MapReduce压缩方式

可回答：1）Hadoop常见的压缩算法有哪些?

问过的一些公司：网易云音乐(2022.11)，阿里(2020.08)

参考答案：

1、MapReduce支持的压缩方式

压缩格式	hadoop自带?	算法	文件扩展名	是否可切分	换成压缩格式后，原来的程序是否需要修改
DEFLATE	是，直接使用	DEFLATE	.deflate	否	和文本处理一样，不需要修改
Gzip	是，直接使用	DEFLATE	.gz	否	和文本处理一样，不需要修改
bzip2	是，直接使用	bzip2	.bz2	是	和文本处理一样，不需要修改
LZO	否，需要安装	LZO	.lzo	是	需要建索引，还需要指定输入格式
Snappy	否，需要安装	Snappy	.snappy	否	和文本处理一样，不需要修改

2、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

3、压缩方式选择

压缩方式选择时重点考虑：压缩/解压缩速度、压缩率（压缩后存储大小）、压缩后是否可以支持切片。

`Gzip压缩`

优点：压缩率比较高，而且压缩/解压速度也比较快；Hadoop本身支持，在应用中处理Gzip格式的文件就和直接处理文本一样；大部分Linux系统都自带Gzip命令，使用方便。

缺点：不支持Split。

应用场景：当每个文件压缩之后在130M以内的（1个块大小），都可以考虑使用Gzip压缩格式。例如一天或者一个小时的日志压缩成一个Gzip文件。

`Bzip2压缩`

优点：支持split；具有很高的压缩率，比gzip压缩率都高；hadoop本身支持，但不支持native；在linux系统下自带bzip2命令，使用方便。

缺点：压缩/解压速度慢；不支持native。

应用场景：适合对速度要求不高，但需要较高的压缩率的时候，可以作为mapreduce作业的输出格式；或者输出之后的数据比较大，处理之后的数据需要压缩存档减少磁盘空间并且以后数据用得比较少的情况；或者对单个很大的文本文件想压缩减少存储空间，同时又需要支持split，而且兼容之前的应用程序（即应用程序不需要修改）的情况。

Lzo压缩

优点：压缩/解压速度也比较快，合理的压缩率；支持split，是hadoop中最流行的压缩格式；可以在linux系统下安装lzop命令，使用方便。

缺点：压缩率比gzip要低一些；hadoop本身不支持，需要安装；在应用中对lzo格式的文件需要做一些特殊处理（为了支持split需要建索引，还需要指定inputformat为lzo格式）。

应用场景：一个很大的文本文件，压缩之后还大于200M以上的可以考虑，而且单个文件越大，lzo优点越越明显。

Snappy压缩

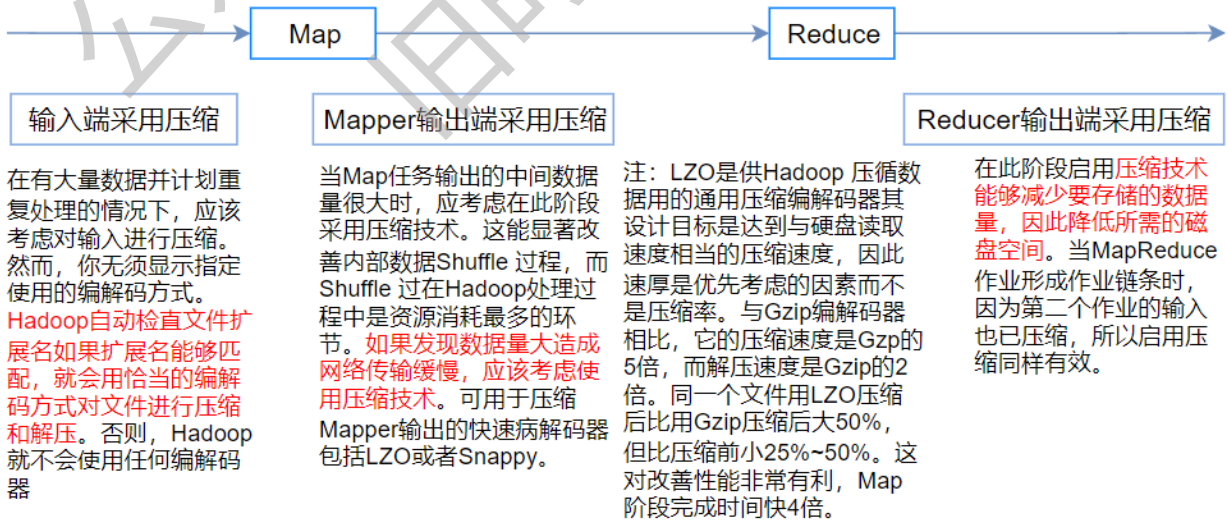
优点：高速压缩速度和合理的压缩率。

缺点：不支持split；压缩率比gzip要低；hadoop本身不支持，需要安装；

应用场景：当Mapreduce作业的Map输出的数据比较大的时候，作为Map到Reduce的中间数据的压缩格式；或者作为一个Mapreduce作业的输出和另外一个Mapreduce作业的输入。

4、压缩位置选择

压缩可以在MapReduce作用的任意阶段启用



输入端采用压缩

在有大量数据并计划重复处理的情况下，应该考虑对输入进行压缩。然而，你无须显示指定使用的编解码方式。Hadoop自动检查文件扩展名，如果扩展名能够匹配，就会用恰当的编解码方式对文件进行压缩和解压。否则，Hadoop就不会使用任何编解码器。

mapper输出端采用压缩

当map任务输出的中间数据量很大时，应考虑在此阶段采用压缩技术。这能显著改善内部数据Shuffle过程，而Shuffle过程在Hadoop处理过程中是资源消耗最多的环节。如果发现数据量大造成网络传输缓慢，应该考虑使用压缩技术。可用于压缩mapper输出的快速编解码器包括LZO或者Snappy。

reducer输出采用压缩

在此阶段启用压缩技术能够减少要存储的数据量，因此降低所需的磁盘空间。当mapreduce作业形成作业链条时，因为第二个作业的输入也已压缩，所以启用压缩同样有效。