**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：**

1.都是面向大数据的并行处理的开源集群计算环境。

2.Spark的计算模式也属于MapReduce，但不局限于Map和Reduce操作。

3.都是将数据分而治之，MapReduce将数据切分为独立的小数据块，被多个Map任务并行处理后将结果传递给Reduce任务。Spark将数据化为RDD（弹性分布数据集），进行“转换”、“动作”、”持久化”运算。

**区别：**

1.MapReduce属于Hadoop生态体系，Spark属于BDAS生态体系。

2. MapReduce通常将中间结果放到HDFS上，Spark是基于内存并行的大数据框架，中间结果存放到内存，对于迭代数据Spark效率更高。

3. MapReduce会消耗大量时间进行排序，但有些应用场景不需要进行排序，Spark可以避免不必要排序带来的开销。

4. Spark是一张有向无环图（从一个点出发最终无法回到该点的一个拓补），并对其进行优化，因此在迭代工作负载上远高于MapReduce。

### 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：

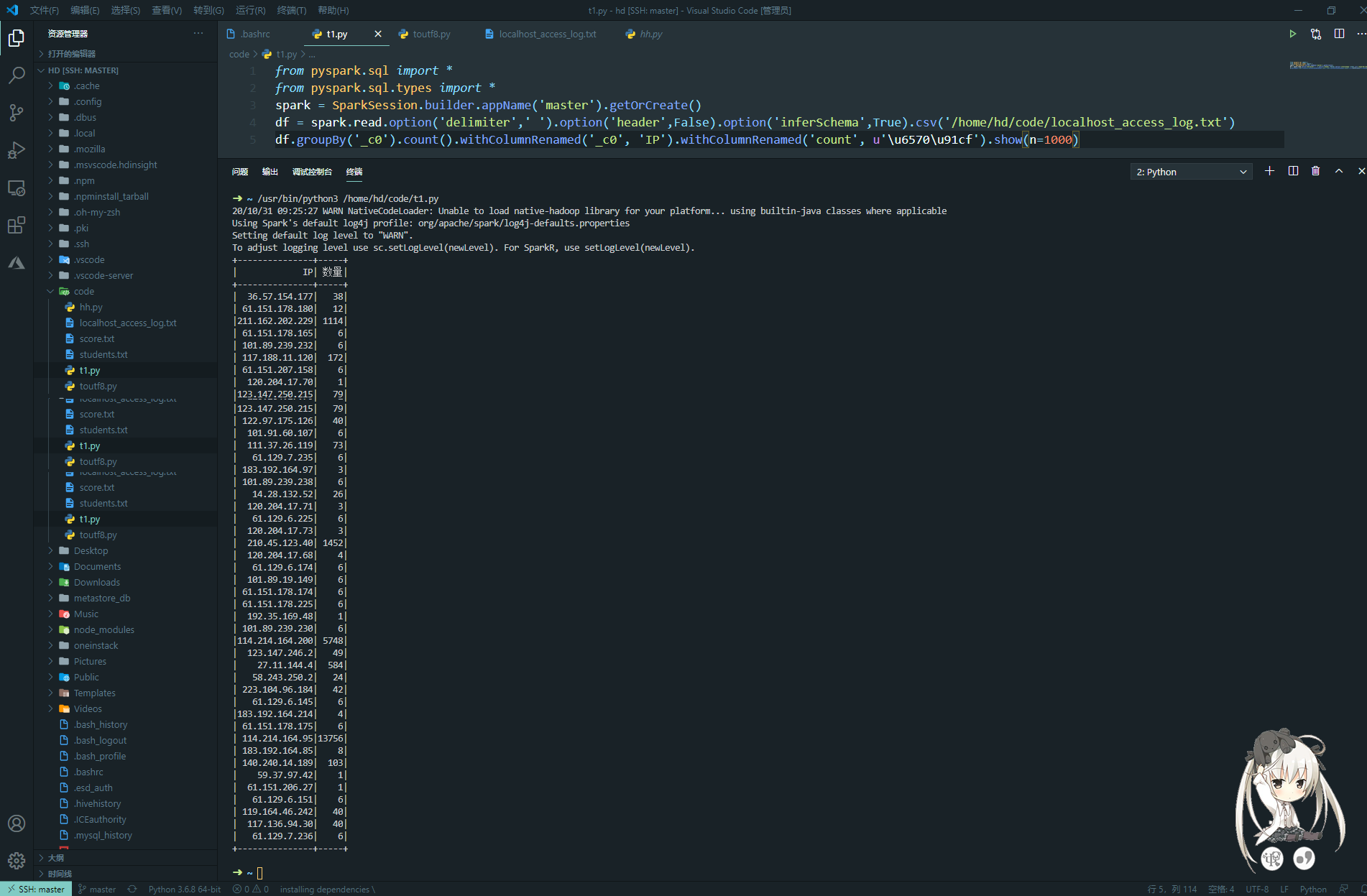


字段分别为 ip 时间 操作方式以及访问地址 返回状态码 数据量字节数

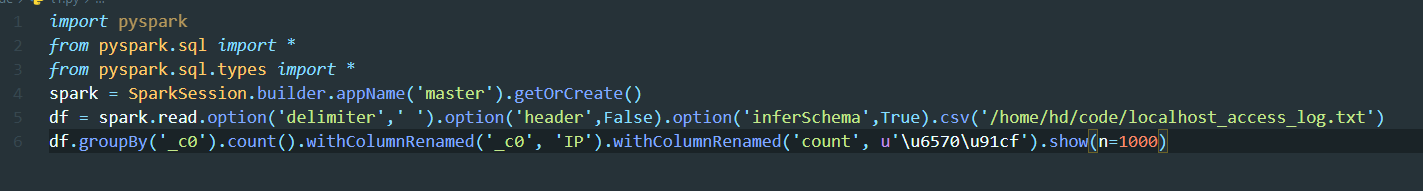
源码：

1：计算独立IP数（2分）

结果截图：

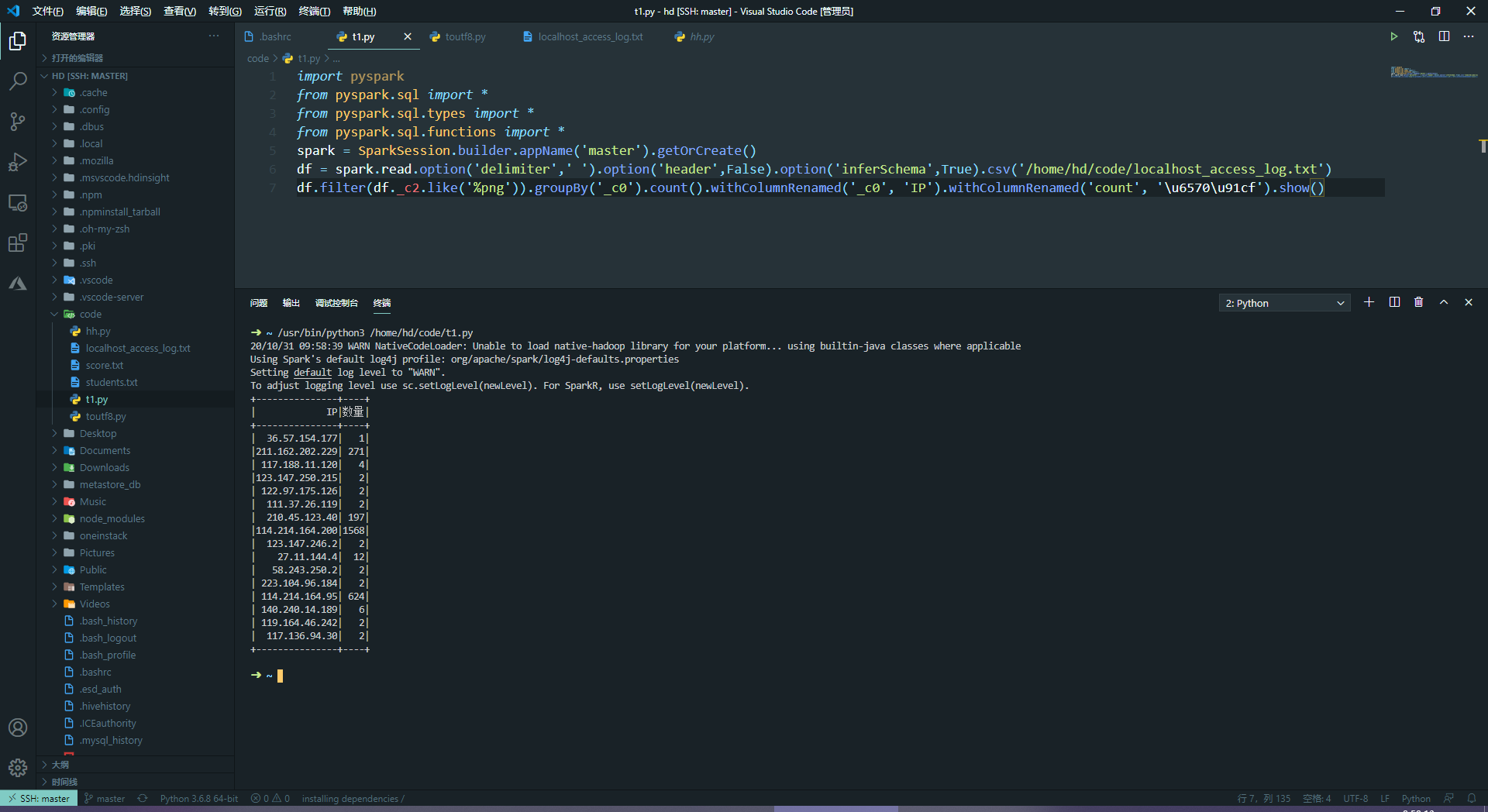


代码截图：

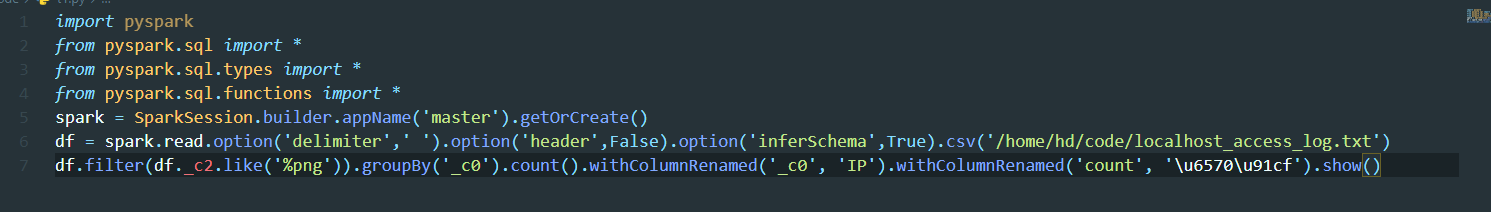


2：统计每个图片独立IP数 （图片为地址中含有.png）（4分）

结果截图：

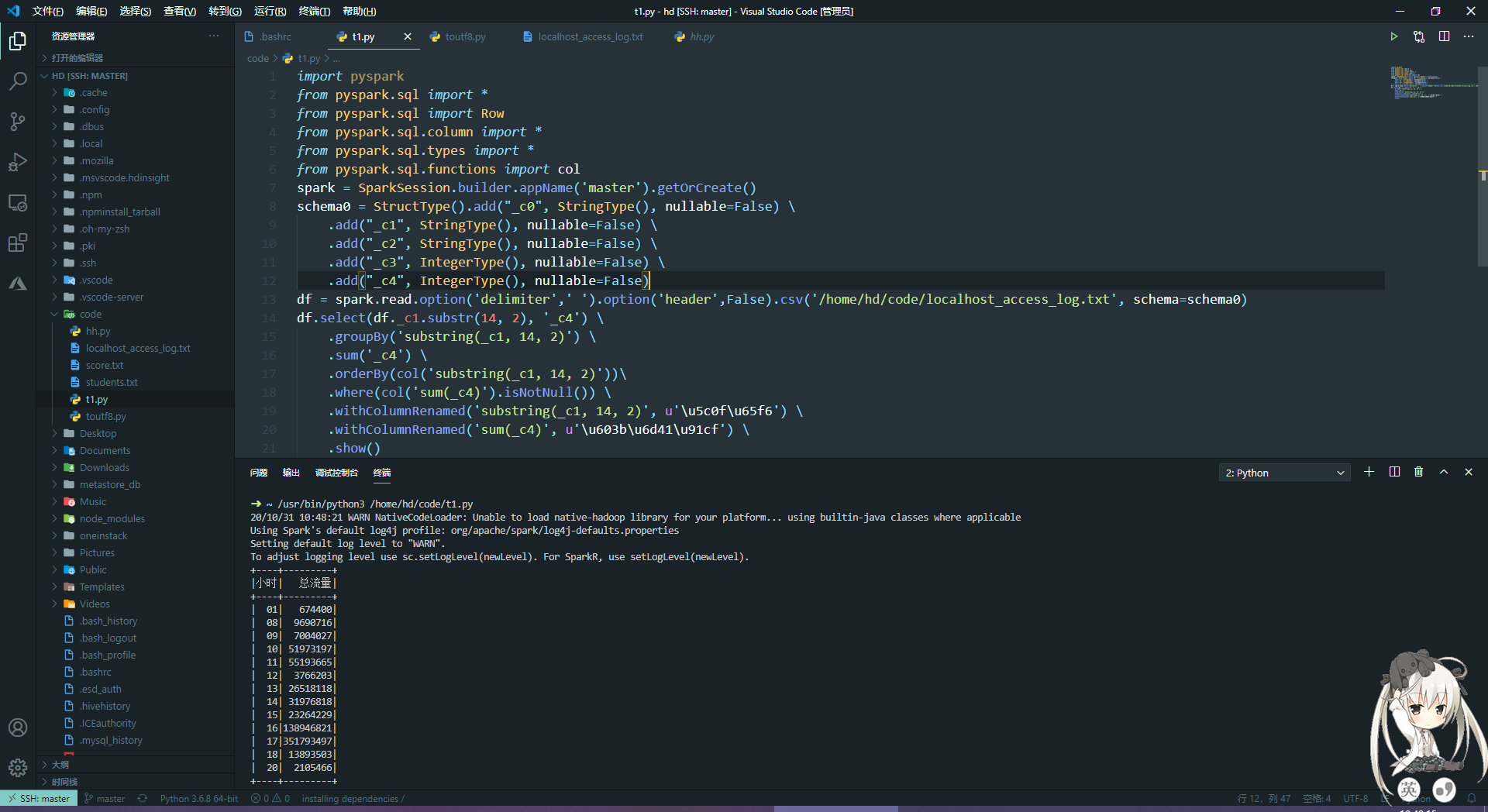


代码截图：



3：统计一天中每个小时的流量（4分）

结果截图：



代码截图：

