大数据开发技术

东北林业大学

卢洋

第七章

Hadoop 2.X 新特性

1. 集群间数据拷贝；

2. 小文件存档；

3. 回收站；

4. 快照管理.

7.2

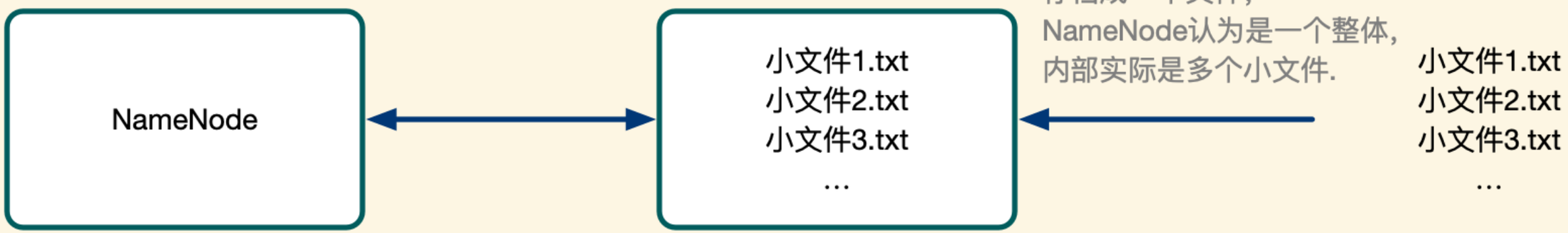
小文件存档

1 HDFS存储小文件的弊端

- 每个文件均按块存储，每个块的元数据存储在NameNode的内存中，因此HDFS存储小文件会非常低效；
- 大量的小文件会耗尽NameNode中的大部分内存；
- 存储小文件所需要的磁盘容量和数据块的大小无关；
- 例如：1MB的文件设置为128MB的块存储，实际使用的是1MB的磁盘空间，而不是128MB.

2 解决存储小文件方法之一

- HDFS存档文件或HAR文件，是一个更高效的文件存档工具；
- 它将文件存入HDFS块，在减少NameNode内存使用的同时，允许对文件进行透明的访问；
- 具体来说，HDFS存档文件对内还是一个一个独立的文件，对NameNode而言却是一个整体，减少了NameNode的内存.


NameNode
小文件1.txt
小文件2.txt
小文件3.txt
…
存档成一个文件，
NameNode认为是一个整体，
内部实际是多个小文件.
小文件1.txt
小文件2.txt
小文件3.txt
…

案例

1. 需要启动YARN进程
   sbin/start-yarn.sh

2. 归档文件
   将/usr/root/input目录下的数据归档成名为input.har的归档文件，并把归档后的文件存储到/usr/root/output路径下.
   bin/hadoop archive -archiveName input.har -p /user/root/input /user/root/output

案例

3. 查看归档

```
hadoop fs -ls -R /user/root/output/input.har
hadoop fs -ls -R har:///user/root/output/
input.har
```

4. 解归档文件

```
hadoop fs -cp har:///user/root/output/
input.har/* /usr/root
```