知道相关数据压缩的类型即可

大数据开发技术

东北林业大学

卢洋

第四章

Hadoop数据压缩

4.1

概述

压缩概述

- 压缩技术能够有效减少底层存储系统(HDFS)读写字节数。
- 压缩提高了网络带宽和磁盘空间的效率。
- 在运行MR程序时，I/O操作、网络数据传输、Shuffle和Merge要花费大量的时间，尤其是数据规模很大和工作负载密集的情况下。
- 因此，使用数据压缩显得非常重要。

压缩概述

- 鉴于磁盘I/O和网络带宽是Hadoop的宝贵资源，数据压缩对于节省资源、最小化磁盘I/O和网络传输非常有帮助。
- 可以在任意MapReduce阶段启用压缩。
- 不过，尽管压缩与解压操作的CPU开销不高，其性能的提升和资源的节约并非没有代价。

压缩策略和原则

也会有副作用

- 压缩是提供Hadoop运行效率的一种优化策略。
- 通过对Mapper、Reducer运行过程的数据进行压缩，以减少磁盘I/O，提高MR程序运行速度。
- 注意：采用压缩技术减少了磁盘I/O，但同时增加了CPU运算负担。所以，压缩特性运用得当能提高性能，但运用不当也可能降低性能。

压缩策略和原则

- 压缩基本原则：

(1) 运算密集型的job，少用压缩；

(2) IO密集型的job，多用压缩.

文件大，一直在写