hive的压缩

1 lzo压缩

1.1 lzo的简介

```
LZO是一个无损的数据压缩库，相比于压缩比它更加追求速度。 查阅
ttp://www.oberhumer.com/opensource/lzo 和http://www.lzop.org 或缺更多有关 LZO的信息
and 查阅压缩数据存储格式 获取有关Hive压缩数据存储信息。

hadoop下各种压缩算法的压缩比，压缩时间，解压时间见下表:
压缩算法     原始文件大小   压缩后的文件大小     压缩速度     解压缩速度
gzip         8.3GB           1.8GB                  17.5MB/s    58MB/s
bzip2        8.3GB           1.1GB                   2.4MB/s    9.5MB/s
LZO-bset     8.3GB           2GB                     4MB/s      60.6MB/s
LZO          8.3GB           2.9GB                   49.3MB/S   74.6MB/s

lzo的压缩率不高，但是压缩、解压速度都比较高。
```

- 启用lzo

```
启用lzo的压缩方式对于小规模集群是很有用处，压缩比率大概能降到原始日志大小的1/3。同时解压缩的速
度也比较快。
```

lzo的官方介绍：

- 安装lzo

```
lzo并不是linux系统原生支持，所以需要下载安装软件包。这里至少需要安装3个软件包：lzo, lzop,
hadoop-gpl-packaging。
```

Version 2.10

01 Mar 2017

Copyright (C) 1996 - 2017
Markus F.X.J. Oberhumer

[News] [Abstract] [Download] [miniLZO] [Links]

News

- LZO 2.10 has been released; a small update that fixes various build issues.

Key Facts

- LZO is a portable lossless data compression library written in ANSI C.
- Offers pretty fast compression and **extremely** fast decompression.
- One of the fastest compression and decompression algorithms around. See the ratings for lzop in the famous Archive Comparison Test .
- Includes slower compression levels achieving a quite competitive compression ratio while still decompressing at this very high speed.
- Distributed under the terms of the GNU General Public License (GPL v2+). Commercial licenses are available through our LZO Professional license program.

Download

LZO is distributed as portable ANSI C source code.

Download LZO (source code, 587 kB, SHA1: 4924676a9bae5db58ef129dc1cebce3baa3c4b5d).

miniLZO

miniLZO is a very lightweight **subset** of the LZO library intended for easy inclusion with your application. It is **generated automatically** from the LZO source code and contains the most important LZO functions.

Very easy to use - it only takes a few **minutes** to add data compression to your application!

Download miniLZO (source code, 62 kB, SHA1: c7432708d49017a3f0b4f44c99d336f8a1be84f5).

Related links

- LZO Professional is our commercial LZO license program.
- If you need better compression you should take a look at the excellent zlib library. zlib is slower and needs more memory, though.
- For even better compression consider using *libbzip2* which is distributed with the bzip2 file compressor.
- The file compressor application lzop uses LZO - it is very similar to gzip but much faster.

hive官网案例：

假设一个有三列的简单数据文件。

- id
- first name
- last name

向这个数据文件中插入4条记录:

```
19630001     john     lennon
19630002     paul     mccartney
19630003     george   harrison
19630004     ringo    starr
```

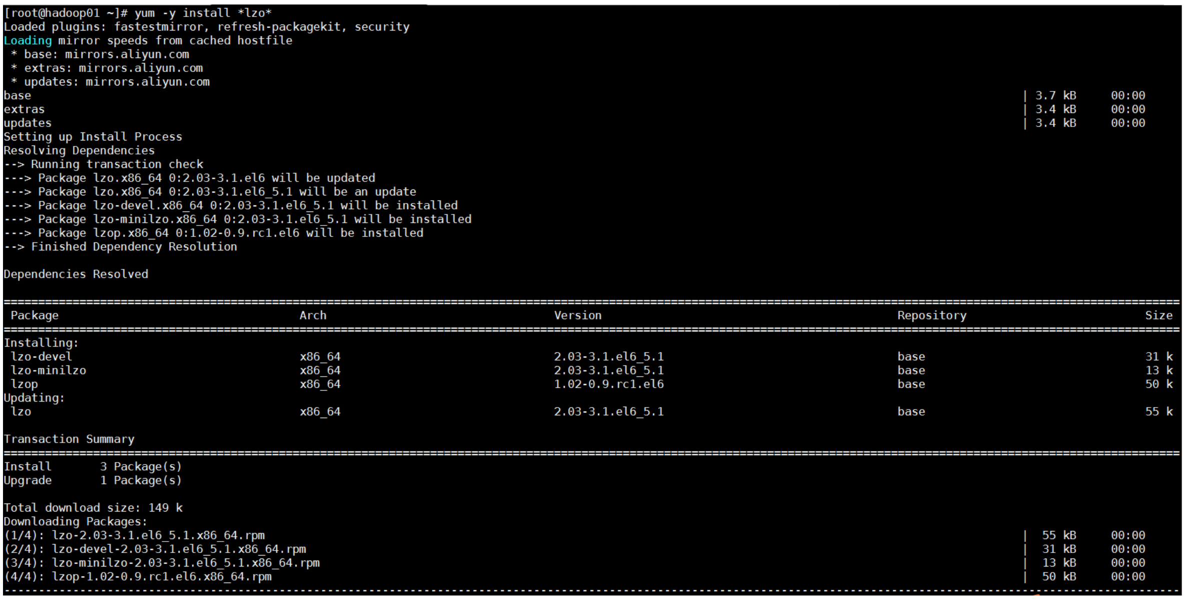
调用这个数据文件 `/home/hivedata/lzodata.txt`.

为了使它成为LZO文件，我们可以使用lzop应用程序，它将创建一个名字类似 `lzodata.txt.lzo` 的文件。把这个文件拷贝到HDFS中。

1.2 lzo的安装测试

要在Hadoop集群中每个节点里安装 `lzo` 和 `lzop` 。安装的细节不在本文档中进行叙述。但是我这里讲解下安装过程。安装lzo和lzop步骤如下：

```
1、在hadoop集群每个节点上安装lzo和lzop及其依赖(主要为解决安装lzop):
[root@hadoop01 ~]# yum -y install *lzo*
```



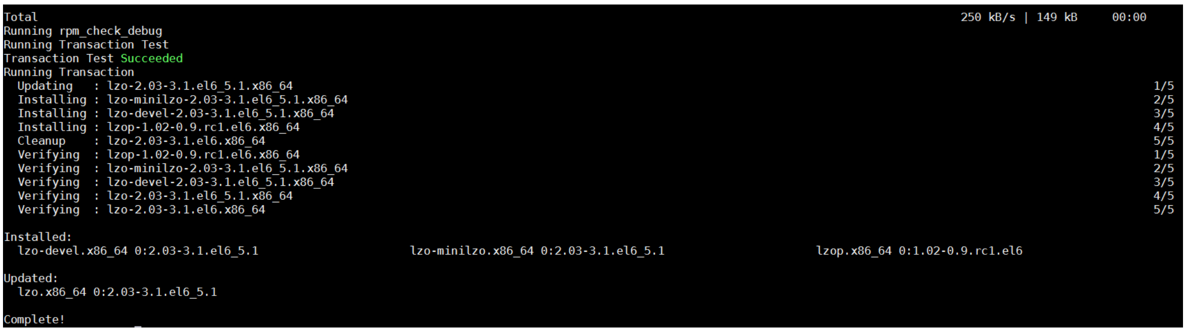
```
[root@hadoop01 ~]# yum -y install *lzo*
Loaded plugins: fastestmirror, refresh-packagekit, security
Loading mirror speeds from cached hostfile
 * base: mirrors.aliyun.com
 * extras: mirrors.aliyun.com
 * updates: mirrors.aliyun.com
base                                                                                                                                        | 3.7 kB     00:00
extras                                                                                                                                      | 3.4 kB     00:00
updates                                                                                                                                     | 3.4 kB     00:00
Setting up Install Process
Resolving Dependencies
--> Running transaction check
---> Package lzo.x86_64 0:2.03-3.1.el6 will be updated
---> Package lzo.x86_64 0:2.03-3.1.el6_5.1 will be an update
---> Package lzo-devel.x86_64 0:2.03-3.1.el6_5.1 will be installed
---> Package lzo-minilzo.x86_64 0:2.03-3.1.el6_5.1 will be installed
---> Package lzop.x86_64 0:1.02-0.9.rc1.el6 will be installed
--> Finished Dependency Resolution

Dependencies Resolved

=========================================================================================================================================================================
 Package                                     Arch                                  Version                                          Repository                         Size
=========================================================================================================================================================================
Installing:
 lzo-devel                                   x86_64                                2.03-3.1.el6_5.1                                 base                               31 k
 lzo-minilzo                                 x86_64                                2.03-3.1.el6_5.1                                 base                               13 k
 lzop                                        x86_64                                1.02-0.9.rc1.el6                                 base                               50 k
Updating:
 lzo                                         x86_64                                2.03-3.1.el6_5.1                                 base                               55 k

Transaction Summary
=========================================================================================================================================================================
Install       3 Package(s)
Upgrade       1 Package(s)

Total download size: 149 k
Downloading Packages:
(1/4): lzo-2.03-3.1.el6_5.1.x86_64.rpm                                                                                                      |  55 kB     00:00
(2/4): lzo-devel-2.03-3.1.el6_5.1.x86_64.rpm                                                                                                |  31 kB     00:00
(3/4): lzo-minilzo-2.03-3.1.el6_5.1.x86_64.rpm                                                                                              |  13 kB     00:00
(4/4): lzop-1.02-0.9.rc1.el6.x86_64.rpm                                                                                                     |  50 kB     00:00
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```



```
Total                                                                                                                              250 kB/s | 149 kB     00:00
Running rpm_check_debug
Running Transaction Test
Transaction Test Succeeded
Running Transaction
  Updating   : lzo-2.03-3.1.el6_5.1.x86_64                                                                                                                       1/5
  Installing : lzo-minilzo-2.03-3.1.el6_5.1.x86_64                                                                                                               2/5
  Installing : lzo-devel-2.03-3.1.el6_5.1.x86_64                                                                                                                 3/5
  Installing : lzop-1.02-0.9.rc1.el6.x86_64                                                                                                                      4/5
  Cleanup    : lzo-2.03-3.1.el6.x86_64                                                                                                                           5/5
  Verifying  : lzop-1.02-0.9.rc1.el6.x86_64                                                                                                                      1/5
  Verifying  : lzo-minilzo-2.03-3.1.el6_5.1.x86_64                                                                                                               2/5
  Verifying  : lzo-devel-2.03-3.1.el6_5.1.x86_64                                                                                                                 3/5
  Verifying  : lzo-2.03-3.1.el6_5.1.x86_64                                                                                                                       4/5
  Verifying  : lzo-2.03-3.1.el6.x86_64                                                                                                                           5/5

Installed:
  lzo-devel.x86_64 0:2.03-3.1.el6_5.1                          lzo-minilzo.x86_64 0:2.03-3.1.el6_5.1                          lzop.x86_64 0:1.02-0.9.rc1.el6

Updated:
  lzo.x86_64 0:2.03-3.1.el6_5.1

Complete!
```

源码编译安装lzo:

```
安装准备:
[root@hadoop01 home]# yum -y install gcc-c++ lzo-devel zlib-devel autoconf
automake libtool

编译安装:
下载路径:http://www.oberhumer.com/opensource/lzo/download/lzo-2.10.tar.gz

解压下载的源码:
[root@hadoop01 home]# tar -zxvf /home/lzo-2.10.tar.gz


[root@hadoop01 home]# cd /home/lzo-2.10/
[root@hadoop01 lzo-2.10]# ./configure -prefix=/usr/local/lzo/
[root@hadoop01 lzo-2.10]# make
[root@hadoop01 lzo-2.10]# make install
```

编译hadoop-lzo源码:

```
1、下载源码
https://github.com/twitter/hadoop-lzo/archive/master.zip
2、上传到服务器，并解压，修改pom.xml

[root@hadoop01 home]# unzip /home/hadoop-lzo-master.zip
[root@hadoop01 home]# cd /home/hadoop-lzo-master

搜索内容hadoop.current并修改版本号:
```

```
<hadoop.current.version>2.7.1</hadoop.current.version>

3、使用maven编译(默认maven已经安装)
 export C_INCLUDE_PATH=/usr/local/lzo/include
 export LIBRARY_PATH=/usr/local/lzo/lib

 4、编译
 [root@hadoop01 hadoop-lzo-master]# mvn package -Dmaven.test.skip=true

5、进入target，将hadoop-lzo-0.4.21-SNAPSHOT.jar放到hadoop的classpath下。如
${HADOOP_HOME}/share/hadoop/common

[root@hadoop01 hadoop-lzo-master]# cp ./target/hadoop-lzo-0.4.21-SNAPSHOT.jar
/usr/local/hadoop-2.7.1/share/hadoop/common/

分发到其它服务器：
[root@hadoop01 hadoop-lzo-master]# scp ./target/hadoop-lzo-0.4.21-SNAPSHOT.jar
hadoop02:/usr/local/hadoop-2.7.1/share/hadoop/common/
hadoop-lzo-0.4.21-SNAPSHOT.jar

[root@hadoop01 hadoop-lzo-master]# scp ./target/hadoop-lzo-0.4.21-SNAPSHOT.jar
hadoop03:/usr/local/hadoop-2.7.1/share/hadoop/common/
hadoop-lzo-0.4.21-SNAPSHOT.jar
```

在core-stie.xml中配置如下，并且同步到每台服务器：

```
<property>
<name>io.compression.codecs</name>
<value>org.apache.hadoop.io.compress.GzipCodec,org.apache.hadoop.io.compress.Def
aultCodec,org.apache.hadoop.io.compress.BZip2Codec,com.hadoop.compression.lzo.Lz
oCodec,com.hadoop.compression.lzo.LzopCodec</value>
</property>
<property>
<name>io.compression.codec.lzo.class</name>
<value>com.hadoop.compression.lzo.LzoCodec</value>
</property>

分发到每台服务器：
[root@hadoop01 hadoop-2.7.1]# scp -r ./etc/hadoop/core-site.xml
hadoop02:/usr/local/hadoop-2.7.1/etc/hadoop/
[root@hadoop01 hadoop-2.7.1]# scp -r ./etc/hadoop/core-site.xml
hadoop03:/usr/local/hadoop-2.7.1/etc/hadoop/

重启集群：
start-all.sh
```

创建lzo的表：

```
CREATE TABLE lzo_test(
id bigint,
firstname string,
lastname string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS  INPUTFORMAT "com.hadoop.mapred.DeprecatedLzoTextInputFormat"
OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat"
;
```

加载数据：

```
将/home/hivedata/lzodata.txt使用lzop生成.lzo文件：
[root@hadoop01 hivedata]# lzop ./lzodata.txt

LOAD DATA Local INPATH '/home/hivedata/lzodata.txt.lzo' INTO TABLE lzo_test;

hive> select * from lzo_test;
OK
19630001        john    lennon
19630002        paul    mccartney
19630003        george  harrison
19630004        ringo   starr
Time taken: 0.097 seconds, Fetched: 4 row(s)
```

索引lzo文件：

```
1. 批量lzo文件修改
[root@hadoop01 hivedata]# hadoop jar /usr/local/hadoop-
2.7.1/share/hadoop/common/hadoop-lzo-0.4.21-SNAPSHOT.jar
com.hadoop.compression.lzo.DistributedLzoIndexer /user/hive/warehouse/lzo_test/

2. 单个lzo文件修改
[root@hadoop01 hivedata]# hadoop jar /usr/local/hadoop-
2.7.1/share/hadoop/common/hadoop-lzo-0.4.21-SNAPSHOT.jar
com.hadoop.compression.lzo.DistributedLzoIndexer
/user/hive/warehouse/lzo_test/lzodata.txt.lzo

注意：
1、使用mr执行，并且会生成索引文件。

2、lzo本身是不支持split的。故如果需要使用lzo，一般有2种办法：
1）合理控制生成的lzo大小，建议不要超过一个block大小。因为如果没有lzo的index文件，该lzo会由一个map处理。如果lzo过大，会导致某个map处理时间过长。
2）配合lzo.index文件使用。好处是文件大小不受限制，可以将文件设置的稍微大点，这样有利于减少文件数目。坏处是生成lzo.index文件本身需要开销。
```

查询：

```
select id,firstname from lzo_test limit 3;
```

修改使用中hive表的输入输出格式:

```
ALTER TABLE lzo_test SET FILEFORMAT
INPUTFORMAT 'com.hadoop.mapred.DeprecatedLzoTextInputFormat'
OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat"
SERDE "org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe";
```

2 snaapy压缩

自带的，直接用即可。

参考：

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+LZO

https://www.cnblogs.com/allthewayforward/p/11131218.html

https://blog.csdn.net/joseph_happy/article/details/50374057