给Hadoop集群中添加Snappy解压缩库

Snappy是用C++开发的压缩和解压缩开发包，旨在提供高速压缩速度和合理的压缩率。Snappy比zlib更快，但文件相对要大20%到100%。在64位模式的Core i7处理器上，可达每秒250~500兆的压缩速度。

Snappy的前身是Zippy。虽然只是一个数据压缩库，它却被Google用于许多内部项目程，其中就包括BigTable，MapReduce和RPC。Google宣称它在这个库本身及其算法做了数据处理速度上的优化，作为代价，并没有考虑输出大小以及和其他类似工具的兼容性问题。Snappy特地为64位x86处理器做了优化，在单个Intel Core i7处理器内核上能够达到至少每秒250MB的压缩速率和每秒500MB的解压速率。

如果允许损失一些压缩率的话，那么可以达到更高的压缩速度，虽然生成的压缩文件可能会比其他库的要大上20%至100%，但是，相比其他的压缩库，Snappy却能够在特定的压缩率下拥有惊人的压缩速度，“压缩普通文本文件的速度是其他库的1.5-1.7倍，HTML能达到2-4倍，但是对于JPEG、PNG以及其他的已压缩的数据，压缩速度不会有明显改善”。

这篇文章主要是用来介绍如何给Hadoop集群中添加Snappy解压缩库。

一、安装snappy

```
yum install snappy snappy-devel
```

二、使得Snappy类库对Hadoop可用

```
ln -sf /usr/lib64/libsnappy.so /usr/lib/hadoop/lib/native/.
```

三、 在$HADOOP_HOME/etc/hadoop/core-site.xml文件中加入snappy配置

```
<property>
  <name>io.compression.codecs</name>
  <value>
    org.apache.hadoop.io.compress.GzipCodec,
    org.apache.hadoop.io.compress.DefaultCodec,
    org.apache.hadoop.io.compress.BZip2Codec,
    org.apache.hadoop.io.compress.SnappyCodec
  </value>
</property>
```

下面是配置在map的输出启用压缩

```
<property>
   <name>mapreduce.map.output.compress</name>
   <value>true</value>
</property>

<property>
   <name>mapreduce.map.output.compress.codec</name>
   <value>org.apache.hadoop.io.compress.SnappyCodec</value>
</property>
```

四、重新启动hadoop的相关进程，使得上面的配置生效

如果你要在Mapreduce程序里面使用Snappy相关类库，可以用下面的方法实现

```
...

Configuration conf = new Configuration();

//对map输出的内容进行压缩
conf.set("mapred.compress.map.output","true");
conf.set("mapred.map.output.compression.codec",
                "org.apache.hadoop.io.compress.SnappyCodec");

//对reduce输出的内容进行压缩
conf.set("mapred.output.compress","true");
conf.set("mapred.output.compression",
                "org.apache.hadoop.io.compress.SnappyCodec");

...
```