Hadoop 2.2.0安装和配置lzo

Hadoop经常用于处理大量的数据，如果期间的输出数据、中间数据能压缩存储，对系统的I/O性能会有提升。综合考虑压缩、解压速度、是否支持split，目前lzo是最好的选择。LZO（LZO是Lempel-Ziv-Oberhumer的缩写）是一种高压缩比和解压速度极快的编码，它的特点是解压缩速度非常快，无损压缩，压缩后的数据能准确还原，lzo是基于block分块的，允许数据被分解成chunk，能够被并行的解压。LZO库实现了许多有下述特点的算法：

（1）、解压简单，速度非常快。
（2）、解压不需要内存。
（3）、压缩相当地快。
（4）、压缩需要64 kB的内存。
（5）、允许在压缩部分以损失压缩速度为代价提高压缩率，解压速度不会降低。
（6）、包括生成预先压缩数据的压缩级别，这样可以得到相当有竞争力的压缩比。
（7）、另外还有一个只需要8 kB内存的压缩级别。
（8）、算法是线程安全的。
（9）、算法是无损的。

本文针对Hadoop 2.2.0，介绍如何安装和使用lzo。

一、下载、解压并编译lzo包

```
[wyp@master ~]$ wget http://www.oberhumer.com/opensource/lzo/download/lzo-2.06.tar.gz
[wyp@master ~]$ tar -zxvf lzo-2.06.tar.gz
[wyp@master ~]$ cd lzo-2.06
[wyp@master ~]$ export CFLAGS=-m64
[wyp@master ~]$ ./configure -enable-shared -prefix=/usr/local/hadoop/lzo/
[wyp@master ~]$ make && sudo make install
```

编译完lzo包之后，会在/usr/local/hadoop/lzo/生成一些文件，目录结构如下：

```
[wyp@master /usr/local/hadoop/lzo]$ ls -l
total 12
drwxr-xr-x 3 root root 4096 Mar 21 17:23 include
drwxr-xr-x 2 root root 4096 Mar 21 17:23 lib
drwxr-xr-x 3 root root 4096 Mar 21 17:23 share
```

将/usr/local/hadoop/lzo目录下的所有文件打包，并同步到集群中的所有机器上。

在编译lzo包的时候，需要一些环境，可以用下面的命令安装好lzo编译环境

```
[wyp@master ~]$ yum -y install  lzo-devel    ₩
            zlib-devel  gcc autoconf automake libtool
```

二、安装Hadoop-LZO

这里下载的是Twitter
hadoop-
lzo，可以用Maven(如何安装Maven请参照本
博客的《Linux命令行下安装Maven与配置》)进行编译。

```
[wyp@master ~]$ wget https://github.com/twitter/hadoop-lzo/archive/master.zip
```

下载后的文件名是master，它是一个zip格式的压缩包，可以进行解压：

```
[wyp@master ~]$ unzip master
```

解压后的文件夹名为hadoop-lzo-master

当然，如果你电脑安装了git，你也可以用下面的命令去下载

```
[wyp@master ~]$ git clone https://github.com/twitter/hadoop-lzo.git
```

hadoop-lzo中的pom.xml依赖了hadoop2.1.0-beta，由于我们这里用到的是Hadoop
2.2.0，所以建议将hadoop版本修改为2.2.0：

```
<properties>
   <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
   <hadoop.current.version>2.2.0</hadoop.current.version>
   <hadoop.old.version>1.0.4</hadoop.old.version>
 </properties>
```

然后进入hadoop-lzo-master目录，依次执行下面的命令

```
[wyp@master hadoop-lzo-master]$ export CFLAGS=-m64
[wyp@master hadoop-lzo-master]$ export CXXFLAGS=-m64
[wyp@master hadoop-lzo-master]$ export C_INCLUDE_PATH=     ₩
                        /usr/local/hadoop/lzo/include
[wyp@master hadoop-lzo-master]$ export LIBRARY_PATH=/usr/local/hadoop/lzo/lib
[wyp@master hadoop-lzo-master]$ mvn clean package -Dmaven.test.skip=true
[wyp@master hadoop-lzo-master]$ cd target/native/Linux-amd64-64
[wyp@master Linux-amd64-64]$ tar -cBf - -C lib . | tar -xBvf - -C ~
[wyp@master ~]$cp ~/libgplcompression* $HADOOP_HOME/lib/native/
[wyp@master hadoop-lzo-master]$cp target/hadoop-lzo-0.4.18-SNAPSHOT.jar   ₩
                        $HADOOP_HOME/share/hadoop/common/
```

其实在tar -cBf - -C lib . | tar -xBvf - -C ~命令之后，会在~目录下生成一下几个文件：

```
[wyp@master ~]$ ls -l
-rw-r--r--  1 libgplcompression.a
-rw-r--r--  1 libgplcompression.la
lrwxrwxrwx  1 libgplcompression.so -> libgplcompression.so.0.0.0
lrwxrwxrwx  1 libgplcompression.so.0 -> libgplcompression.so.0.0.0
-rwxr-xr-x  1 libgplcompression.so.0.0.0
```

其中libgplcompression.so和libgplcompression.so.0是链接文件，指向libgplcompression.so.0.0.0，将刚刚生成的libgplcompression*和target/hadoop-lzo-0.4.18-SNAPSHOT.jar同步到集群中的所有机器对应的目录。

三、配置Hadoop环境变量

1、在Hadoop中的$HADOOP_HOME/etc/hadoop/hadoop-env.sh加上下面配置：

```
export LD_LIBRARY_PATH=/usr/local/hadoop/lzo/lib
```

2、在$HADOOP_HOME/etc/hadoop/core-site.xml加上如下配置：

```
<property>
 <name>io.compression.codecs</name>
 <value>org.apache.hadoop.io.compress.GzipCodec,
       org.apache.hadoop.io.compress.DefaultCodec,
       com.hadoop.compression.lzo.LzoCodec,
       com.hadoop.compression.lzo.LzopCodec,
       org.apache.hadoop.io.compress.BZip2Codec
       </value>
</property>

<property>
 <name>io.compression.codec.lzo.class</name>
 <value>com.hadoop.compression.lzo.LzoCodec</value>
</property>
```

3、在$HADOOP_HOME/etc/hadoop/mapred-site.xml加上如下配置

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

<property>
 <name>mapred.map.output.compression.codec</name>
 <value>com.hadoop.compression.lzo.LzoCodec</value>
</property>

<property>
 <name>mapred.child.env</name>
 <value>LD_LIBRARY_PATH=/usr/local/hadoop/lzo/lib</value>
</property>
```

将刚刚修改的配置文件全部同步到集群的所有机器上，并重启Hadoop集群，这样就可以在Hadoop中使用lzo。

四、如何使用

这里在Hive中使用一下lzo，在hive中创建一个lzo表：

```
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
```

如果在创建lzo表出现了如下错误：

```
FAILED: Error in metadata: Class not found:    ₩
com.hadoop.mapred.DeprecatedLzoTextInputFormat
FAILED: Execution Error,
return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask
```

请检查你的环境是否配置好。

然后在本地用lzo压缩一个文件，先看看users.txt的内容：

```
[wyp@master ~]$ cat users.txt
1^Awyp
2^Azs
3^Als
4^Aww
5^Awyp2
6^Awyp3
7^Awyp4
8^Awyp5
9^Awyp6
10^Awyp7
11^Awyp8
12^Awyp5
13^Awyp9
14^Awyp20
[wyp@master ~]$ lzop users.txt
[wyp@master ~]$ ls -l users.txt*
-rw-r--r-- 1 wyp wyp 97  Mar 25 15:40 users.txt
-rw-r--r-- 1 wyp wyp 154 Mar 25 15:40 users.txt.lzo
```

将users.txt.lzo的数据导入到lzo表里面：

```
hive> load data local inpath '/home/wyp/users.txt.lzo' into table lzo;
Copying data from file:/home/wyp/users.txt.lzo
Copying file: file:/home/wyp/users.txt.lzo
Loading data to table default.lzo
Table default.lzo stats: [num_partitions: 0, num_files: 1,
            num_rows: 0, total_size: 154, raw_data_size: 0]
OK
Time taken: 0.49 seconds
hive> select * from lzo;
OK
1 wyp
2 zs
3 ls
4 ww
5 wyp2
6 wyp3
7 wyp4
8 wyp5
9 wyp6
10 wyp7
11 wyp8
12 wyp5
13 wyp9
14 wyp20
Time taken: 0.244 seconds, Fetched: 14 row(s)
```

好了，我们可以在Hadoop中使用lzo了！！（完）