大数据开发技术

东北林业大学

卢洋

第四章

Hadoop数据压缩

4.3

压缩方式的选择

Gzip

- 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130MB内的(一个块大小内)，都可以考虑用Gzip格式;
  例如: 一天或者一个小时的日志压缩成一个Gzip文件.

Bzip2

可切片，还不用改

- 优点:
  支持Split; 具有较高的压缩率，比Gzip压缩率高; Hadoop自带，使用方便.
- 缺点: 压缩/解压速度慢.
- 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; 或者输出之后的数据需要压缩存档减少磁盘空间并且以后数据使用较少的情况; 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的情况.

Lzo

- 优点:
  压缩/解压速度比较快，合理的压缩率；支持Split，是Hadoop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优势越明显.

Snappy

- 优点:
  高压缩速率和合理的压缩率.

- 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作业的输出和另一个MapReduce作业的输入.