大数据开发技术

东北林业大学

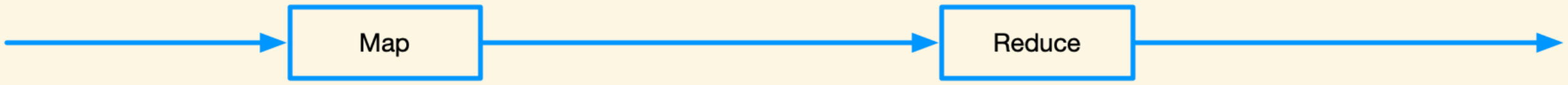
卢洋

第四章

Hadoop数据压缩

4.4

压缩位置的选择



输入端采用压缩

在有大量数据并计划重复处理的情况下，应该考虑对输入数据进行压缩。无须显式地指定使用的压缩编码方式。Hadoop自动检测文件扩展名，如扩展名能够匹配，就会用恰当的编解码方式对文件进行压缩和解压。否则，Hadoop不会使用任何编解码器。

Mapper输出采用压缩

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Mapper输出的快速编解码器包括Lzo和Snappy。

注：LZO是供Hadoop压缩数据用的通用压缩编解码器。其设计目标是达到与硬盘读取速度相当的压缩速度，因此速度是优先考虑的因素，而不是压缩率。与Gzip编解码器相比，它的压缩速度是Gzip的5倍，而解压速度与Gzip的2倍。同一个文件用LZO压缩后比用Gzip压缩后大50%，但比压缩前小25%~50%，这对改善性能非常有利，Map阶段完成时间快4倍。

Reducer输出采用压缩

在此阶段启用压缩技术能够减少待存储的数据量，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。