

HDFS应用实践

——HDFS的基本操作



联创中控（北京）科技有限公司

本节要点



大纲

1 / HDFS命令行操作

2 / HDFS的Web界面

3 / HDFS命令详解

HDFS的命令行操作

命令行接口是HDFS的交互方式

- 简单直观
- 便于使用
- 可以进行一些基本操作



HDFS的命令行操作

1、列出HDFS文件

通过 `-ls` 命令列出HDFS下的文件：

```
# hadoop fs -ls
```

在HDFS中，没有当前目录这样一个概念，也没用`cd`这个命令。

2、列出HDFS目录下某个文档中的文件

通过 “`-ls 文件名`” 命令浏览HDFS下名为`master1-file`的目录中的文件：

```
# hadoop fs -ls /mater1-file
```



HDFS的命令行操作

3、上传文件到HDFS

通过“-put 文件1 文件2”命令将“/usr/hadoop/hadoop”下的data文件上传到HDFS上并重命名为test:

```
# hadoop fs -put /usr/hadoop/hadoop/data /test
```

4、将HDFS中文件复制到本地系统中

通过“-get 文件1 文件2”命令将HDFS中的“master1-file”文件复制本地系统并命名为“getout”:

```
# hadoop fs -get /master1-file getout
```



HDFS的命令行操作

5、查看HDFS下某个文件

通过“-cat 文件”命令查看master1-file文件中的内容：

```
# hadoop fs -cat /master1-file
```

6、删除HDFS下的文档

通过“-rmr 文件”命令删除HDFS下的“master1-file”文档。

```
# hadoop fs -rmr /master1-file
```



本节要点



大纲

1 / HDFS命令行操作

2 / HDFS的Web界面

3 / HDFS命令详解

HDFS的web界面

http : //localhost:50070

HDFS的 Web界 面提供的 功能

01

提供了基本的文件系统信息，其中包括集群启动时间、版本号、编译时间及是否又升级。

02

提供了文件系统的基本功能：Browse the filesystem，增加了对文件系统的可读性。

03

可以直接通过Web界面访问文件内容。

本节要点



大纲

1 / HDFS命令行操作

2 / HDFS的Web界面

3 / HDFS命令详解

HDFS的命令详解

Hadoop提供了一组shell命令在命令行终端对Hadoop进行操作。这些操作包括诸如格式化文件系统、上传和下载文件、启动DataNode、查看文件系统使用情况、运行JAR包等几乎所有和Hadoop相关的操作。



Distcp命令介绍

Distcp命令的引入:

HDFS提供了一个非常实用的程序——`distcp`，用来在Hadoop文件系统中并行地复制大量文件。`distcp`一般适用于在两个HDFS集群间传送数据的情况。

将master1下的master1-file文件拷贝到master2-file文件下

```
hadoop distcp hdfs://master1/master1-file hdfs://master2/master2-file
```



Distcp命令介绍

如果尝试使用distcp进行HDFS集群间的复制，使用HDFS模式之后，HDFS运行在不同的Hadoop版本之上，复制将会因为RPC系统的不匹配而失败。



可以使用基于HTTP的HFTP进行访问

```
hadoop distcp hftp://master1:50070/master1-file  
hdfs://master2/ master2-file
```


HDFS文件系统平衡设置

问题出现:

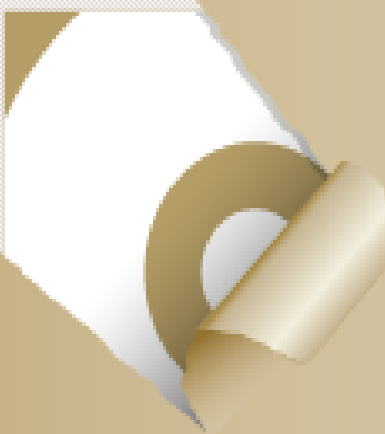
当复制大规模数据到HDFS时, 要考虑的一个重要因素是文件系统的平衡。

问题解决:

可以通过设置更多的Map任务来避免不平衡情况的发生。
HDFS提供了一个工具balancer来改变集群中的文件块存储的平衡。

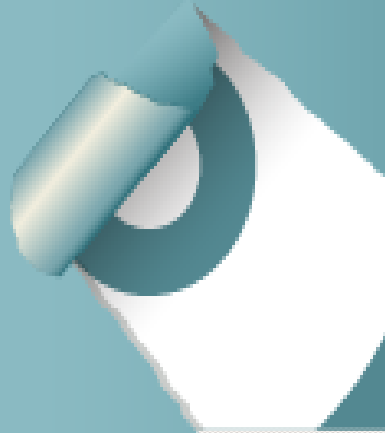


使用Hadoop归档文件



文件块的元数据信息会被存在NameNode的内存中，对HDFS来说，大规模存储小文件显然是低效的，很多小文件会耗尽NameNode的大部分内存。

Hadoop归档文件和HAR文件可以将文件高效地放入HDFS块中的文件存档设备，Hadoop归档文件是通过archive命令工具根据文件集合创建的。



使用Hadoop归档文件示例

```
hadoop archive -archiveName 参数1 -p 参数2 参数3
```

第一个参数是归档文件的名称，这里是file.har文件；第二个参数是要归档的文件源，这里只归档一个源文件夹；最后一个参数是HAR文件的输出目录。

```
#hadoop archive -archiveName files.har -p /master1-file /
```



HAR文件的不足

当创建一个归档文件时，还会创建原始文件的一个副本，这样就需要额外的磁盘空间（尽管归档完成后会删除原始文件）。

归档文件一旦创建就不能改变，要增加或删除文件，就要重新创建。

HAR文件可以作为MapReduce的一个输入文件，没有一个基于归档的InputFormat可以将多个文件打包到一个单一的MapReduce中去。

本节要点



大纲

1 / HDFS命令行操作

2 / HDFS的Web界面

3 / HDFS命令详解

HDFS的其他命令总结

- NameNode -format : 格式化DFS文件系统
- SecondaryNameNode : 运行DFS的
- SecondaryNameNode进程
- NameNode : 运行DFS的NameNode进程
- DataNode : 运行DFS的DataNode进程
- dfsadmin : 运行DFS的管理客户端
- mradmin : 运行MapReduce的管理客户端
- fsck : 运行HDFS的检测进程

- fs : 运行一个文件系统工具
- balancer : 运行一个文件系统平衡进程
- jobtracker:运行一个JobTracker进程
- job : 管理运行中的MapReduce任务
- queue : 获得运行中的MapReduce队列的信息
- version : 打印版本号
- jar : 运行一个jar文件
- daemonlog : 读取/设置守护进程的日志记录级别



HDFS的其他命令总结

Hadoop相关命令的统一格式如下：

hadoop command [genericOptions] [commandOptions]

其中只有dfsadmin、fsck、fs具有选项genericOptions及commandOptions，其余的命令只有commandOptions。



只有commandOptions选项的命令

- distcp。Distcp命令用于Distcp分布式复制。用于在集群内部及集群之间复制数据。
- archives。Archives命令是Hadoop定义的档案格式。archive对应一个文件系统，它的扩展名是har，包含元数据及数据文件。
- 使用如下命令将Hadoop回滚到前一个版本，它的用法如下：

hadoop DataNode [-rollback]



只有commandOptions选项的命令

nameNode命令稍微复杂一些，它的用法如下：

hadoop nameNode

[-format] //格式化NameNode

[-upgrade] //在hadoop升级后，应该使用这个命令启动NameNode

[-rollback] //s使用NameNode回滚前一个版本

[-finalize] //删除文件系统的前一个状态，这会导致系统不能回滚到前一个状态

[-importCheckpoint]//复制备份checkpoint的状态到当前checkpoint



只有commandOptions选项的命令

sencondaryNameNode的命令用法如下：

hadoop secondaryNameNode

[-checkpoint [force]] //当editlog超过规定大小（默认64MB）时，启动检查secondaryNameNode的checkpoint过程；如果启用force选项，则强制执行checkpoint过程。

[-geteditsize] //在终端上显示editlog文件的大小



只有commandOptions选项的命令

当集群中添加新的DataNode时，可以使用balancer这个命令来进行负载均衡。
其用法如下：

```
hadoop balancer
```



有genericOptions选项的命令

dfsadmin。在dfsadmin命令中可以执行一些类似Windows中高级用户才能执行的命令，比如升级、回滚等。其用法如下：

hadoop dfsadmin [CENERIC_OPTIONS]

[-report] //在终端上显示文件系统的基本信息

[-safemode enter | leave | get | wait]//Hadoop的安全模式及相关维护；在安全模式中系统是只读的，数据块也不可以删除或复制

[-refreshNodes] [-finalizeUpgrade] //重新读取hosts和exclude文件，将新的被云迹加入到集群中的DataNode连入，同时断开与那些从集群出去的DataNode的选择

[-upgradeProgress status | details | force] //获得当前系统的升级状态、细节，或者强制执行升级过程

[-metasave filename]//保存NameNode的主要数据结构到指定目录下

[-setQuota <quota> <dirname> ...<dirname>]//为每个目录设定配额

[-setSpaceQuota <quota> <dirname> ...<dirname>]//为每个目录设置配额空间

[-clrSpaceQuota <diraname> ...<dirname>]//清除这些目录的配额空间

[-help [cmd]]//显示命令的帮助信息



有genericOptions选项的命令

fsck。fsck在HDFS中被用来检查系统中的不一致情况。与Linux不同，这个命令只能用于检查，不能进行修复。其使用方法如下：

```
hadoop fsck [GENERIC_OPTIONS] <path> [-move|-delete | -  
openforwrite] [-files [-blocks [-locations | -racks]]]
```

//<path> 检查的起始目录

//-move 移动受损文件到/lost+found

//-delete 删除受损文件

//-openforwrite 在终端上显示被写打开的文件

//-files 在终端上显示正在检查的文件

//-blocks 在中断上显示块信息

//-location 在终端上显示每个块的位置

//-rack 显示DataNode的网络拓扑结构图



有genericOptions选项的命令

fs是HDFS最常用的命令。可以使用这些命令查看HDFS上的目录结构、文件、上传和下载文件、创建文件夹、复制文件等、使用方法如下：

hadoop fs [genericOptions]



有genericOptions选项的命令

fs命令参数

`[-ls <path>]` //显示目标路径当期目录下的所有文件

`[-lsr <path>]` //递归显示目录路径下的所有目录及文件（深度优先）

`[-du <path>]` //以字节为单位显示目录中所有文件的大小，或该文件的大小（如果目标为文件）

`[-dus <path>]` //以字节为单位显示目标文件大小

`[-count [-q] <path>]` //将目录的大小、包含文件个数信息输出到屏幕 `[-mv <src> <dst>]` //把文件或目录移动到目标路径。 `[-rm [-skipTrash] <path>]` //删除文件，这个命令不能删除文件夹

`[-rmr [-skipTrash] <path>]` //删除文件夹及其下的所有文件

`[-expunge]`

`[-put <localsrc> ... <dst>]` //从本地文件系统上传文件到HDFS中



有genericOptions选项的命令

fs命令参数

`[-copyFromLocal <localsrc> ... <dst>]`//与put相同

`[-moveFromLocal <localsrc> ... <dst>]`//与put相同，但是文件上传之后会从本地文件系统中移除

`[-get [-ignoreCrc] [-crc] <src> <localdst>]`//复制文件到本地文件系统，这个命令可以选择是否忽略校验和，忽略校验和下载主要用于挽救那些已经发生错误的文件

`[-getmerge <src> <localdst> [addn1]]`//将源目录中的所有文件进行排序并写入目标文件中，文件之间以换行符分隔

`[-cat <src>]`//在终端显示（标准输出 stdout）文件中的内容，类似Linux系统中的cat

`[-text <src>]`

`[-copyToLocal [-ignoreCrc] [-crc] <src> <localdst>]`//与get相同

`[-moveToLoal [-crc] <src> <localdst>]`

`[-mkdir <path>]`//创建文件夹

`[-setrep [-R] [-w] <rep> <path/file>]`//改变一个文件的副本个数，参数-R可以递归地对该目录下的所有文件做统一操作



有genericOptions选项的命令

fs命令参数

`[-touchz <path>]` //类似Linux中的touch。创建一个空文件

`[-test [ezd] <path>]` //将源文件输出为文本格式显示到终端上，通过这个命令可以查看
TextRecordInputStream（SequenceFile等）或zip文件

`[-stat [format] <path>]` //以指定格式返回路径的信息

`[-stat [-f] <file>]` //在终端上显示（标注输出）文件的最后1kb内容。

`[-chmod [-R] <MODE[,MODE]... [OCTALMODE> PATH ...]` //改变文件的权限，只有文件的所有者
或是超级用户才能使用这个命令。

`[-chown [-R] [OWNER] [:[GROUNP]] PATH ...]` //改变文件的拥有者，-R可以递归地改变文件夹内
所有文件的拥有者。

`[-chgrp [-R] GROUP PATH ...]` //改变文件所属的组，-R可以递归地改变文件夹内所有文件所属的组。

`[-help [cmd]]` //这是命令的帮助信息



THANKS

谢 谢 观 看



联创中控（北京）科技有限公司

地址：北京市昌平区北京国际信息产业基地高新四街6号院5层

电话：400-659-9866

网址：<http://www.uicctech.com>