HDFS技术原理

www.huawei.com

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

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

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

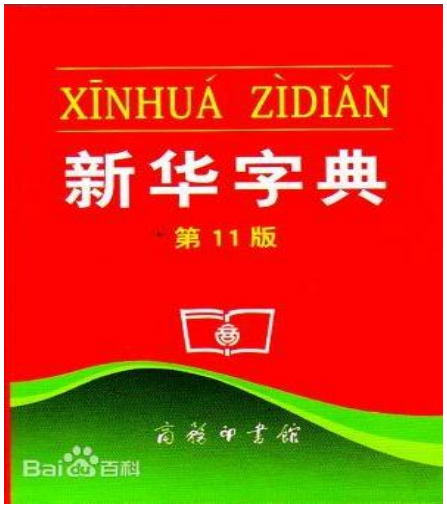
HUAWEI

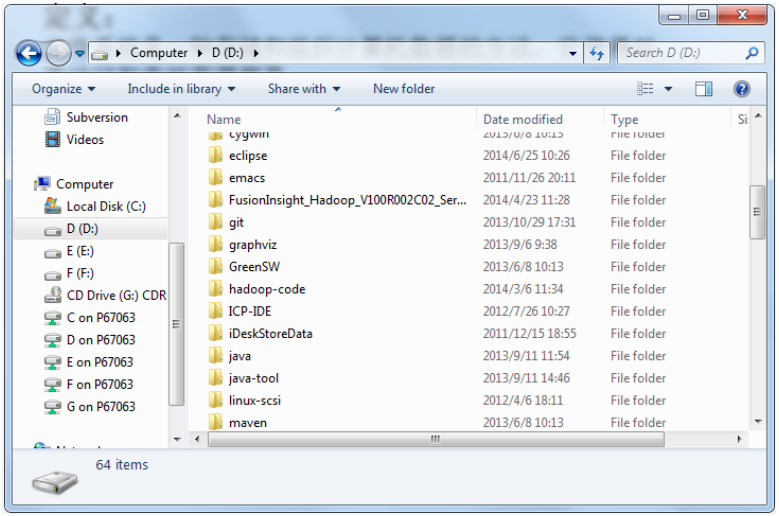
目录

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

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**）

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

HDFS概述

- **HDFS(Hadoop Distributed File System)**基于**Google**发布的**GFS**论文设计开发，运行在通用硬件上的分布式文件系统。
- 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、流式数据访问

HDFS不适合做什么？
大量小文件、随机写入、低延迟读取

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

HDFS应用场景举例

HDFS是**Hadoop**技术框架中的分布式文件系统，对部署在多台独立物理机器上的文件进行管理。

可应用于以下几种场景：

- 网站用户行为数据存储
- 生态系统数据存储
- 气象数据存储

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

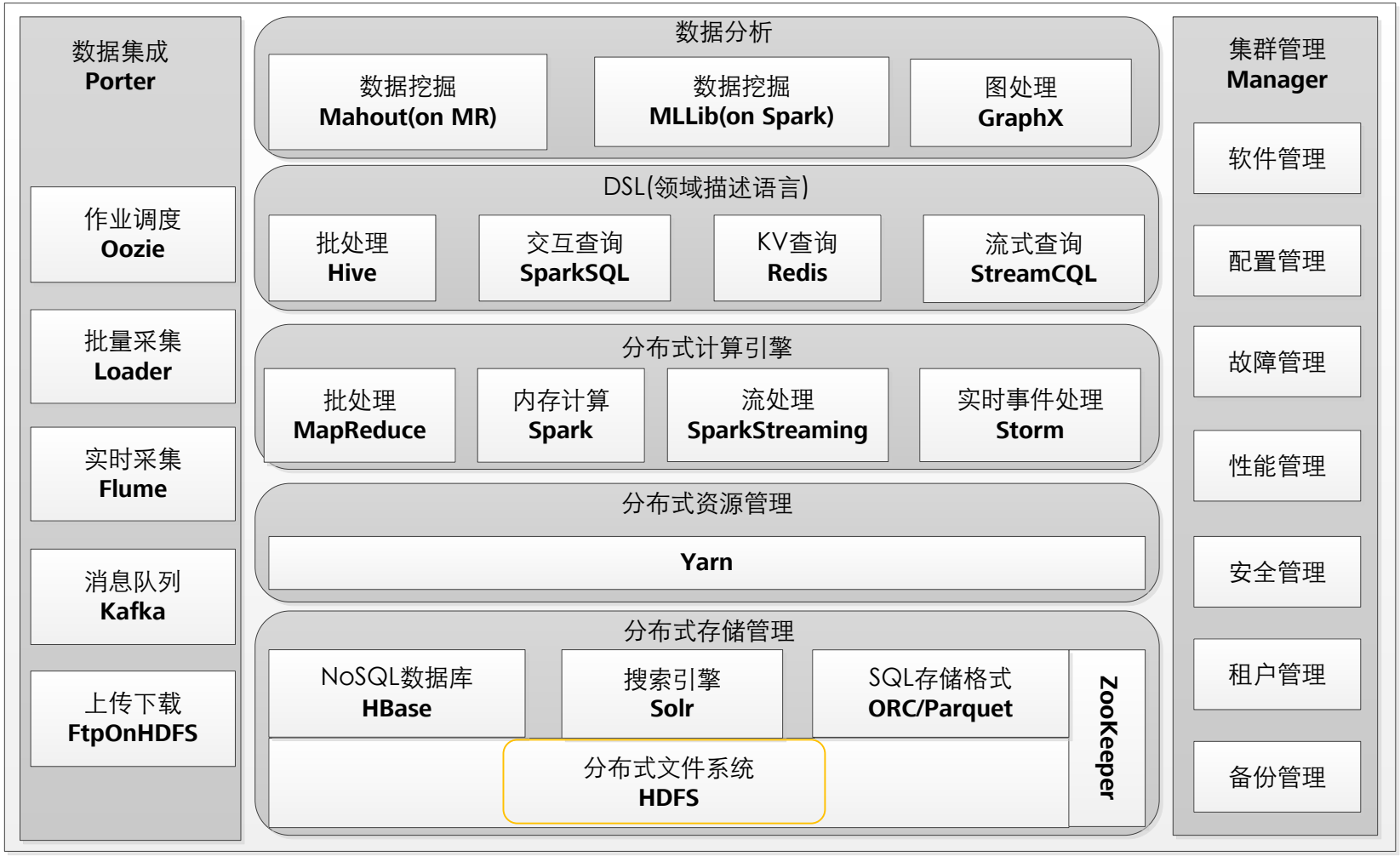
HUAWEI

目录

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

HDFS在FusionInsight产品的位置



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

目录

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

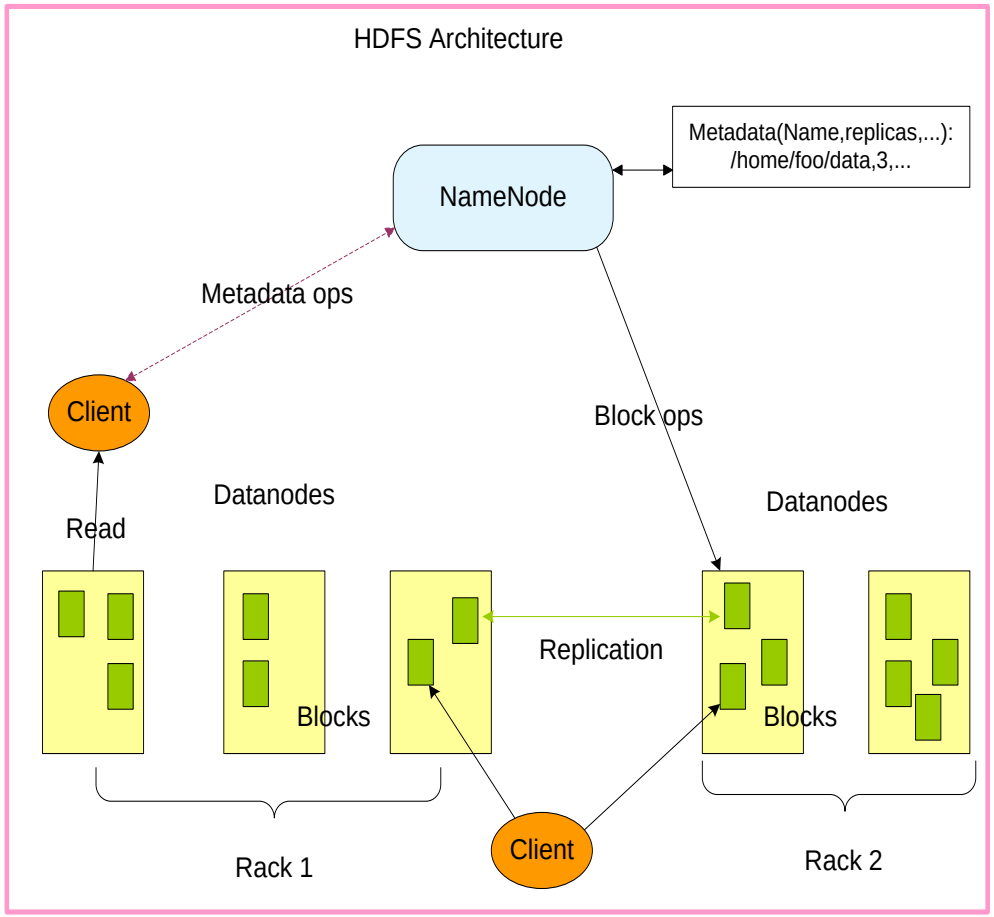
系统设计目标

硬件失效	流式数据访问	存储数据较大	数据一致性	多硬件平台	移动计算能力
• 硬件的异常比软件的异常更加常见。 • 对于有上百台服务器的数据中心来说，认为总有服务器异常，硬件异常是常态。 • **HDFS**需要监测这些异常，并自动恢复数据。	• 基于**HDFS**的应用仅采用流式方式读数据。 • 运行在**HDFS**上的应用并非以通用业务为目的的应用程序。 • 应用程序关注的是吞吐量，而非响应时间。 • 非**POSIX**标准接口的数据访问。	• 运行在**HDFS**的应用程序有较大的数据需要处理。 • 典型的文件大小为**GB**到**TB**级别。	• 应用程序采用**WORM**（**Write Once Read Many**）的数据读写模型。 • 文件仅支持追加，而不允许修改。	• **HDFS**可运行在不同的硬件平台上。	• 计算和存储采用就近原则，计算离数据最近。 • 就近原则将有效减少网络的负载，降低网络拥塞。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

基本系统架构



HDFS架构包含三个部分：

NameNode，DataNode，Client

- **NameNode**：**NameNode**用于存储、生成文件系统的元数据。运行一个实例。
- **DataNode**：**DataNode**用于存储实际的数据，将自己管理的数据块上报给**NameNode** ，运行多个实例。
- **Client**：支持业务访问**HDFS**，从**NameNode ,DataNode**获取数据返回给业务。多个实例，和业务一起运行。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

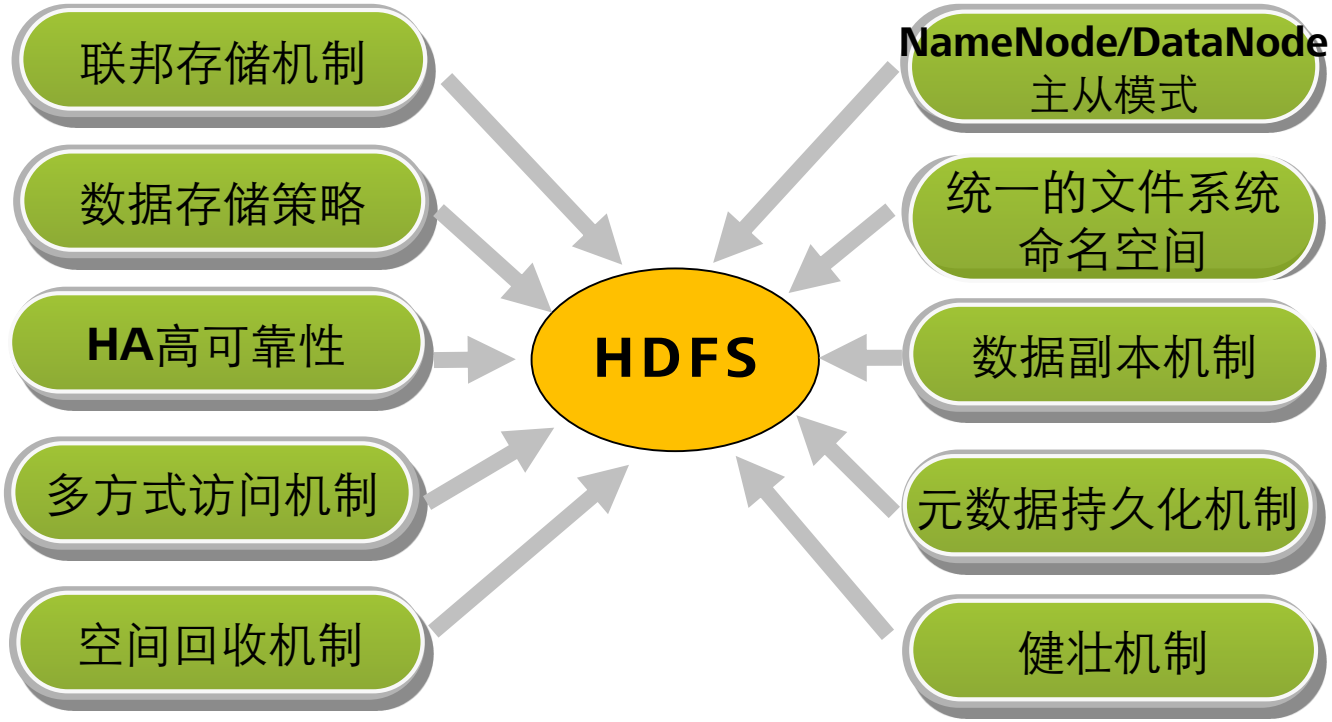
HUAWEI

目录

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

HDFS架构关键设计



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HDFS数据读取流程

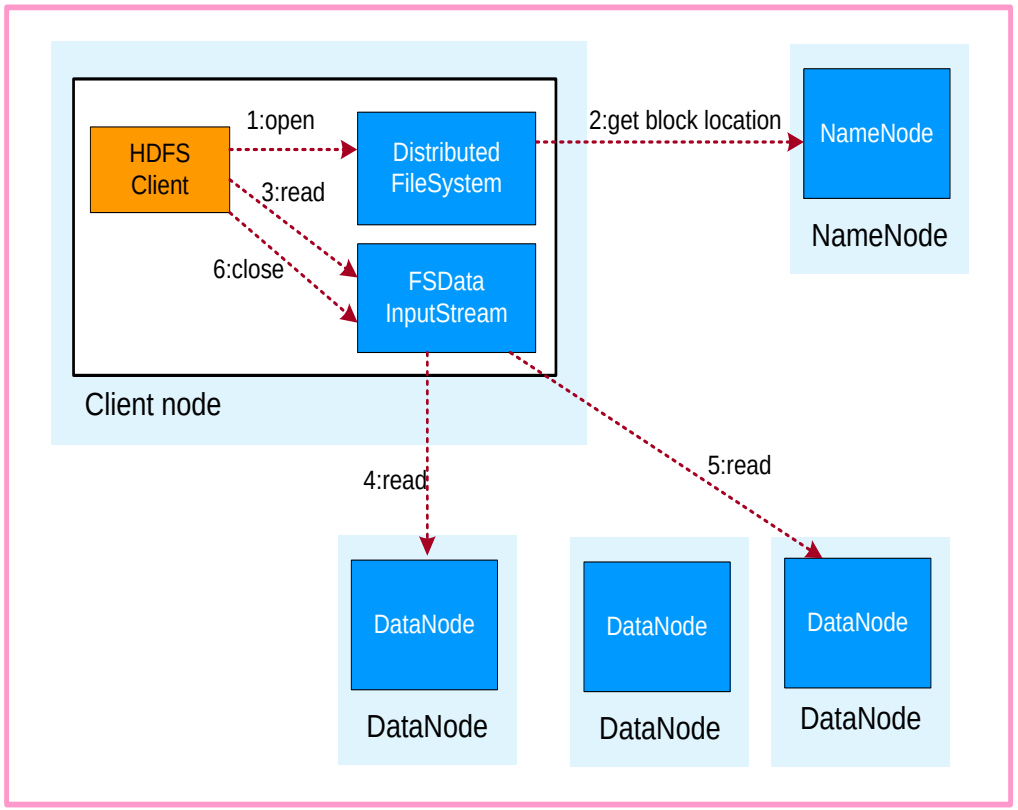
HDFS数据读取流程如下：

1. 业务应用调用**HDFS Client**提供的**API**打开文件。

2. HDFS Client联系**NameNode**，获取到文件信息（数据块、**DataNode**位置信息）。

3. 业务应用调用**read API**读取文件。

4. HDFS Client根据从**NameNode**获取到的信息，联系**DataNode**，获取相应的数据块。**(Client**采用就近原则读取数据**)**。



5. HDFS Client会与多个**DataNode**通讯获取数据块。

6. 数据读取完成后，业务调用**close**关闭连接。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HDFS数据写入流程

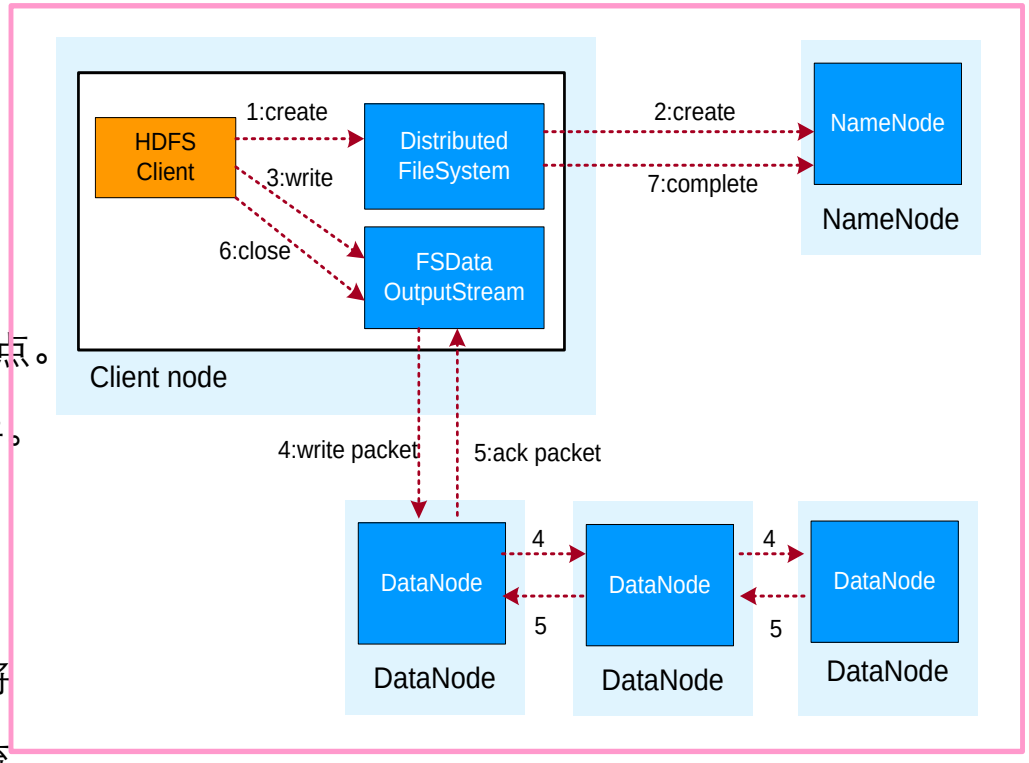
HDFS数据写入流程如下：

1. 业务应用调用**HDFS Client**提供的**API**创建文件，请求写入。

2. HDFS Client联系**NameNode**，**NameNode**在元数据中创建文件节点。

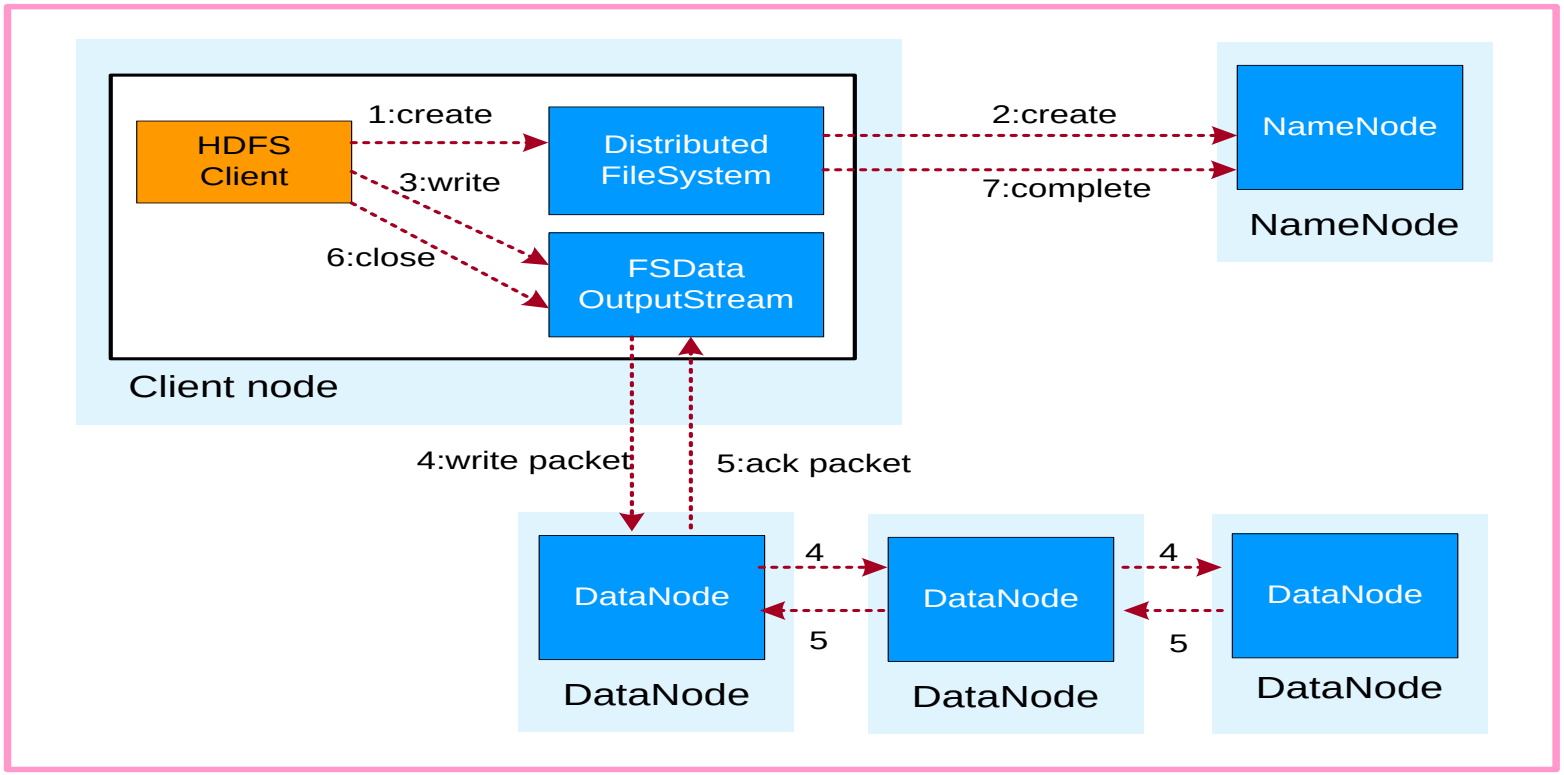
3. 业务应用调用**write API**写入文件。

4. HDFS Client收到业务数据后，从**NameNode**获取到数据块编号、位置信息后，联系**DataNode**，并将需要写入数据的**DataNode**建立起流水线，完成后，客户端再通过自有协议写入数据到**DataNode1**，再由**DataNode1**复制到**DataNode2, DataNode3**。



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HDFS数据写入流程（续）



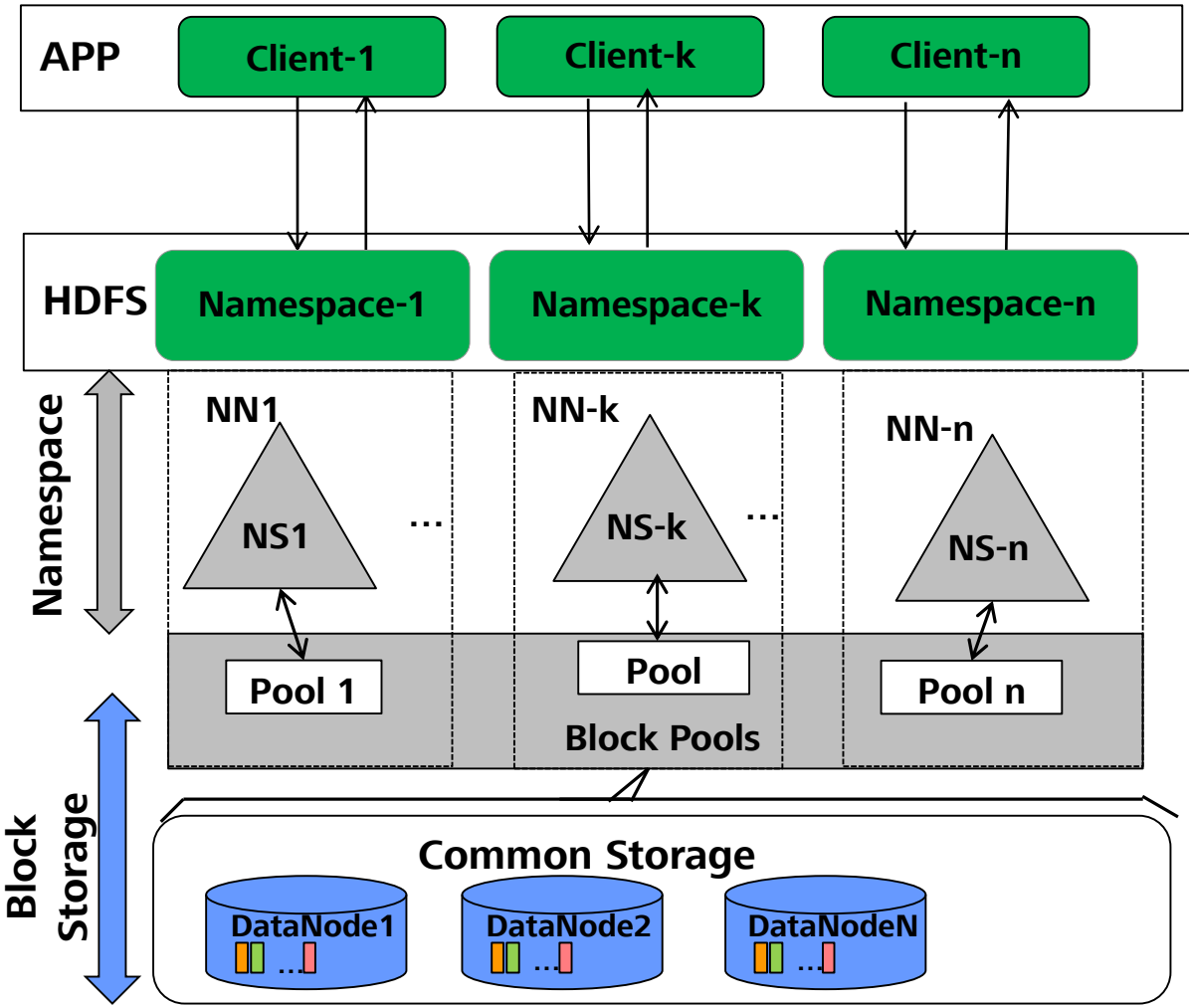
5. 写完的数据，将返回确认信息给**HDFS Client**。

6. 所有数据确认完成后，业务调用**HDFS Client**关闭文件。

7. 业务调用**close,flush**后**HDFS Client**联系**NameNode**，确认数据写完成，**NameNode**持久化元数据。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HDFS联邦（Federation）



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

数据副本机制

副本距离计算公式：

- **Distance(Rack1/D1, Rack1/D1)=0**

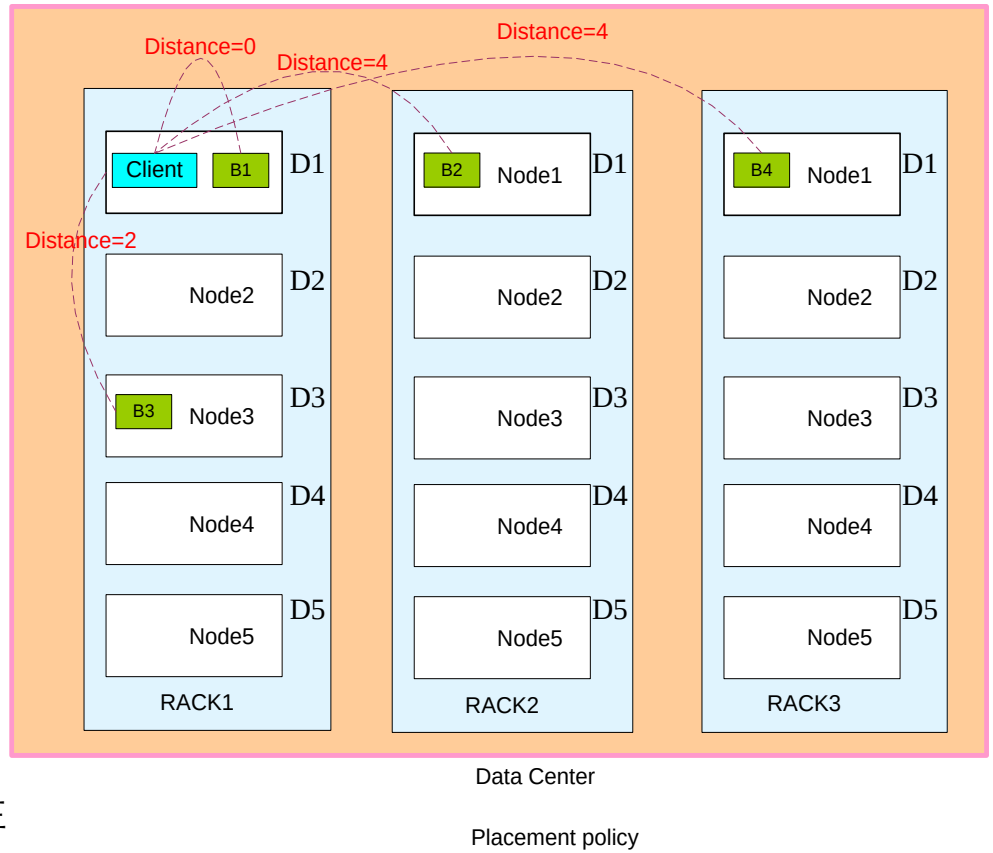
  同一台服务器的距离为**0**

- **Distance(Rack1/D1, Rack1/D3)=2**

  同一机架不同的服务器距离为**2**

- **Distance(Rack1/D1, Rack2/D1)=4**

  不同机架的服务器距离为**4**



副本放置策略：

- 第一个副本在本地机器
- 第二个副本在远端机架的节点
- 第三个副本看之前的两个副本是否在同一机架，如果是则选择其他机架，否则选择和第一个副本相同机架的不同节点，第四个及以上，随机选择副本存放位置。

Placement policy

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

元数据持久化

元数据持久化的流程如下：

1.备**NameNode**通知主**NameNode**生成新的日志文件，以后的日志写到**Editlog.new**中，并获取旧的**Editlog**。

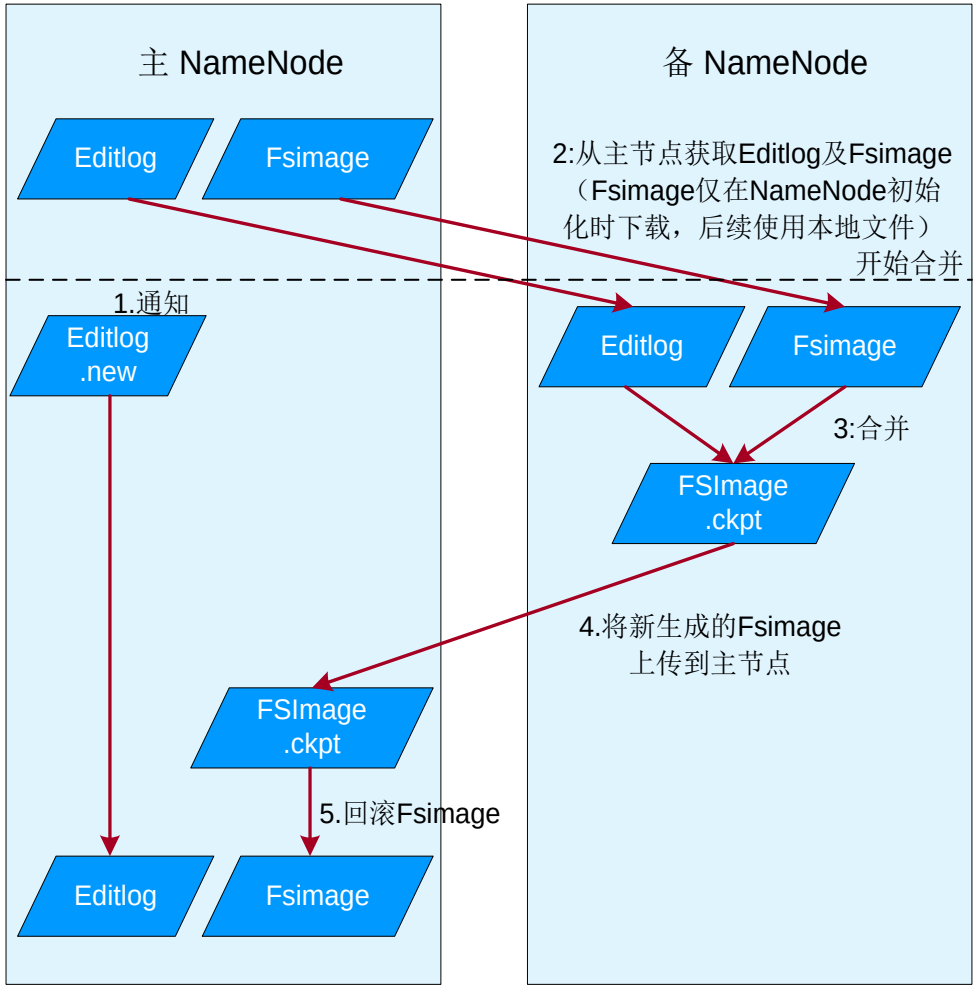
2.备**NameNode**从主**NameNode**上获取**FSImage**文件及旧的**EditLog**。

3.备**NameNode**将日志和旧的元数据合并，生成新的元数据**FSImage.ckpt**。

4.备**NameNode**将元数据上传到主**NameNode**。

5.主**NameNode**将上传的元数据进行回滚。

6.循环步骤**1**。



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

元数据持久化健壮机制

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。
- 数据有效性保证
  - **DataNode**数据在读取时校验失败，则从其他数据节点读取数据。
- 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。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

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。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

配置HDFS数据存储策略-分级存储

配置**DataNode**使用分级存储：

HDFS的异构分级存储框架提供了**RAM_DISK**（内存虚拟硬盘）、**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>

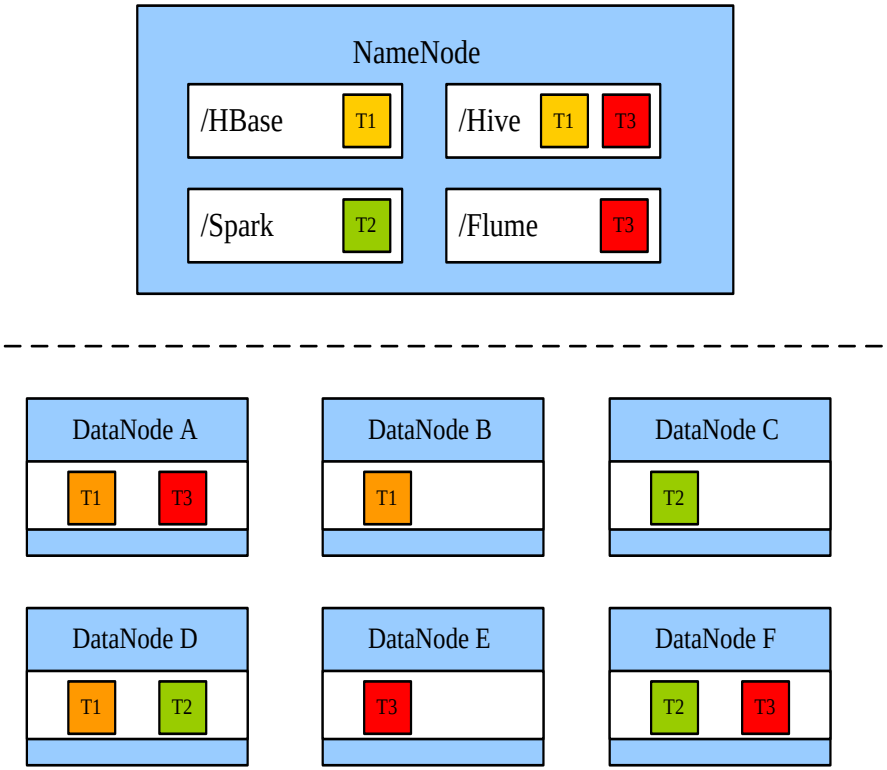
Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

配置HDFS数据存储策略-标签存储

配置**DataNode**使用标签存储：

用户通过数据特征灵活配置**HDFS**数据块摆放策略，即为一个**HDFS**目录设置一个标签表达式，每个**DataNode**可以对应一个或多个标签；当基于标签的数据块摆放策略为指定目录下的文件选择**DataNode**节点进行存放时，根据文件的标签表达式选择出将要存放的**DataNode**节点范围，然后在这个**DataNode**节点范围内，遵守下一个指定的数据块摆放策略进行存放。



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

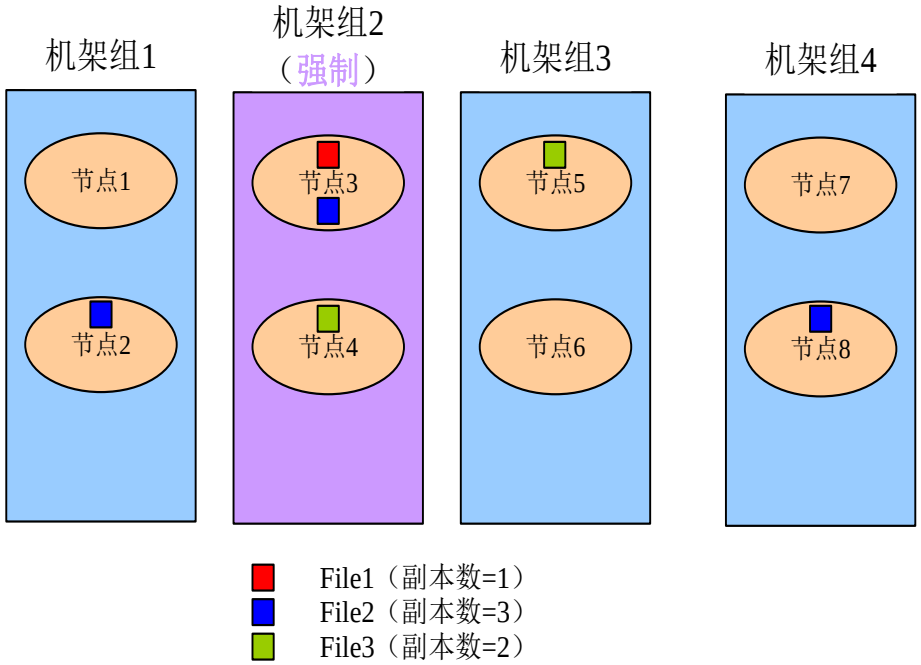
配置HDFS数据存储策略-节点组存储

配置DataNode使用节点组存储：

关键数据根据实际业务需要保存在具有高度可靠性的节点中，此时**DataNode**组成了异构集群。通过修改**DataNode**的存储策略，系统可以将数据强制保存在指定的节点组中。

使用约束：

- 第一份副本将从强制机架组（机架组**2**）中选出，如果在强制机架组中没有可用节点，则写入失败。
- 第二份副本将从本地客户端机器或机架组中的随机节点中**(**当客户端机器机架组不为强制机架组时**)**选出。
- 第三份副本将从其他机架组中选出。
- 各副本应存放在不同的机架组中。

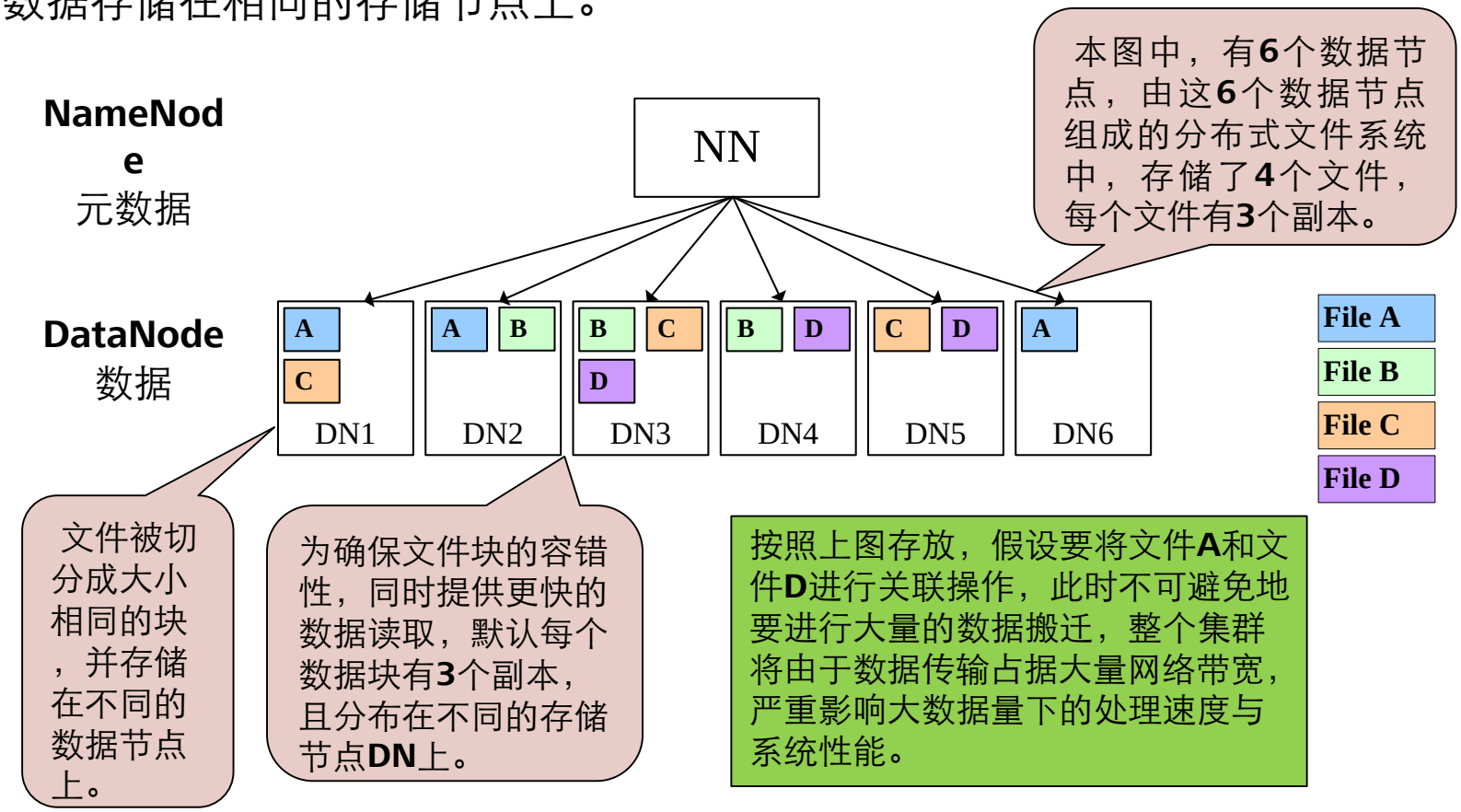


如果所需副本的数量大于可用的机架组数量，则会将多出的副本存放在随机机架组中。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

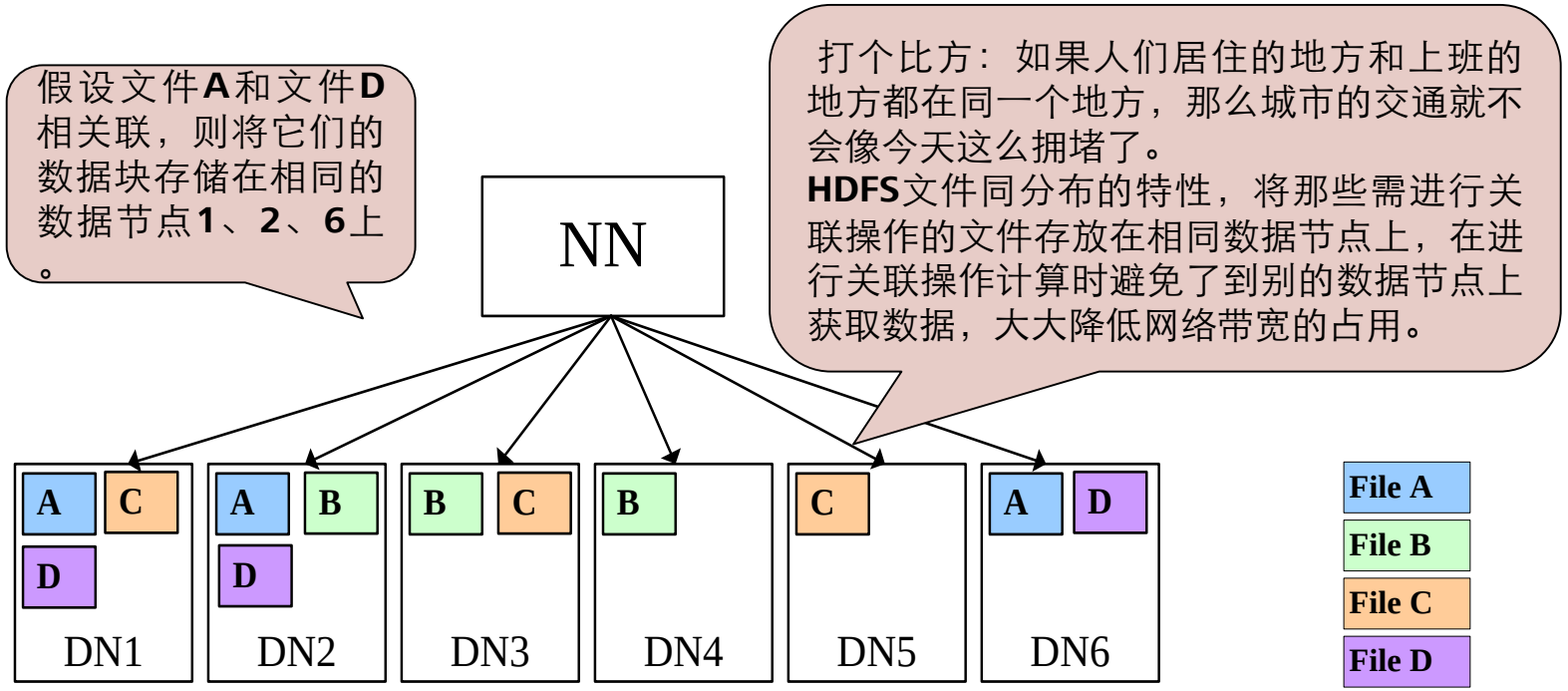
Colocation同分布

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。



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

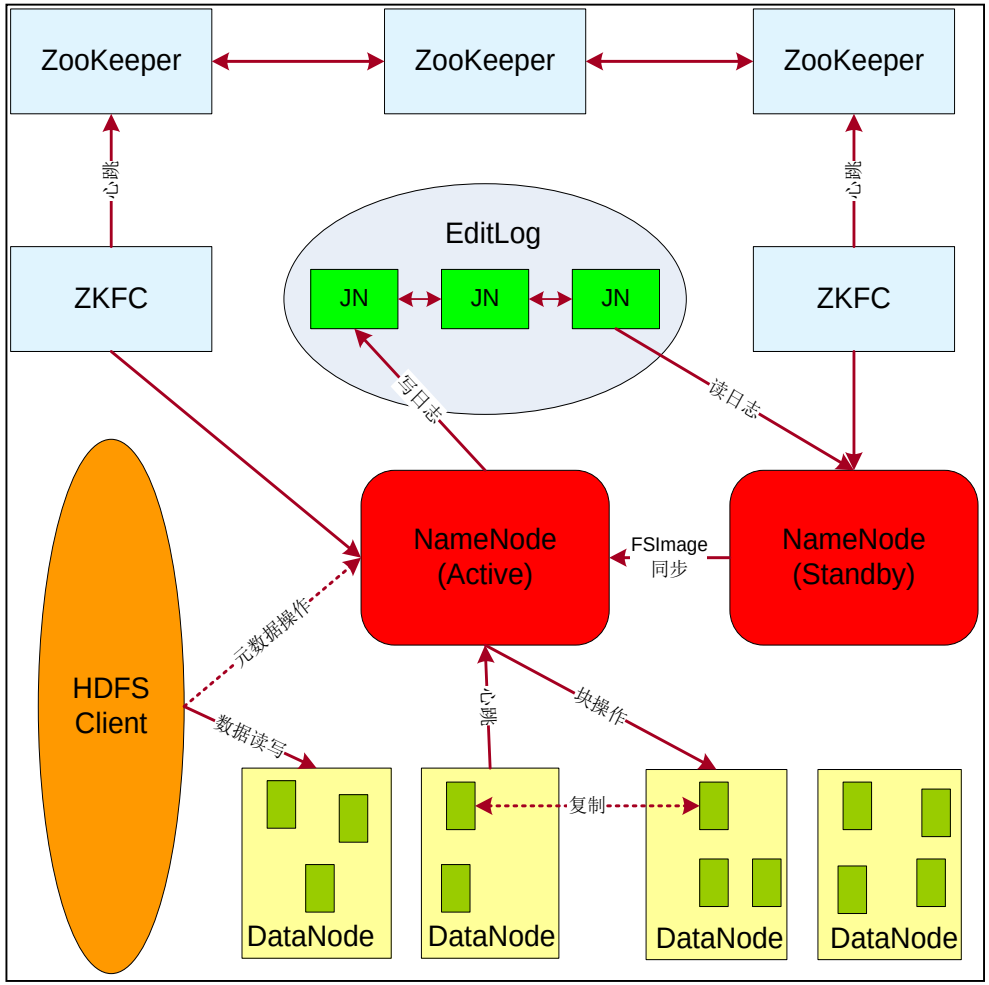
Colocation同分布效果图



Hadoop 实现文件级同分布，即存在相关联的多个文件的所有块都分布在同一存储节点上。文件级同分布实现文件的快速访问，避免了因数据搬迁带来的大量网络开销。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HDFS高可靠性



HDFS的高可靠性（HA）架构在基本架构上增加了以下组件：

- **ZooKeeper**
  分布式协调，主要用来存储**HA**下的状态文件，主备信息。**ZK**个数建议**3**个及以上且为奇数个。
- **NameNode**主备
  NameNode主备模式，主提供服务，备合并元数据并作为主的热备。
- **ZKFC**
  ZKFC(ZooKeeper Failover Controller)用于控制**NameNode**节点的主备状态。
- **JN**
  JN(JournalNode)用于共享存储**NameNode**生成的**Editlog**。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

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。
- 统一的通讯协议：
  - 统一采用**RPC**方式通信。**NameNode**被动的接收**Client, DataNode**的**RPC**请求。
- 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。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

HDFS支持接口

接口类别	接口举例	接口说明
JAVA	mkdirs(Path f)	通过该接口可在**HDFS**上创建文件夹，其中**f**为文件夹的完整路径。
	create(Path f)	通过该接口可在**HDFS**上创建文件，其中**f**为文件的完整路径。
HTTP	curl -L --negotiate -u : "http://<HOST>:<PORT>/webhdfs/v1/<PATH>?op=OPEN"	打开并读取**HDFS**文件内容。
	curl -i -X DELETE --negotiate -u : "http://<host>:<port>/webhdfs/v1/<path>?op=DELETE"	删除指定的文件。
SHELL	hdfs dfs [COMMAND [COMMAND_OPTIONS]]	在**HDFS**文件系统上运行**filesystem**命令。
	hdfs fsck <path> [COMMAND_OPTIONS]	运行**HDFS**文件系统检查工具。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

总结

- **HDFS**概述及应用场景
- **HDFS**在**FusionInsight**产品的位置
- **HDFS**系统架构
- 关键特性介绍

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

思考题

1. **HDFS**是什么样的系统，适合于做什么？
2. **HDFS**的设计目标是什么？
3. **HDFS**包含哪些组件？
4. 请简述**HDFS**的读写流程。
5. **HDFS**元数据是如何持久化的？

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

习题

1. HDFS是基于流数据模式访问和处理超大文件的需求而开发的，具有高容错、高可靠性、高可扩展性、高吞吐率等特征，适合的读写任务是（ ）。

A．一次写入，少次读　　B．多次写入，少次读

C．多次写入，多次读　　D．一次写入，多次读

2. 以下对于HDFS描述不正确的是（ ）

A. HDFS是一个使用java编写的分布式文件系统。

B. HDFS由NameNode，DataNode，client组成。

C. HDFS不支持标准的POSIX文件系统接口。

D. HDFS支持对已有数据进行修改。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

习题

3. HDFS不适合的功能有（ ）

A. 多副本方式存储数据。

B. 存储TB-PB级别的数据。

C. 文件的随机写入。

D. 硬件故障的容错处理。

4. （多选）有关HDFS说法正确的有（ ）

A. HDFS不适合存储大量小文件。

B. HDFS不适合有低延迟数据访问要求的业务。

C. HDFS适合流式数据的访问。

D. 基于HDFS的应用应该使用WORM的数据读写模型编程。

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

学习推荐

- 华为**Learning**网站
  - **http://support.huawei.com/learning/Index!toTrainIndex**
- 华为**Support**案例库
  - **http://support.huawei.com/enterprise/servicecenter?lang=zh**
- 华为大数据论坛
  - **http://support.huawei.com/ecommunity/bbs/list_1069,1420018021.html?l=zh**

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Thank you

www.huawei.com