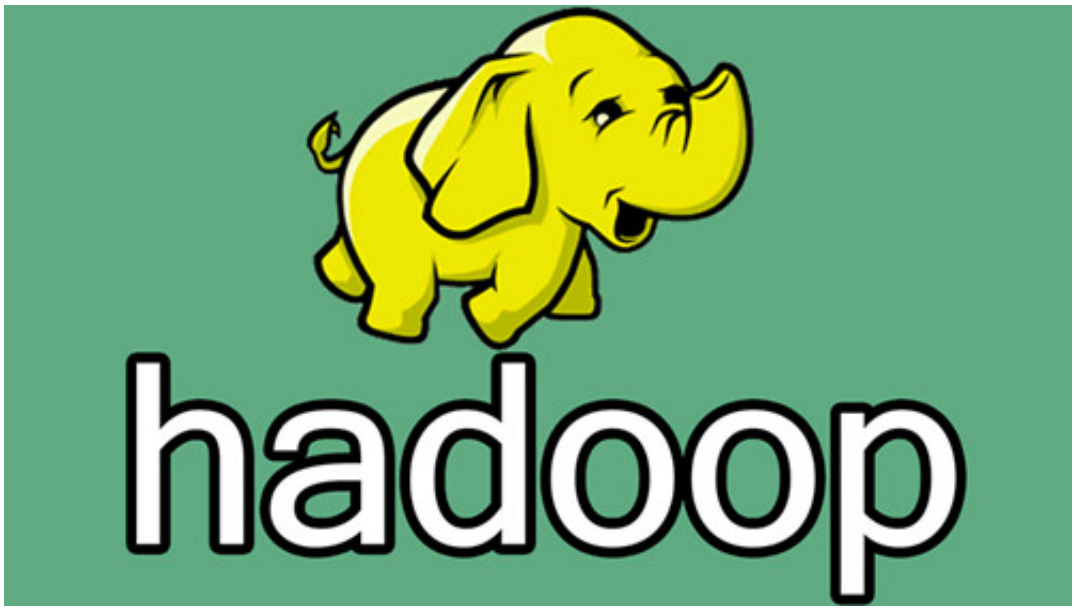
Hadoop Archives 编程指南

概述

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 文件包含了档案中文件的文件名和位置信息。



如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

如何创建 Archive

可以使用下面命令实现：

hadoop archive -archiveName name -p <parent> [-r <replication factor>] <src>* <dest>

-archiveName 选项指定你要创建的归档名字。比如 foo.har。归档文件的扩展名应该是 *.har。parent 参数用于指定文件应归档到的相对路径。 用法如下：

-p /foo/bar a/b/c e/f/g

其中 /foo/bar 是父路径，a/b/c 和 e/f/g
是对应于父路径的相对路径。注意创建归档文件的是一个 Map/Reduce 作业。你应该在 map reduce 集群上运行这个命令。
-r 表示希望复制因子；如果这个可选参数没有指定，那么副本因子将会是 3。

如果你仅仅归档 /foo/bar 单个目录，那么你可以仅仅使用下面命令：

hadoop archive -archiveName zoo.har -p /foo/bar -r 3 /outputdir

如果指定加密区域中的源文件，那么这些文件将会被解密，并且写入到归档文件中。如果 har 文件不在加密区（encryption zone）中，则它们将以被解密形式存储。 如果 har 文件位于加密区域，则将会以加密形式存储。

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

如果没提供模式，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

如何解压归档文件

由于归档中的所有 fs shell 命令工作都是透明的，因此解压归档文件只是复制的操作。

串行解压归档文件可以如下：

hdfs dfs -cp har:///user/zoo/foo.har/dir1 hdfs:/user/zoo/newdir

并行解压归档文件可以使用 DistCp:

hadoop distcp har:///user/zoo/foo.har/dir1 hdfs:/user/zoo/newdir

归档示例

创建归档文件

hadoop archive -archiveName foo.har -p /user/hadoop -r 3 dir1 dir2 /user/zoo

上面的例子使用 /user/hadoop 作为创建归档的相对归档目录。/user/hadoop/dir1 和 /user/hadoop/dir2 目录将会归档到 /user/zoo/foo.har 里面。归档操作并不会删除输入文件。如果你想在创建归档文件之后删除这些输入文件，你需要自己做。在这个例子中，因为我们指定了 -r 3，那么副本因子为3将会被使用。

查找文件

在 hadoop 档案中查找文件就像在文件系统上执行 ls 一样简单。在我们归档完 /user/hadoop/dir1 和 /user/hadoop/dir2
目录，如果我们想查看归档里面有哪些文件，你仅仅需要使用下面命令：

hdfs dfs -ls -R har:///user/zoo/foo.har/

要理解-p 参数的重要性，让我们再看一遍上面的例子。 如果您只是在 hadoop 存档上使用 ls（而不是lsr）

hdfs dfs -ls har:///user/zoo/foo.har

输出如下：

har:///user/zoo/foo.har/dir1
har:///user/zoo/foo.har/dir2

您可以回忆一下使用以下命令创建存档

hadoop archive -archiveName foo.har -p /user/hadoop dir1 dir2 /user/zoo

如果我们将上面命令修改为下：

hadoop archive -archiveName foo.har -p /user/ hadoop/dir1 hadoop/dir2 /user/zoo

那么在 Hadoop 归档上如下使用 ls 命令：

hdfs dfs -ls har:///user/zoo/foo.har

那么你会得到如下结果：

har:///user/zoo/foo.har/hadoop/dir1
har:///user/zoo/foo.har/hadoop/dir2

请注意，已归档文件已相对于 /user/ 而不是/ user/hadoop 进行归档。

Hadoop Archives 和 MapReduce

在 MapReduce 中使用 Hadoop Archives 就像使用默认文件系统中的文件一样简单。 如果我们在 HDFS 上的 /user/zoo/foo.har 路径里面存储了 Hadoop 归档文件，那么在 MapReduce 里面将它作为输入文件可以使用 har:///user/zoo/foo.har。

更多的关于 Hadoop 归档文件请参见：Hadoop Archives Guide