HDP 配置 lzo

笔记本： 大数据
创建时间： 2015/8/20 14:50 **更新时间：** 2015/8/21 11:21

HDP 配置 lzo

1. 在各个节点安装 lzo (namenode & datanode)
   yum install lzo lzo-devel hadoop-lzo hadoop-lzo-native

2. 配置 lzo
   1) core-site.xml
      io.compression.codecs:

org.apache.hadoop.io.compress.GzipCodec,org.apache.hadoop.io.compress.DefaultCodec,org.apache.hadoop.io.compress.SnappyCodec,com.hadoop.compression.lzo.LzoCodec,com.hadoop.compression.lzo.LzopCodec

      io.compression.codec.lzo.class:
      com.hadoop.compression.lzo.LzoCodec

   2) mapred-site.xml
      mapreduce.map.output.compress: true

      mapreduce.map.output.compress.codec: com.hadoop.compression.lzo.LzoCodec

      *** mapred.child.env: 每个子进程传递的环境变量 -- old, 已经遗弃的属性名称, 用 mapreduce.map.env & mapreduce.reduce.env 替代
      LD_LIBRARY_PATH=/usr/hdp/current/share/lzo/0.6.0/lib

      mapreduce.map.env: LD_LIBRARY_PATH=/usr/hdp/current/share/lzo/0.6.0/lib
      mapreduce.reduce.env: LD_LIBRARY_PATH=/usr/hdp/current/share/lzo/0.6.0/lib

3. hadoop 集群内测试 lzo
   1) 安装 lzop
      $ wget http://www.lzop.org/download/lzop-1.03.tar.gz
      $ tar zxvf lzop-1.03.tar.gz && cd lzop-1.03
      $ export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/lib

      $ ./configure
      $ make && make install

   2) 利用 lzop 压缩文件
      $ lzop -U -9 ./2015-05-01.dat
      $ hadoop fs -put ./2015-05-01.dat.lzo /rawdata/position/860100010020300001/2015-05-01

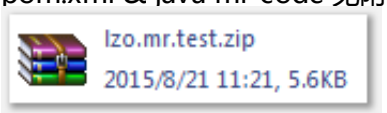
$ hadoop jar /usr/hdp/current/share/lzo/0.6.0/lib/hadoop-lzo-0.6.0.jar com.hadoop.compression.lzo.DistributedLzoIndexer /rawdata/position/860100010020300001/2015-05-01/2015-05-01.dat.lzo

3) 创建 Hive table, 测试读取压缩数据

```
## create table in hive
USE mining;
CREATE EXTERNAL TABLE IF NOT EXISTS rawdata_position (
        bubble_date STRING,
        mac STRING,
        build_id STRING,
        floor_id STRING,
        axis_x STRING,
        axis_y STRING)
PARTITIONED BY (p_build_id STRING, p_date STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
-- STORED AS INPUTFORMAT "com.hadoop.mapreduce.LzoTextInputFormat"
STORED AS INPUTFORMAT "com.hadoop.mapred.DeprecatedLzoTextInputFormat"
OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat"
LOCATION '/rawdata/position';

## add partition for testing data
$ hive -e "use mining;alter table rawdata_position add if not exists partition(p_build_id='860100010020300001',p_date='2015-05-01') location '/rawdata/position/860100010020300001/2015-05-01';"

## test hive table with lzo data
$ hive -e "use mining; select * from rawdata_position where p_build_id = '860100010020300001' and p_date = '2015-05-01' limit 100;"
```

4) 测试 Mapreduce & streaming

```
## 本地 hadoop-lzo jar 导入到 maven 中
$ mvn install:install-file -Dfile=/usr/hdp/current/share/lzo/0.6.0/lib/hadoop-lzo-0.6.0.jar \
                        -DgroupId=com.hadoop.gplcompression \
                        -DartifactId=hadoop-lzo \
                        -Dversion=0.6.0 \
                        -Dpackaging=jar

# pom.xml & java mr code 见附件
```




```
## Hadoop streaming 处理 lzo 文件。注: -input 指定还有 lzo 文件的目录, .index 文件会
```

被忽略

```
$ hadoop jar /usr/hdp/2.2.6.0-2800/hadoop-mapreduce/hadoop-streaming.jar \
        -inputformat com.hadoop.mapred.DeprecatedLzoTextInputFormat \
        -input /user/root/input/ncdc \
        -output /user/root/output \
        -mapper /bin/cat \
        -reducer wc

# 使用 DeprecatedLzoTextInputFormat 输入格式,
# 会把行号当作 key 传送到 reduce, 去掉行号, 可以用下面方法
$ hadoop jar /usr/hdp/2.2.6.0-2800/hadoop-mapreduce/hadoop-streaming.jar \
        -Dstream.map.input.ignoreKey=true \
        -Dmapreduce.job.reduces=0 \
        -inputformat com.hadoop.mapred.DeprecatedLzoTextInputFormat \
        -input /user/root/input/ncdc \
        -output /user/root/output \
        -mapper /bin/cat
```

5) 安装 impala-lzo, 让 impala 支持读取 lzo 表

1) 搭建本地 CDH gplextras repo

ref: http://archive.cloudera.com/gplextras5/redhat/6/x86_64/gplextras/
http://archive.cloudera.com/cdh5/repo-as-tarball/
http://abcve.com/use-cdh5-install-hadoop-cluster/

2) 安装 impala-lzo 库

```
$ yum install impala-lzo
```

4. reference docs

Hadoop 2.2.0安装和配置lzo
http://www.iteblog.com/archives/992

Hadoop-2.2.0使用lzo压缩文件作为输入文件
http://www.iteblog.com/archives/996

hive对lzo文件并行处理的关键点
http://www.cnblogs.com/cloudma/archive/2012/11/15/hadoop-lzo-index.html

others
http://www.dedecms.com/knowledge/servers/others/2012/1015/15198.html
http://www.dedecms.com/knowledge/servers/others/2012/1015/15198_2.html