大数据开发技术

东北林业大学

卢洋

第三章

MapReduce框架原理

3.1

InputFormat数据输入

1 切片与MapTask并行度决定机制

(1) 问题引出

MapTask的并行度决定Map阶段的任务处理并发度，进而影响整个Job的处理速度。

思考：1G的数据，启动8个MapTask，可以提高集群的并发处理能力。那么，1K的数据也启动8个MapTask，会提高性能嘛？MapTask是否越多越好？哪些因素影响MapTask并行度？

(2) MapTask并行度决定机制

数据块：Block是HDFS物理上把数据分成一块一块。

数据切片：数据切片只是在逻辑上对输入进行切分，并不会在磁盘上将其切分成片进行存储。

数据切片是逻辑上的概念，与数据块无关

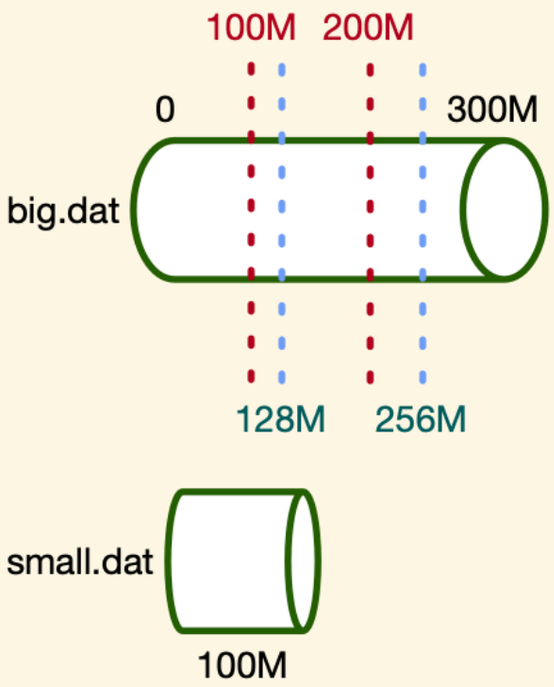
切片与块的大小一致，方便网络运输

并行度由切片数量决定，切片数量由文件大小等决定

1 切片与MapTask并行度决定机制

1 假设切片大小设置为100M

2 假设切片大小设置为128M



1）一个Job的Map阶段并行度由客户端在提交Job时的切片数决定

2）每一个Split切片分配一个MapTask并行实例处理

3）默认情况下，切片大小=BlockSize

4）切片时不考虑数据集整体，而是逐个对每一个文件进行切片