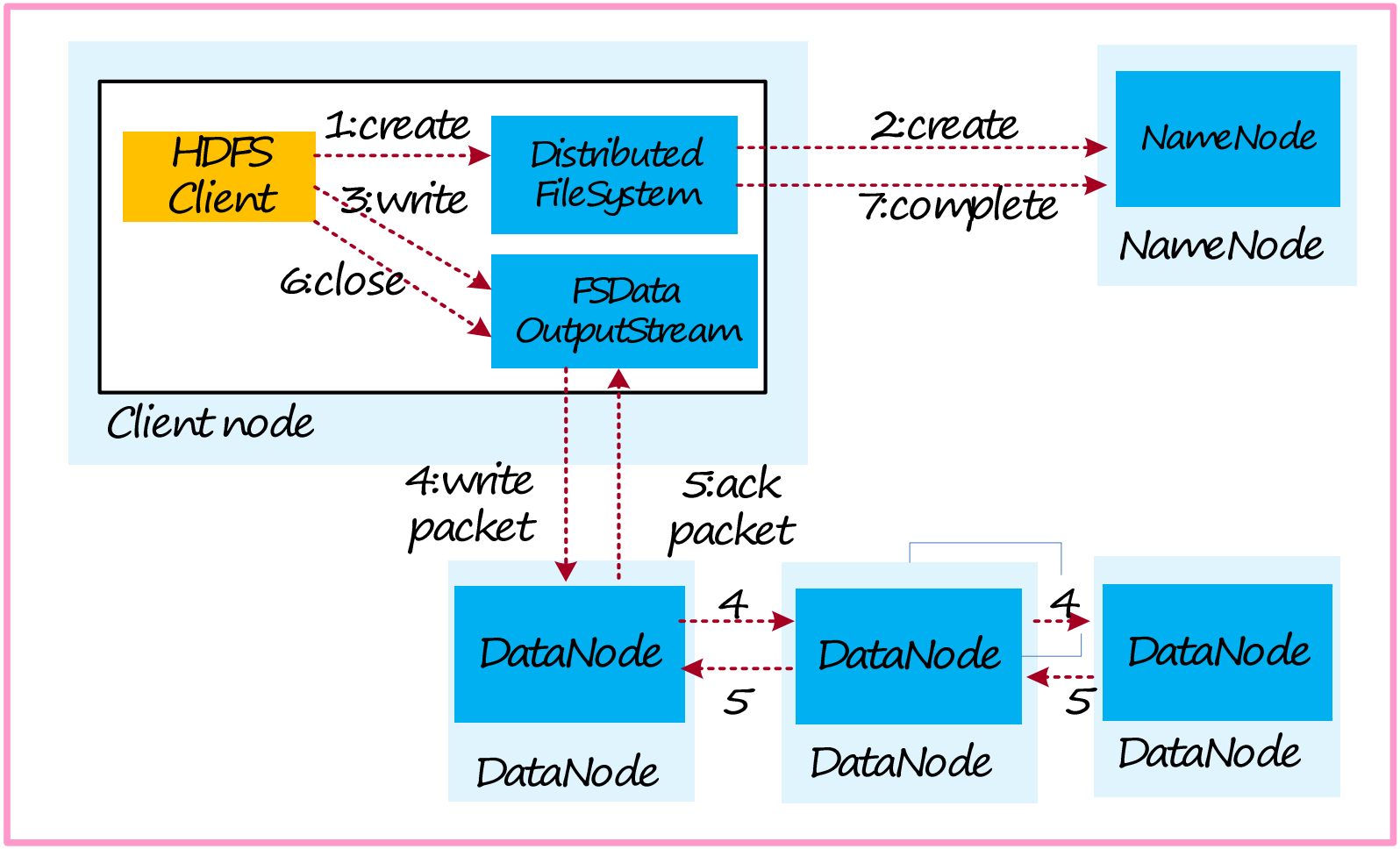
HDFS作为Hadoop的基础存储设施，实现了一个分布式、高容错、可线性扩展的文件系统。

目录

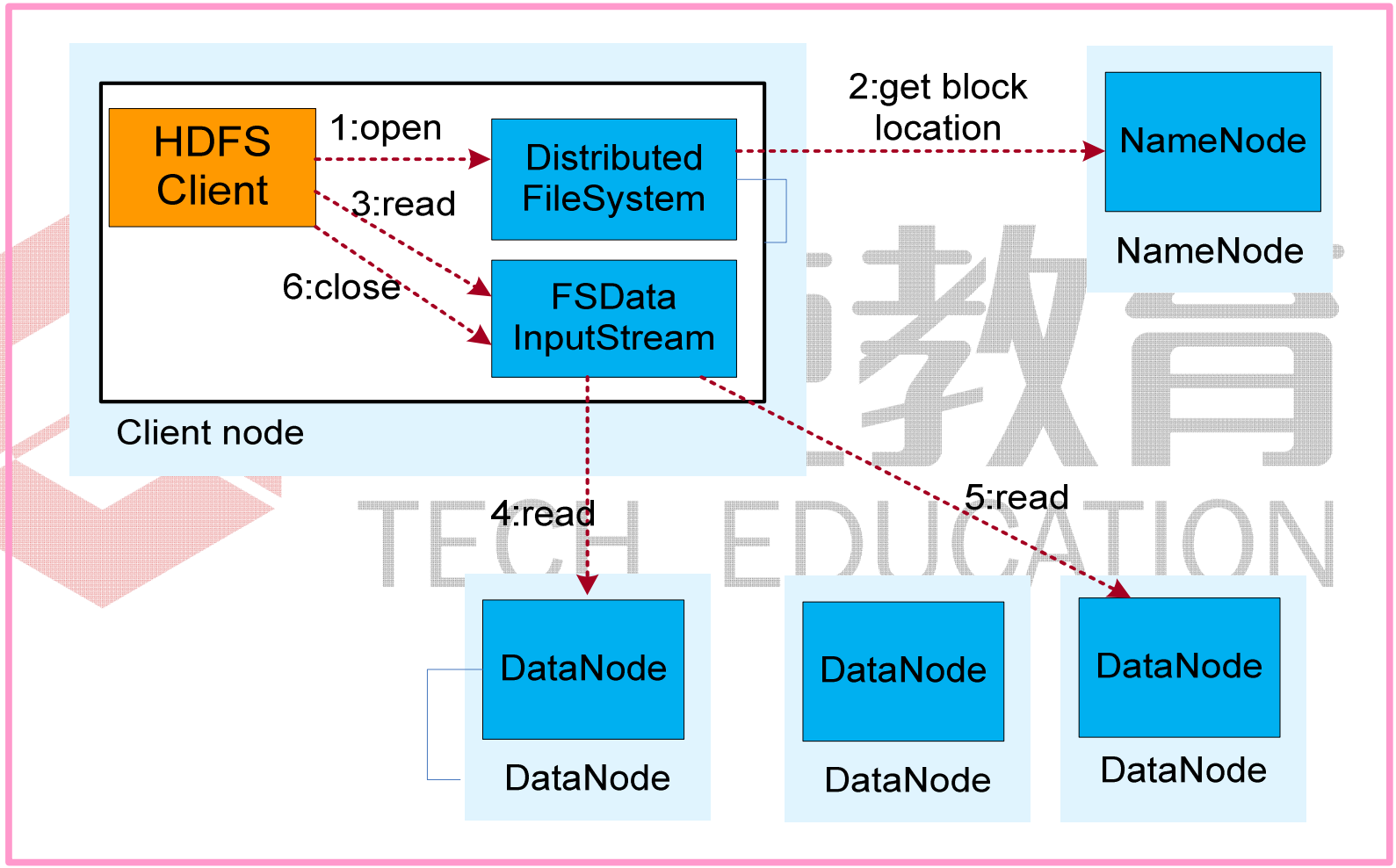
基本系统架构



HDFS数据写入流程

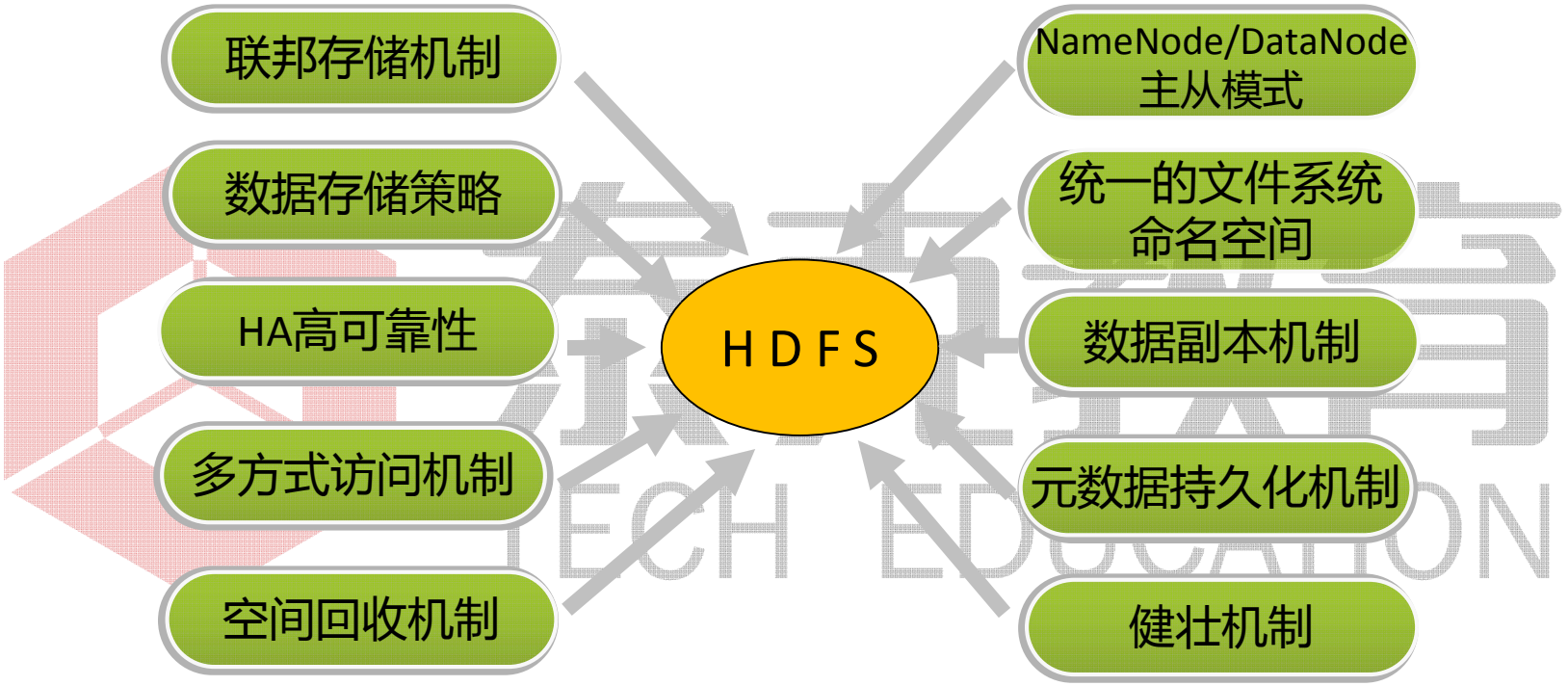


HDFS数据读取流程

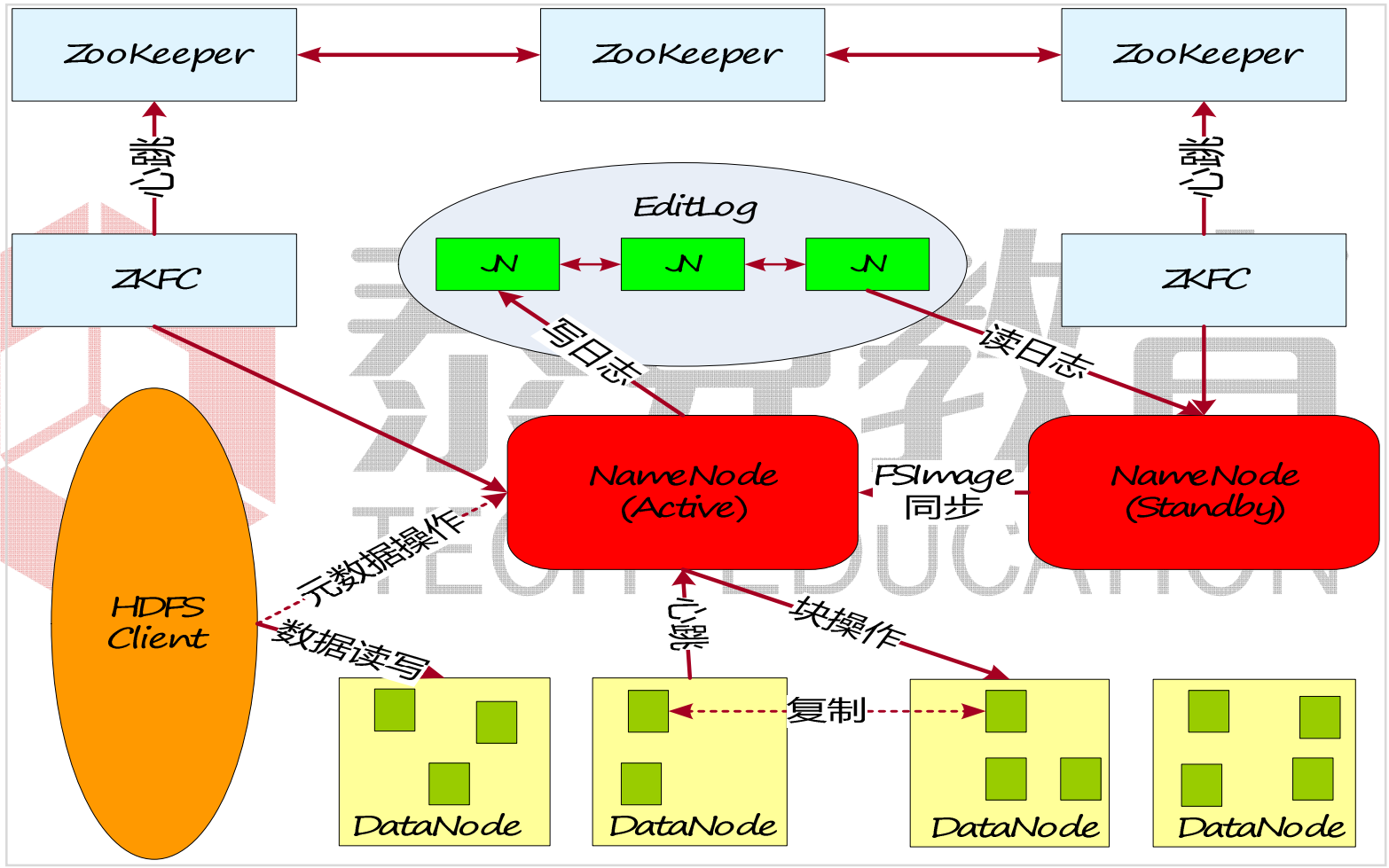


目录

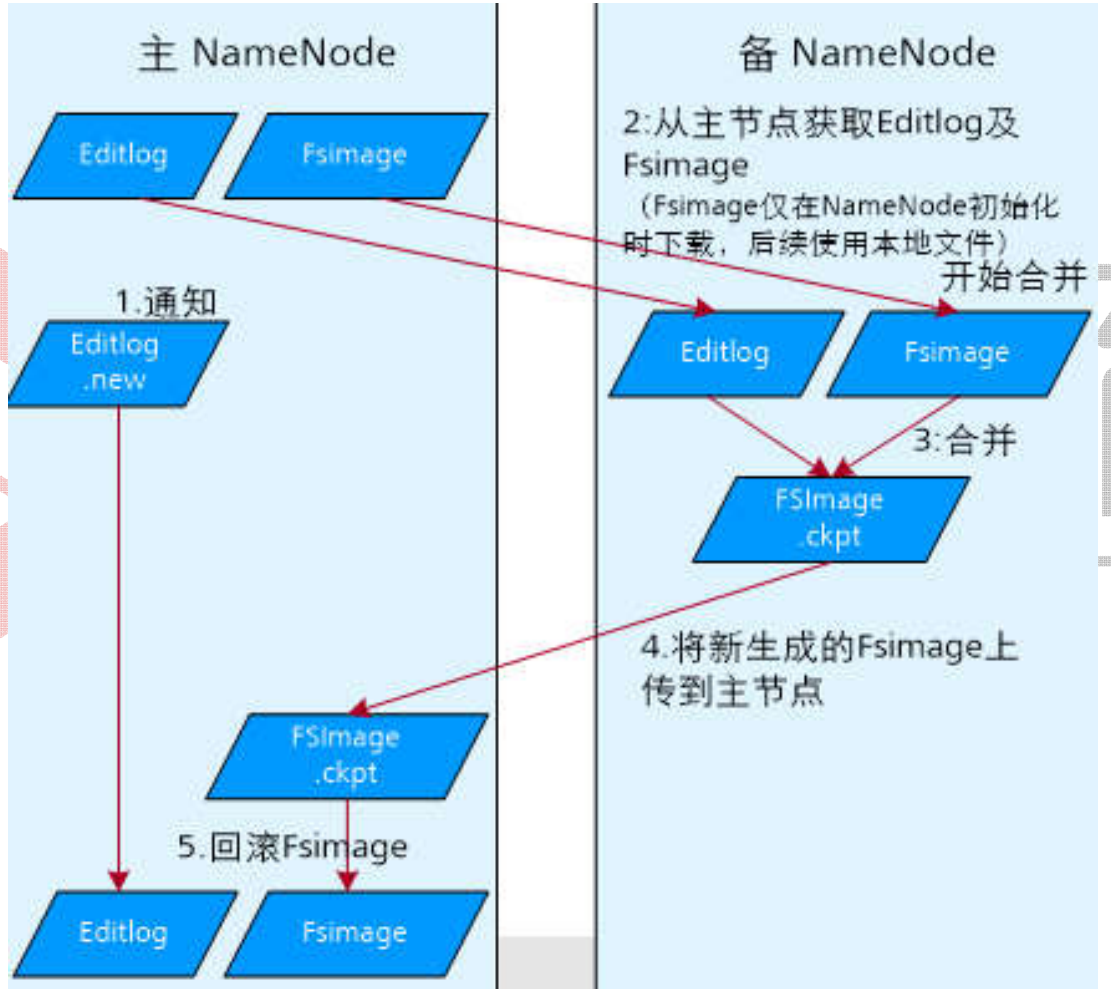
HDFS架构关键设计



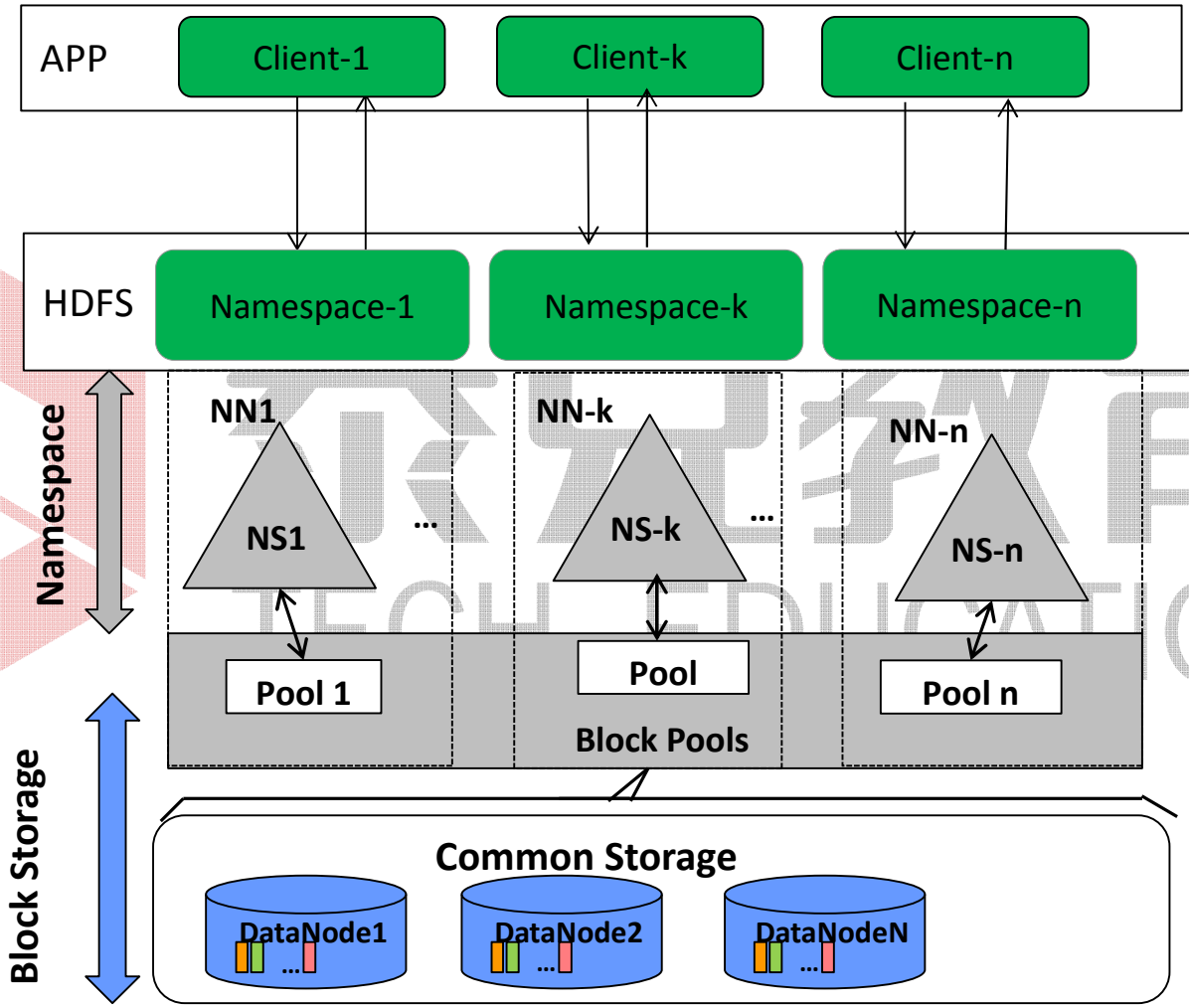
HDFS高可靠性（HA）



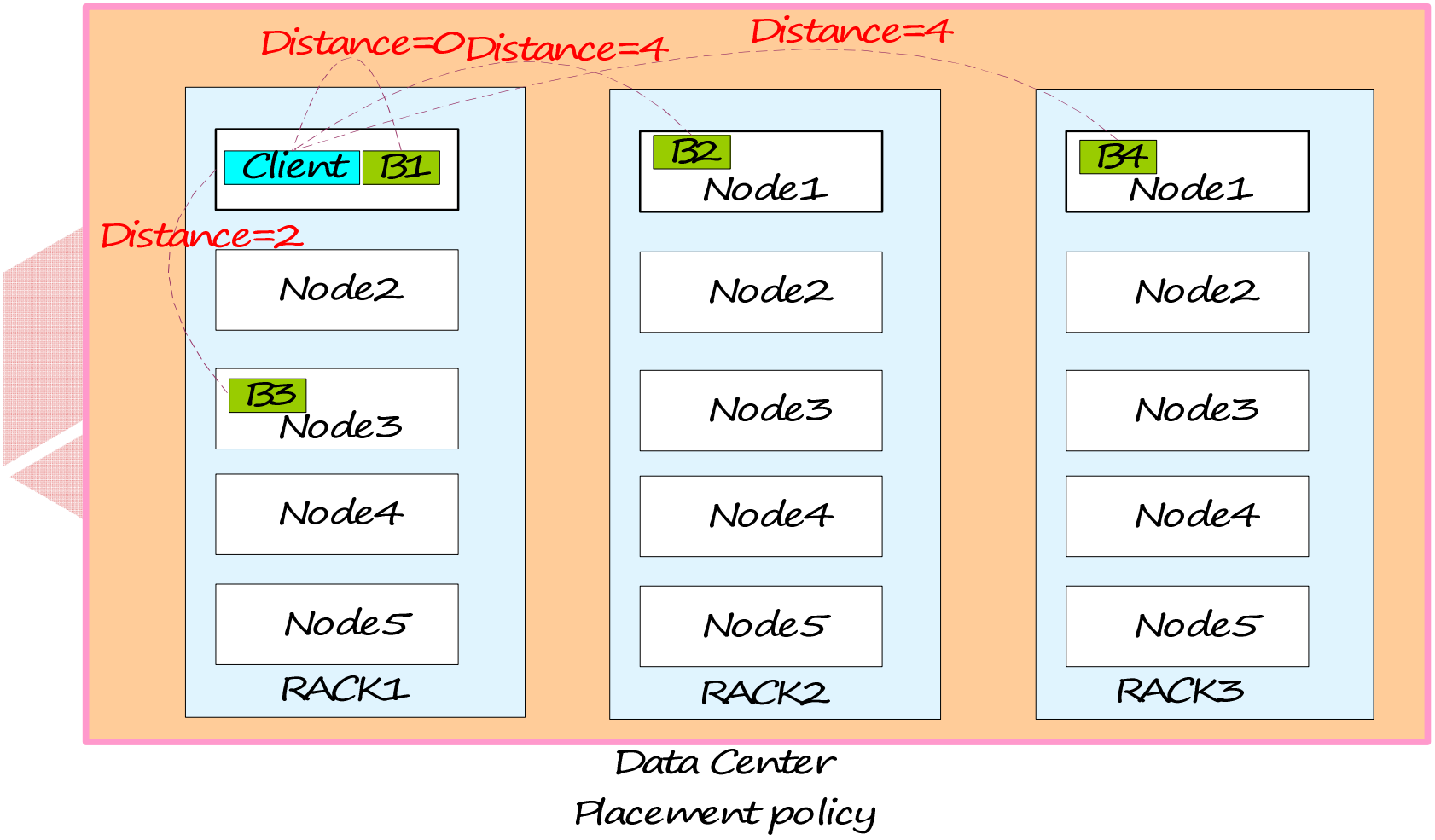
元数据持久化



HDFS联邦（Federation）



数据副本机制


Distance=0
Distance=4
Distance=4
Client
B1
Distance=2
B2
Node1
B4
Node1
Node2
Node2
Node2
B3
Node3
Node3
Node3
Node4
Node4
Node4
Node5
Node5
Node5
RACK1
RACK2
RACK3
Data Center
Placement policy

配置HDFS数据存储策略

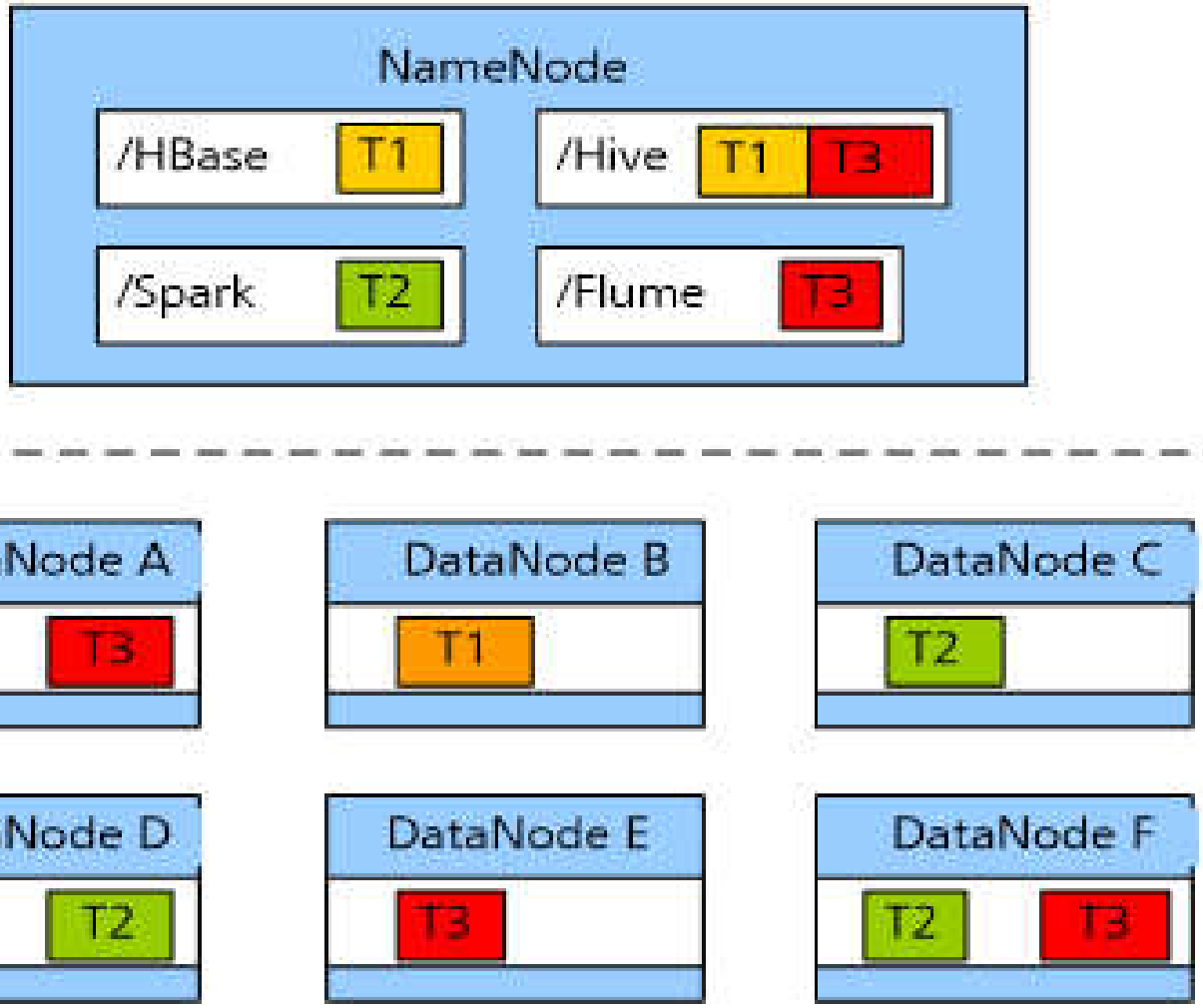
- 默认情况下，HDFS NameNode自动选择DataNode保存数据的副本。在实际业务中，存在以下场景：
  - DataNode上存在的不同的存储设备，数据需要选择一个合适的存储设备分级存储数据。
  - DataNode不同目录中的数据重要程度不同，数据需要根据目录标签选择一个合适的DataNode节点保存。
  - DataNode集群使用了异构服务器，关键数据需要保存在具有高度可靠性的节点组中。

配置HDFS数据存储策略 – 分级存储

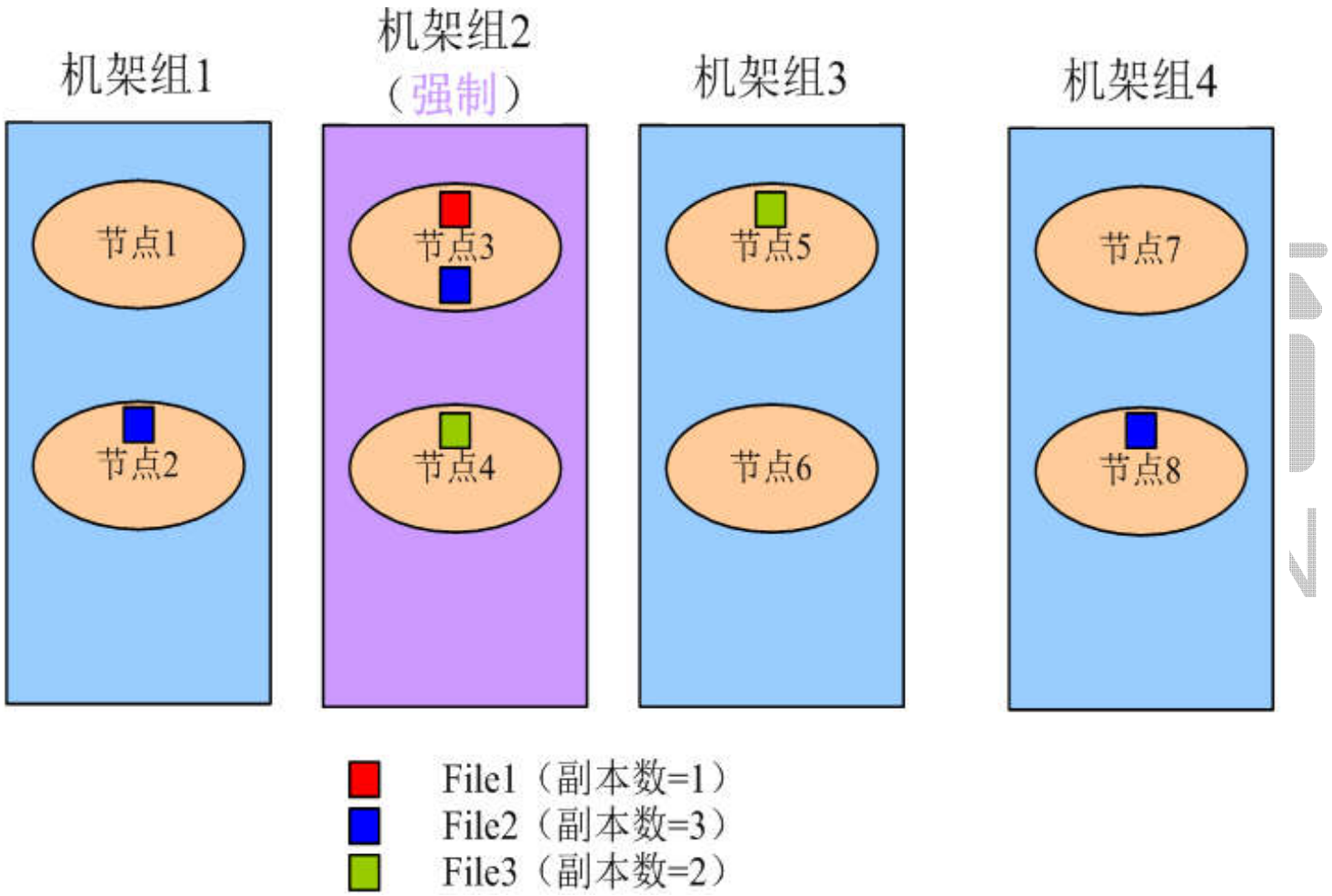
- 配置DataNode使用分级存储
  - HDFS的分级存储框架提供了RAM_DISK（内存盘）、DISK（机械硬盘）、ARCHIVE（高密度低成本存储介质）、SSD（固态硬盘）四种存储类型的存储设备。
  - 通过对四种存储类型进行合理组合，即可形成适用于不同场景的存储策略。

策略ID	名称	Block放置位置（副本数）	备选存储策略	副本的备选存储策略
15	LAZY_PERSIST	RAM_DISK:1, DISK:*n*-1	DISK	DISK
12	All_SSD	SSD:*n*	DISK	DISK
10	ONE_SSD	SSD:1, DISK:*n*-1	SSD, DISK	SSD, DISK
7	HOT (default)	DISK:*n*	<none>	ARCHIVE
5	WARM	DISK:1, ARCHIVE:*n*-1	ARCHIVE, DISK	ARCHIVE, DISK
2	COLD	ARCHIVE:*n*	<none>	<none>

配置HDFS数据存储策略 – 标签存储

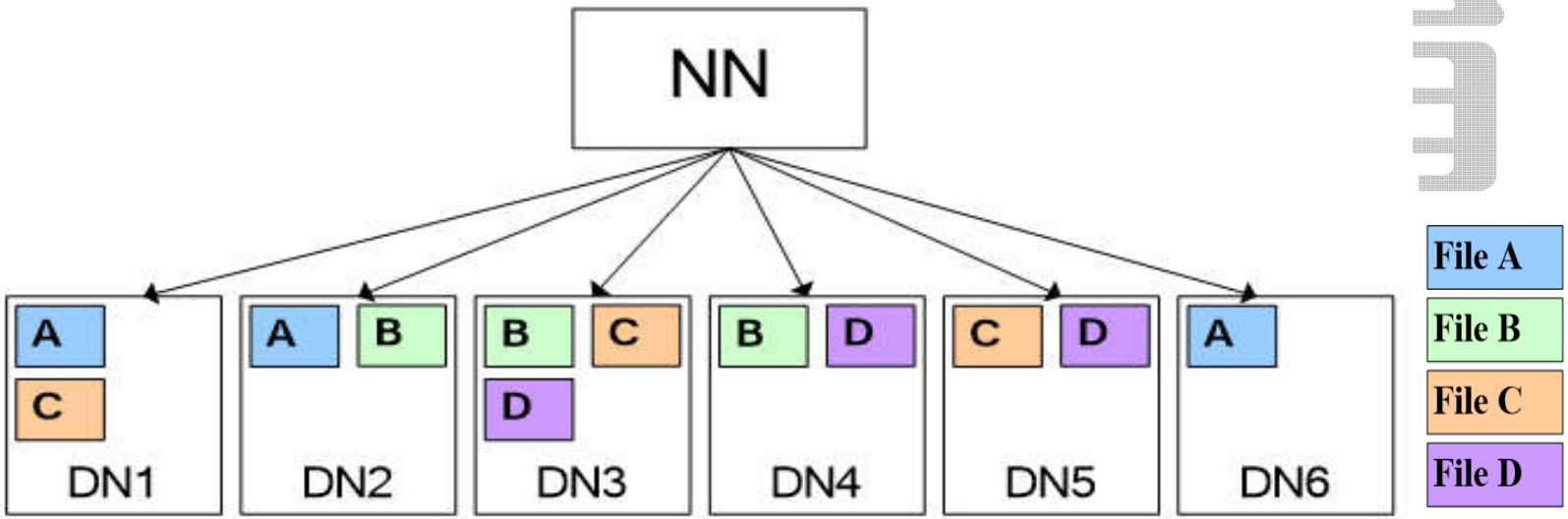


配置HDFS数据存储策略 - 节点组存储



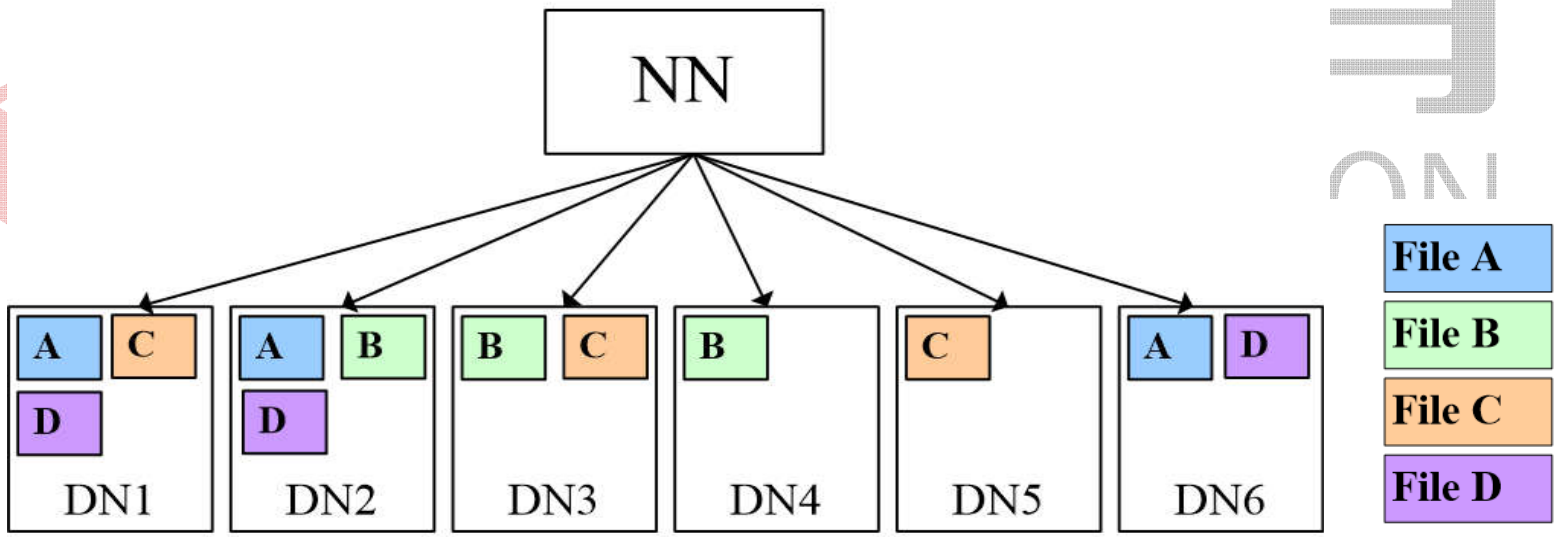
Colocation同分布

- 同分布(Colocation)的定义：将存在关联关系的数据或可能要进行关联操作的数据存储在相同的存储节点上。
- 按照下图存放，假设要将文件A和文件D进行关联操作，此时不可避免地要进行大量的数据搬迁，整个集群将由于数据传输占据大量网络带宽，严重



Colocation同分布效果图

- HDFS文件同分布的特性，将那些需进行关联操作的文件存放在相同数据节点上，在进行关联操作计算时避免了到其他的数据节点上获取数据，大大降低网络带宽的占用。
- 使用同分布特性，文件A、D进行join时，由于其对应的block都在相同节点，因此大大降低资源消耗。



HDFS数据完整性保障

- HDFS主要目的是保证存储数据完整性，对于各组件的失效，做了可靠性处理。
- 重建失效数据盘的副本数据
  - DataNode向NameNode周期上报失败时，NameNode发起副本重建动作以恢复丢失副本。
- 集群数据均衡
  - HDFS架构设计了数据均衡机制，此机制保证数据在各个DataNode上分布是平均的。
- 元数据可靠性保证
  - 采用日志机制操作元数据，同时元数据存放在主备NameNode上。
  - 快照机制实现了文件系统常见的快照机制，保证数据误操作时，能及时恢复。
- 安全模式
  - HDFS提供独有安全模式机制，在数据节点故障，硬盘故障时，能防止故障扩散。

HDFS架构其他关键设计要点说明

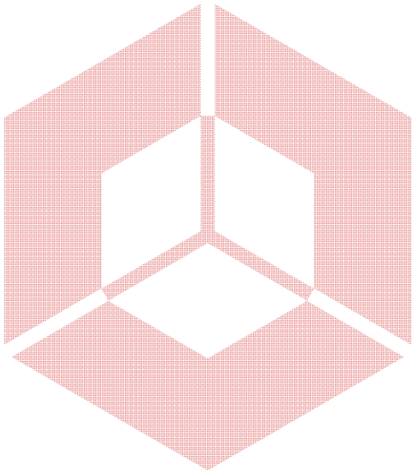
- 统一的文件系统：
  - HDFS对外仅呈现一个统一的文件系统。
- 空间回收机制：
  - 支持回收站机制，以及副本数的动态设置机制。
- 数据组织：
  - 数据存储以数据块为单位，存储在操作系统的HDFS文件系统上。
- 访问方式：
  - 提供JAVA API，HTTP方式，SHELL方式访问HDFS数据。

常用shell命令

命令类别	命令	命令说明
dfs	-cat	显示文件内容
	-ls	显示目录列表
	-rm	删除文件
	-put	上传目录/文件到HDFS
	-get	从HDFS下载目录/文件到本地
	-mkdir	创建目录
	-chmod/-chown	改变文件属组
	…	…
dfsadmin	-safemode	安全模式操作
	-report	报告服务状态

本章总结

- 本章对HDFS概念及应用场景进行了介绍，然后阐述了HDFS系统架构原理及其关键特性。

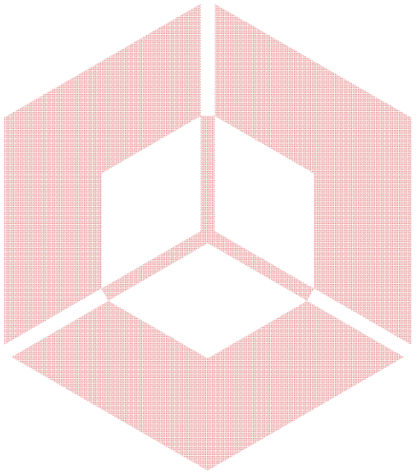


思考题

1. HDFS是什么，适合于做什么？
2. HDFS包含哪些角色？
3. 请简述HDFS的读写流程。

更多信息

- 下载培训资料：
  - http://support.huawei.com/learning/trainFaceDetailAction?lang=zh&pbiPath=term1000025185&courseId=Node1000009072
- eLearning课程：
  - http://support.huawei.com/learning/nodeQueryAction!loadTrainProjectInfo?lang=zh&pbiPath=term1000025185&courseId=Node1000009421&navId=MW000001
- 考试大纲：
  - http://support.huawei.com/learning/Certificate!toExamOutlineDetail?lang=zh&nodeId=Node1000003516
- 模拟考试：
  - http://support.huawei.com/learning/Certificate!toSimExamDetail?lang=zh&nodeId=Node1000004285
- 认证流程：
  - http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_40



谢谢

www.huawei.com