Hadoop Archives

目录

1. 什么是Hadoop archives?

Hadoop archives是特殊的档案格式。一个Hadoop archive对应一个文件系统目录。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。

2. 如何创建archive?

用法: hadoop archive -archiveName name <src>* <dest>

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 /user/hadoop/dir1 /user/hadoop/dir2 /user/zoo/

在上面的例子中， /user/hadoop/dir1 和 /user/hadoop/dir2 会被归档到这个文件系统目录下 -- /user/zoo/foo.har。当创建archive时，源文件不会被更改或删除。

3. 如何查看archives中的文件?

archive作为文件系统层暴露给外界。所以所有的fs shell命令都能在archive上运行，但是要使用不同的URI。 另外，archive是不可改变的。所以重命名，删除和创建都会返回错误。Hadoop Archives 的URI是

har://scheme-hostname:port/archivepath/fileinarchive

如果没提供scheme-hostname，它会使用默认的文件系统。这种情况下URI是这种形式

har:///archivepath/fileinarchive

这是一个archive的例子。archive的输入是/dir。这个dir目录包含文件filea，fileb。 把/dir归档到/user/hadoop/foo.bar的命令是

hadoop archive -archiveName foo.har /dir /user/hadoop

获得创建的archive中的文件列表，使用命令

hadoop dfs -lsr har:///user/hadoop/foo.har

查看archive中的filea文件的命令-

hadoop dfs -cat har:///user/hadoop/foo.har/dir/filea