**课后扩展题作业**

# 实操一、西游记如来出现了多少次。

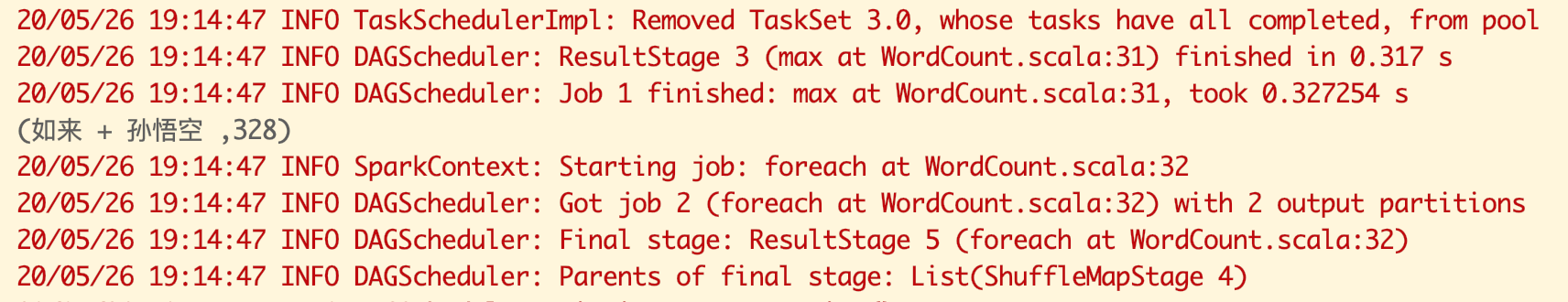
1. 扩展题

改一改代码，把如来和孙悟空出现的次数加起来，并把结果打印出来。例如：如来 + 孙悟空，234

1. 关键代码

*//从西游记小说中读取数据* **val** rdd = sc.textFile(**"/Users/boscoli/Downloads/大数据实训\_品高/2/西游记.txt"**)  
 .map { x =>  
 **val** str = **if** (x.length > 0)  
 *//开始分词，将结果使用逗号隔开*ToAnalysis.*parse*(x).recognition(filter).toStringWithOutNature(**","**)  
 str.toString  
 }.flatMap(x => x.split(**","**))  
 **val** result = rdd.map((\_, 1)).groupBy(\_.\_1).mapValues(\_.size).filter(x => StringUtils.*isNotBlank*(x.\_1))  
*// result.foreach(println)  
 //统计如来出现的次数并输出结果  
// result.filter(x => x.\_1.equals("如来")).foreach(println)  
 //统计如来和孙悟空共同出现的次数并输出结果* **val** a = result.filter(x => x.\_1.equals(**"如来"**))  
 **val** b = result.filter(x => x.\_1.equals(**"孙悟空"**))  
 *println*(**"如来 + 孙悟空 "**, a.max().\_2 + b.max().\_2)  
 result.filter(x => x.\_1.equals(**"如来"**)).values.foreach(*println*)

1. 结果截图



# 实操二、大数据计算和存储分离会怎样？

1. 扩展题

word\_count\_result.txt文件，请大家各自上传到S3的bucket下，通过外部表查询里面的内容，单独把如来的出现次数查出来。

1. 建表SQL脚本
2. 创建外表

CREATE EXTERNAL TABLE tt

(mz STRING, cs STRING)

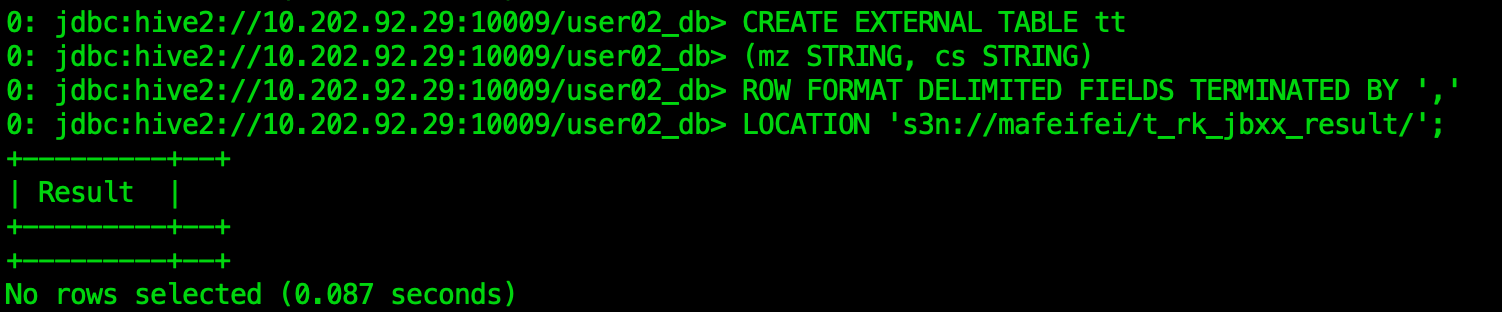
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION 's3n://mafeifei/t\_rk\_jbxx\_result/';

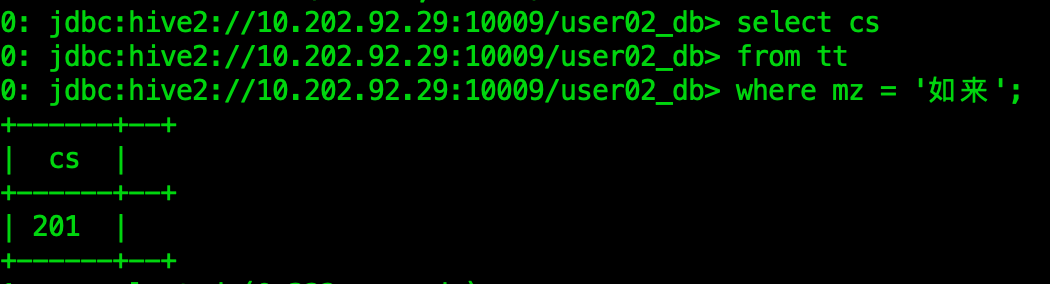
1. SQL查询语句

Select cs from tt where mz = ‘如来’；

1. 结果截图
2. 创建外表



2）查询结果



# 实操三、大数据框架打擂台。

1. 扩展题

使用SparkSQL或Greenplum计算框架，查询人数排名前五的职业是哪几个？

1. 查询SQL语句

**Select t t.zy as “职业”**

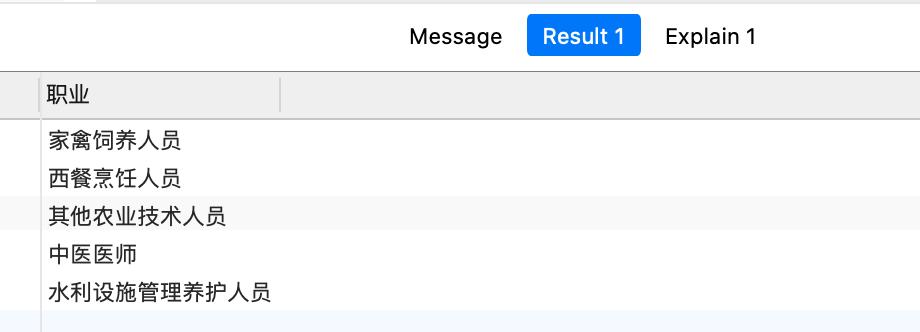
**From t\_rk\_jbxx t**

**Group by t.zy**

**Order by “count”(t.zy) DESC**

**LIMIT 5**

1. 结果截图



# 实操四、把大数据整合到你的软件。

1. 扩展题

使用JDBC的方式，从Greenplum中选取一个表，数据读出来。

1. 关键代码

**import** java.sql.DriverManager  
**val** url = **"jdbc:postgresql://bigdata28.depts.bingosoft.net:25432/user02\_db"  
val** properties = **new** Properties()  
properties.setProperty(**"driverClassName"**, **"org.postgresql.Driver"**)  
properties.setProperty(**"user"**, **"user02"**)  
properties.setProperty(**"password"**, **"pass@bingo2"**)  
  
**val** connection = DriverManager.*getConnection*(url, properties)  
**val** statement = connection.createStatement

*println*(**"表数据："**)  
**var** resultSet = statement.executeQuery(**"select** *\** **from t\_rk\_jbxx limit 10"**)  
**while** (resultSet.next) {  
 **val** asjbh = resultSet.getString(4)  
 **val** ajmc = resultSet.getString(1)  
 **val** bamjbh = resultSet.getString(8)  
 *//输出表数据  
 println*(**s"$**asjbh **｜$**ajmc **｜$**bamjbh**"**)

1. 结果截图

