大数据开发技术

东北林业大学

卢洋

第三章

MapReduce框架原理

3.4

CombineTextInputFormat机制

- 框架默认的TextInputFormat切片机制是对任务按文件规划切片；
- 不管文件大小，都会是一个单独的切片，都会交给一个MapTask；
- 如果有大量的小文件，就会产生大量的MapTask，效率极低.

1 应用场景

- CombineTextInputFormat用于小文件过多的场景；
- 可以将多个小文件从逻辑上规划到一个切片中；
- 这样，多个小文件就可以交给一个MapTask处理.

2 虚拟存储切片最大值设置

- CombineTextInputFormat.setMaxInputSplitSize(job, 4194304);

单位比特，4m

- 注意：虚拟存储切片最大值设置最好根据实际的小文件大小情况来设置具体的值.

setMaxInputSplitSize值为4M

10m，先切出去4m,剩6m,6<2*4，故均分成两块。故切三块，4，3，3

虚拟存储过程

切片过程

a.txt 1.7M

1.7M<4M划分一块

b.txt 5.1M

5.1M>4M但是小于2*4M，划分两块：
块1：2.55M；块2：2.55M

c.txt 3.4M

3.4M<4M划分一块

d.txt 6.8M

6.8M>4M但是小于2*4M，划分两块：
块1：3.4M；块2：3.4M

最终存储文件：
1.7M
2.55M
2.55M
3.4M
3,4M
3.4M

(1) 判断虚拟存储的文件大小是否大于setMaxInputSplitSize值，大于等于则单独形成一个切片。

(2) 如果不大于，则跟下一个虚拟存储文件进行合并，共同形成一个切片。

最终会形成3个切片，大小分别为：
(1.7+2.55)M
(2.55+3.4)M
(3.4+3.4)M

最后形成三个切片<4