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)：

```
1  $ $HADOOP_HOMOE/bin/hadoop jar
2 $HADOOP_HOMOE/share/hadoop/common/hadoop-lzo-0.4.20-SNAPSHOT.jar
3 com.hadoop.compression.lzo.DistributedLzoIndexer
4 /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：

```
1 $ $HADOOP_HOMOE/bin/hadoop jar
2 $HADOOP_HOMOE/share/hadoop/common/hadoop-lzo-0.4.20-SNAPSHOT.jar
3 com.hadoop.compression.lzo.LzoIndexer
4 /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：

```
1 job.setInputFormatClass(TextInputFormat.class);
2
3 修改为
```

```
4
5 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：

```
1 <dependency>
2         <groupId>com.hadoop.gplcompression</groupId>
3         <artifactId>hadoop-lzo</artifactId>
4         <version>0.4.19</version>
5 </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：

```
1 $ /home/q/hadoop-2.2.0/bin/hadoop jar
2                   statistics2.jar com.wyp.Sts
3                   -Dmapreduce.job.queuename=queue1
4                   /home/wyp/input
5                   /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：


```
1 $ bin/hadoop jar
2         $HADOOP_HOMOE/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar
3         -inputformat com.hadoop.mapred.DeprecatedLzoTextInputFormat
4         -input /home/wyp/input
5         -output /home/wyp/results
6         -mapper /bin/cat
7         -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：

```
1 $ bin/hadoop jar
2         $HADOOP_HOMOE/share/hadoop/tools/lib/hadoop-streaming-2.2.0.jar
3         -inputformat com.hadoop.mapred.DeprecatedLzoTextInputFormat
4         -input /home/wyp/input
5         -D stream.map.input.ignoreKey=true
6         -output /home/wyp/results
7         -mapper /bin/cat
8         -reducer wc
```

3、对于 Hive，需要在建表的时候注意，如下：

```
1 hive> create table lzo(
2     > id int,
3     > name string)
4     > STORED AS INPUTFORMAT 'com.hadoop.mapred.DeprecatedLzoTextInputFormat'
5     > OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat';
6 OK
7 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。

云凡教育大数据学院

关注云凡教育微信公众号 yfteach，第一时间获取公开课信息。