**2020年安徽省大数据与人工智能应用竞赛A卷**

**学校名称： 队伍名称： 队伍编号：**

# 第六部分：综合题目（15分）

### 1、Mapreduce和Spark的都是并行计算，那么他们有什么相同和区别（5分）

答：

相同点：

Mapreduce和spark都是大数据分析工具，使用场景都是在海量数据情况，都是一个计算框架。Mapreduce和spark都是先map读取每一行数据，再分发到reduce集中处理

不同点：

Mapreduce是在硬盘中读取数据并计算，而spark是基于内存的计算，将数据读入到内存中进行计算所以更快，而Mapreduce从硬盘一条一条读取数据所以比spark慢。

Spark使用的是RDD，分布式弹性数据集，其操作更灵活，使用的是计算逻辑，也就是说，在使用spark构建计算逻辑的时候，都只是建立计算的逻辑，并不会真正的对RDD进行操作，只有当最后输出结果的时候才会从原始数据开始计算，知道输出结果，如果中间计算中有计算数据丢失，则可以通过计算逻辑从原始数据导出，因为原始数据一直保存在硬盘中，因此spark不用担心计算出错，每一步的操作不会改变原始的RDD，只是会在原始RDD之上生成新的RDD，非常类似于装饰者模式，对原始数据不操作，生成新的数据。

Spark比mapreduce方便编写。

### 2、利用spark完成下面的统计分析（10分）

数据为某网站的访问日志：

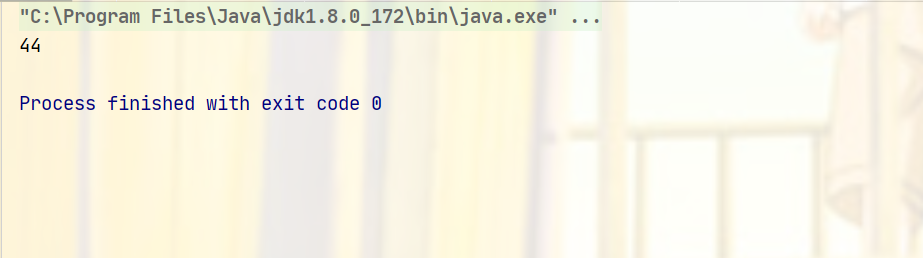


字段分别为 ip 时间 操作方式以及访问地址 返回状态码 数据量字节数

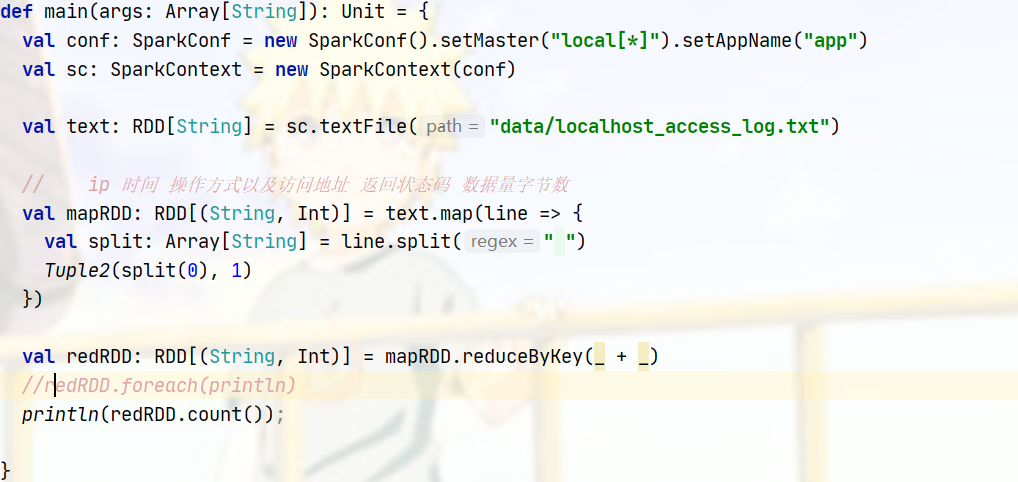
源码：

1：计算独立IP数（2分）

结果截图：

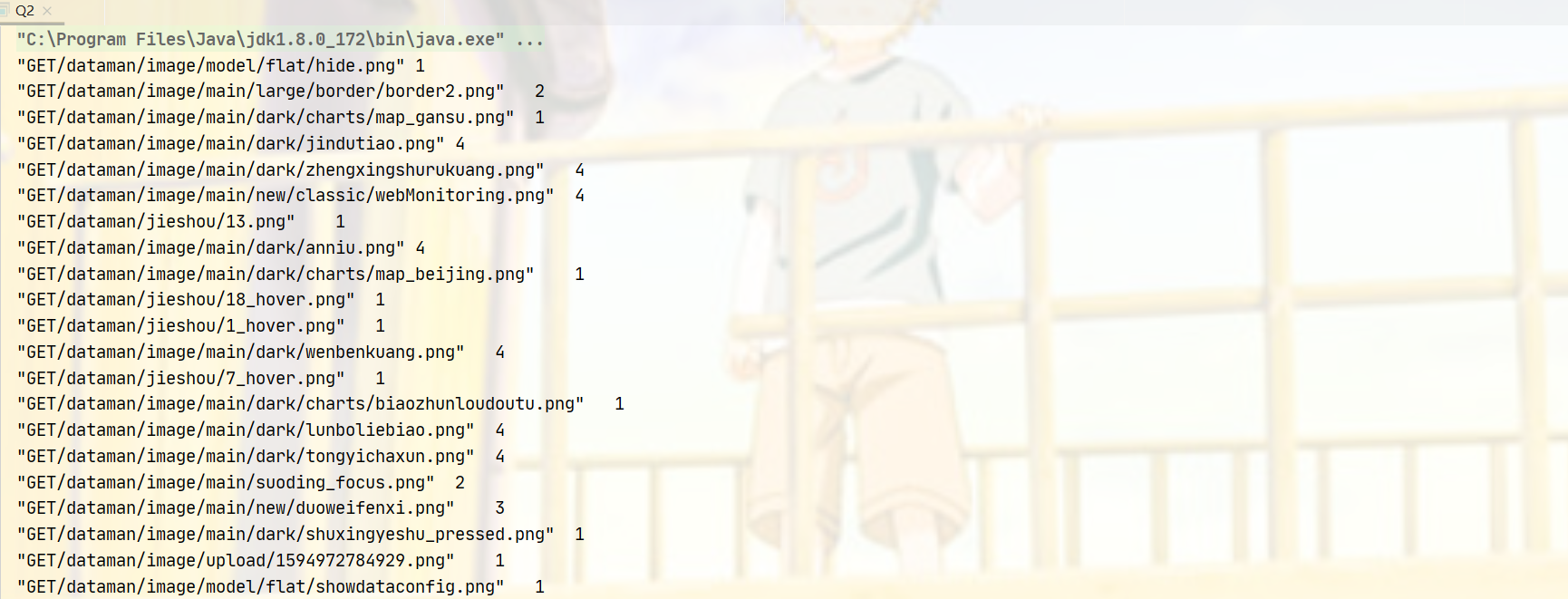


代码截图：



2：统计每个图片独立IP数 （图片为地址中含有.png）（4分）

结果截图：



代码截图：



3：统计一天中每个小时的流量（4分）

结果截图：



代码截图：

