大数据开发技术

东北林业大学

卢洋

第一章

MapReduce概述

1.8

WordCount案例

1 需求

- 在给定的文本文件中统计输出每一个单词出现的总次数.

输入数据

[hello.txt]

root root

student student

teacher teacher

one

two

three

hadoop

期望输出

root	2
student	2
teacher	2
one	1
two	1
three	1
hadoop	1

2 需求分析

- 按照MapReduce编程规范，分别编写Mapper、Reducer、Driver.

1 输入数据:

root root
student student
teacher teacher
one
two
three
hadoop

2 输出数据:

root 2
student 2
teacher 2
one 1
two 1
three 1
hadoop 1

3 Mapper

3.1 将MapTask输入的文本内容转换成String

root root

3.2 根据空格将这一行切分为单词

root
root

3.3 将单词输出为<单词, 1>

root, 1
root, 1

4 Reducer

4.1 汇总各个Key的个数

root, 1
root, 1

4.2 输出该Key的总次数

root, 2

5. Driver

5.1 获取配置信息，获取Job对象实例

5.2 指定本程序的jar包所在的本地路径

5.3 关联Mapper/Reducer业务类

5.4 指定Mapper输出数据的KV类型

5.5 指定最终输出的数据的KV类型

5.6 指定Job的输入原始文件所在路径

5.7 指定Job的输出结果所在路径

5.8 提交作业

3 环境准备

- 创建maven工程;
- 填写依赖;
- 配置log4j.properties;
- 创建package.

4 编写程序

- Mapper;
- Reducer;
- Driver.