Hadoop学习第二天之hadoop命令操作

DFS：

1. (原因)数据越来越多，一个操作系统已经无法管辖
2. (方式)允许文件通过网络在多台主机间共享
3. (易用)通透性，内层机理对用户和程序透明，仿佛在本地磁盘操作。
4. (能力)容错，即使有某些机器脱机，整体仍然能正常运行
5. (适用)Hdfs适合一次写入多次查询，不支持并发写，不适合小文件

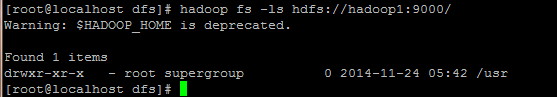
Hadoop命令操作：

格式：hadoop fs

可以在/etc/profile 中添加

alias hf=’hadoop fs’ #简化每次命令的书写

-ls <path> 表示对hdfs下一级目录的查看



-lsr <path> 表示对hdfs目录的递归查看

-mkdir <path> 创建目录

-put <src> <des> 从linux上传文件到hdfs

-get <src> <des> 从hdfs下载文件到linux

-text <path> 查看文件内容

-rm <path> 删除文件

-rmr <path> 表示递归删除文件, 删除文件(夹)

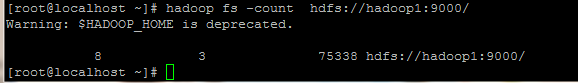
-touchz <path> //创建一个空文件

-help [cmd] //显示命令的帮助信息

-ls(r) <path> //显示当前目录下所有文件

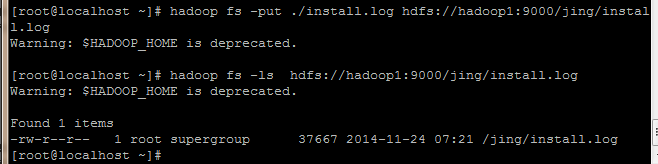
-du(s) <path> //显示目录中所有文件大小

-count[-q] <path> //显示目录中文件数量

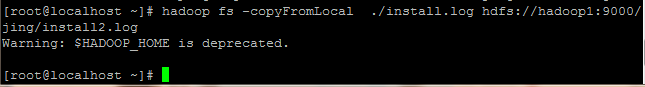


-mv <src> <dst> //移动多个文件到目标目录(都是hdfs目录)

-cp <src> <dst> //复制多个文件到目标目录(都是hdfs目录)



-copyFromLocal //同put



-moveFromLocal //从本地文件移动到hdfs

-getmerge <src> <localdst> //将源目录中的所有文件排序合并到一个文件中

-cat <src> //在终端显示文件内容

-text <src> //在终端显示文件内容

-copyToLocal [-ignoreCrc] <src> <localdst> //复制到本地

-moveToLocal <src> <localdst>

hdfs在对数据存储进行block划分时，如果文件大小超过block，那么按照block大小进行划分；不到block size的，划分为一个块，是实际数据大小。

Hadoop集群用户的常用命令

1. 创建一个hadoo档案文件

用法：hadoop archive – archiveName NAME -p <parent path> <src>\* <dest>

|  |  |
| --- | --- |
| **命令选项** | **描述** |
| -p | 后跟父路径 |
| -archiveName NAME | 要创建的档案的名字。 |
| src | 文件系统的路径名，和通常含正则表达的一样。 |
| dest | 保存档案文件的目标目录。 |

 Hadoop archive的扩展名是\*.har。Hadoop archive包含元数据（形式是\_index和\_masterindx）和数据（part-\*）文件。\_index文件包含了档案中的文件的文件名和位置信息。

由-archiveName选项指定你要创建的archive的名字。比如foo.har。archive的名字的扩展名应该是\*.har。输入是文件系统的路径名，路径名的格式和平时的表达方式一样。创建的archive会保存到目标目录下。注意创建archives是一个Map/Reduce job。你应该在map reduce集群上运行这个命令。下面是一个例子：

#hadoop archive -archiveName foo.har -p / . mytest2/

将hdfs /下的所有文件 复制到 /mytest2中

1. 递归的拷贝文件和目录。DistCp（分布式拷贝）

用法：hadoop distcp <srcurl> <desturl>

DistCp最常用在集群之间的拷贝：

bash$ hadoop distcp hdfs://nn1:8020/foo/bar \

hdfs://nn2:8020/bar/foo

这条命令会把nn1集群的/foo/bar目录下的所有文件或目录名展开并存储到一个临时文件中，这些文件内容的拷贝工作被分配给多个map任务， 然后每个TaskTracker分别执行从nn1到nn2的拷贝操作。注意DistCp使用绝对路径进行操作。

命令行中可以指定多个源目录：

bash$ hadoop distcp hdfs://nn1:8020/foo/a \

hdfs://nn1:8020/foo/b \

hdfs://nn2:8020/bar/foo

或者使用-f选项，从文件里获得多个源：

bash$ hadoop distcp -f hdfs://nn1:8020/srclist \

hdfs://nn2:8020/bar/foo

其中srclist 的内容是

hdfs://nn1:8020/foo/a

hdfs://nn1:8020/foo/b

Namenode ：

是整个文件系统的管理节点，维护者整个文件系统的文件目录树，文件、目录的元信息和每个文件对应的数据块列表。接收用户的操作请求。

文件包括：

fsimage:元数据镜像文件。存储某一时段NameNode内存元数据信息。

edits:操作日志文件。

fstime:保存最近一次checkpoint的时间

以上这些文件是保存在linux的文件系统中

DataNode:

提供真实的文件数据的存储服务；

文件块(block)：最基本的存储单位。

如果一个文件小于一个数据块的大小，并不占用真个数据块存储空间

SecondaryNameNode：

HA的一个解决方案。但不支持热备。配置即可。

执行过程：从NameNode上下载元数据信息（fsimage,edits），然后把二者合并，生成新的fsimage，在本地保存，并将其推送到NameNode，同时重置NameNode的edits.

默认在安装在NameNode节点上