Hadoop-2.2.0使用lzo压缩文件作为输入文件

在《Hadoop 2.2.0安装和配置lzo》文章中介绍了如何基于 Hadoop 2.2.0安装lzo。里面简单介绍了如果在Hive里面使用lzo数据。今天主要来说说如何在Hadoop 2.2.0中使用lzo压缩文件当作的数据。

lzo压缩默认的是不支持切分的，也就是说，如果直接把lzo文件当作Mapreduce任务的输入，那么Mapreduce只会用一个Map来处理这个输入文件，这显然不是我们想要的。其实我们只需要对lzo文件建立索引，这样这个lzo文件就会支持切分，也就可以用多个Map来处理lzo文件。我们可以用《Hadoop 2.2.0安装和配置lzo》文章中编译的hadoop-lzo-0.4.20-SNAPSHOT.jar包来对lzo文件建立索引(假如在/home/wyp/input目录下有个cite.txt.lzo文件，这个目录是在HDFS上)：

```
$ $HADOOP_HOMOE/bin/hadoop jar
        $HADOOP_HOMOE/share/hadoop/common/hadoop-lzo-0.4.20-SNAPSHOT.jar
        com.hadoop.compression.lzo.DistributedLzoIndexer
        /home/wyp/input/cite.txt.lzo
```

生成出来的索引文件后缀为.index，并存放在lzo同一目录下.在本例中产生的索引文件是存放在/home/wyp/input目录下，名称为cite.txt.lzo.index。

我们也可以用下面的方法对lzo文件来建立索引：

```
$ $HADOOP_HOMOE/bin/hadoop jar
        $HADOOP_HOMOE/share/hadoop/common/hadoop-lzo-0.4.20-SNAPSHOT.jar
        com.hadoop.compression.lzo.LzoIndexer
        /home/wyp/input/cite.txt.lzo
```

这个方法和上面方法产生出来的索引文件是一样的；但是上面的方法是通过启用Mapreduce任务来执行的，而这里的方法只在一台客户机上运行，效率很慢！

那么，如何在Mapreduce任务中使用lzo文件。下面分别对Mapreduce程序、Streaming程序以及Hive分别进行说明：

1、对于Mapreduce程序，我们需要把程序中所有的TextInputFormat修改为LzoTextInputFormat，如下：

```
job.setInputFormatClass(TextInputFormat.class);
```

修改为

```
job.setInputFormatClass(LzoTextInputFormat.class);
```

LzoTextInputFormat类需要引入相应的包，如果你是使用pom文件，可以引入以下依赖：

```
<dependency>
        <groupId>com.hadoop.gplcompression</groupId>
        <artifactId>hadoop-lzo</artifactId>
        <version>0.4.19</version>
</dependency>
```

如果你的输入格式不是LzoTextInputFormat类，那么Mapreduce程序将会把.index文件也当作是数据文件！修改完之后，需要重新编译你的Mapreduc程序。这样在运行Mapreduce程序的时候，将lzo文件所在的目录当作输入即可，Mapreduce程序会识别出.index文件的：

```
$ /home/q/hadoop-2.2.0/bin/hadoop jar
                        statistics2.jar com.wyp.Sts
                        -Dmapreduce.job.queuename=queue1
                        /home/wyp/input
                        /home/wyp/resluts
```

2、对于Streaming程序来说，可以通过-inputformat指定输入的文件格式，使用如下：

```
$ bin/hadoop jar
        $HADOOP_HOMOE/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar
        -inputformat com.hadoop.mapred.DeprecatedLzoTextInputFormat
        -input /home/wyp/input
        -output /home/wyp/results
        -mapper /bin/cat
        -reducer wc
```

对应Streaming作业还需要注意的是，使用DeprecatedLzoTextInputFormat输入格式，会把文本的行号当作key传送到reduce的，所以我们需要将行号去掉，可以用下面方法实现：

```
$ bin/hadoop jar
        $HADOOP_HOMOE/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar
        -inputformat com.hadoop.mapred.DeprecatedLzoTextInputFormat
        -input /home/wyp/input
        -D stream.map.input.ignoreKey=true
        -output /home/wyp/results
        -mapper /bin/cat
        -reducer wc
```

3、对于Hive，需要在建表的时候注意，如下：

```
hive> create table lzo(
    > id int,
    > name string)
    > STORED AS INPUTFORMAT 'com.hadoop.mapred.DeprecatedLzoTextInputFormat'
    > OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat';
OK
Time taken: 3.423 seconds
```

注意4,5行代码。这样就可以使用lzo文件了，并支持分割。