Hive总结4 数据压缩

> 在实际工作当中，hive当中处理的数据，一般都需要经过压缩，前期我们在学习hadoop的时候，已经配置过hadoop的压缩，我们这里的hive也是一样的,可以使用压缩来节省我们的MR处理的网络带宽。

MR支持的压缩编码

压缩格式	工具	算法	文件扩展名	是否可切分
DEFAULT	无	DEFAULT	.deflate	否
Gzip	gzip	DEFAULT	.gz	否
bzip2	bzip2	bzip2	.bz2	是
LZO	lzop	LZO	.lzo	否
LZ4	无	无	.lz4	否
Snappy(常用)	无	Snappy	.snappy	否

为了支持多种压缩/解压缩算法，Hadoop引入了编码/解码器，如下表所示：

压缩格式	对应的编码/解码器
DEFLATE	org.apache.hadoop.io.compress.DefaultCodec
gzip	org.apache.hadoop.io.compress.GzipCodec
bzip2	org.apache.hadoop.io.compress.BZip2Codec
LZO	com.hadoop.compression.lzo.LzopCodec
LZ4	org.apache.hadoop.io.compress.Lz4Codec
Snappy	org.apache.hadoop.io.compress.SnappyCodec

压缩性能比较

压缩算法	原始文件大小	压缩文件大小	压缩速度	解压速度
gzip	8.3GB	1.8GB	17.5MB/s	58MB/s
bzip2	8.3GB	1.1GB	2.4MB/s	9.5MB/s
LZO	8.3GB	2.9GB	49.3MB/s	74.6MB/s

`snappy`压缩达到了`250MB/s`,解压达到了`500MB/s`,这性能直接碾压上面所列举的那几个!所以`snappy`也常作为企业数据压缩格式!

压缩参数设置

参数	默认值	阶段	建议
io.compression.codecs （在core-site.xml中配置）	org.apache.hadoop.io.compress.DefaultCodec, org.apache.hadoop.io.compress.GzipCodec, org.apache.hadoop.io.compress.BZip2Codec, org.apache.hadoop.io.compress.Lz4Codec	输入压缩	Hadoop使用文件扩展名判断是否支持某种编解码器
mapreduce.map.output.compress	false	mapper输出	这个参数设为true启用压缩
mapreduce.map.output.compress.codec	org.apache.hadoop.io.compress.DefaultCodec	mapper输出	使用LZO、LZ4或snappy编解码器在此阶段压缩数据
mapreduce.output.fileoutputformat.compress	false	reducer输出	这个参数设为true启用压缩
mapreduce.output.fileoutputformat.compress.codec	org.apache.hadoop.io.compress. DefaultCodec	reducer输出	使用标准工具或者编解码器，如gzip和bzip2
mapreduce.output.fileoutputformat.compress.type	RECORD	reducer输出	SequenceFile输出使用的压缩类型：NONE和BLOCK

开启Map输出阶段压缩

开启 `map` 输出阶段压缩可以减少 `job` 中 `map` 和 `Reduce task` 间数据传输量。具体配置如下：

案例实操
<1>开启hive中间传输数据压缩功能

```
hive(default)>sethive.exec.compress.intermediate=true;
```

<2>开启mapreduce中map输出压缩功能

```
hive(default)>setmapreduce.map.output.compress=true;
```

<3>设置mapreduce中map输出数据的压缩方式

```
hive(default)>setmapreduce.map.output.compress.codec=
org.apache.hadoop.io.compress.SnappyCodec;
```

<4>执行查询语句

```
select count(1) from score;
```

开启Reduce输出阶段压缩

当`Hive`将输出写入到表中时，输出内容同样可以进行压缩。属性`hive.exec.compress.output`控制着这个功能。用户可能需要保持默认设置文件中的默认值`false`，这样默认的输出就是非压缩的纯文本文件了。用户可以通过在查询语句或执行脚本中设置这个值为`true`，来开启输出结果压缩功能。

案例实操

1. 开启hive最终输出数据压缩功能

```
hive (default)>set hive.exec.compress.output=true;
```

2.开启mapreduce最终输出数据压缩

```
hive(default)>set mapreduce.output.fileoutputformat.compress=true;
```

3. 设置mapreduce最终数据输出压缩方式

```
hive (default)> set mapreduce.output.fileoutputformat.compress.codec =
org.apache.hadoop.io.compress.SnappyCodec;
```

4. 设置mapreduce最终数据输出压缩为块压缩

```
hive(default)>set mapreduce.output.fileoutputformat.compress.type=BLOCK;
```

5. 测试一下输出结果是否是压缩文件

```
insert overwrite local directory '/export/servers/snappy' select * from score
distribute by s_id sort by s_id desc;
```