大数据开发技术

东北林业大学

卢洋

第三章

MapReduce框架原理

Combiner

选择填空重点

Combiner合并

(1) Combiner是MR程序中Mapper和Reducer之外的一种组件；

(2) Combiner组件的父类是Reducer；

(3) Combiner和Reducer的区别在于运行的位置：

Combiner是在每一个MapTask所在节点运行；

运行过程中

Reducer是接收全局所有Mapper的输出结果；

(4) Combiner的意义就是对每一个MapTask的输出进行局部汇总，以减少网络传输量；

Reducer是全局汇总

(5) Combiner能够应用的前提是不能影响最终的业务逻辑；而且，Combiner的输出KV应该与Reducer的输入KV类型对应。

什么叫不影响业务逻辑?

(1) Mapper

3 5 7 -> (3+5+7)/3=5

2 6 -> (2+6)/2=4

什么叫不影响业务逻辑?

(1) Mapper

3 5 7 -> (3+5+7)/3=5

2 6 -> (2+6)/2=4

(2) Reducer

(3+5+7+2+6)/5=23/5 ≠ (5+4)/2=9/2

自定义Combiner的实现步骤

(1) 自定义一个Combiner，继承Reducer，重写reduce()方法；

(2) 在Job类中设置

job.setCombinerClass(WordcountCombiner.class);