**基于大众点评数据的商业选址推荐系统**

**一、应用场景及业务需求**

大众点评作为一款消费点评网站，不仅为用户提供商户信息、消费点评及消费优惠等信息服务，同时亦提供团购、餐厅预订、外卖及电子会员卡等O2O（Online To Offline）交易服务。截止到2015年第一季度，大众点评月活跃用户数超过2亿，收录商户数量超过1400万家，覆盖全国2500多个城市及美国、日本、法国等近百个热门旅游国家和地区。除上海总部之外，大众点评已经在北京、广州、天津、杭州、南京等160多座城市设立分支机构。大众点评月综合浏览量（网站及移动设备）超过150亿，其中移动客户端的浏览量超过85%，移动客户端累计独立用户数超过2亿。

如此的流量基数使得从中爬取的数据可以作为商铺选址的理论依据支撑。在当下快消快餐行业火爆的大环境下，各种火锅、烧烤、奶茶新店雨后春笋，却不少只是昙花一现，开业短短数月便亏得什么也不剩。餐饮品牌固然是一方面，选址更为重要。同一商圈的同质化餐饮评分差异，同一餐饮品牌不同商圈的流量差异，不同商圈消费人群消费水平的差异等，都可以在一定程度上反映出选址的优劣。本系统在大众点评数据基础上，利用pyspark进行数据分析计算，计算的结果动态化展示。同时利用这些数据搭建了推荐系统，面向商户和顾客开放两项功能，1.面向商户综合给出商铺最优化选址；2.面向顾客给出商圈推荐。

**二、数据来源及数据内容**

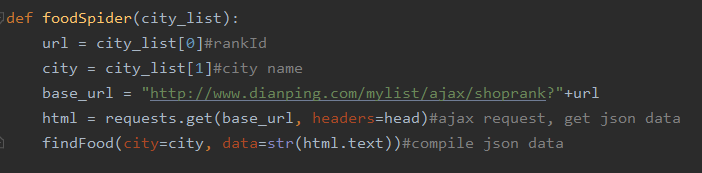
1．从大众点评网站页面上爬取用于获取各城市大众点评排行榜的ajax请求参数，作为spark streaming的输入流，然后向大众点评网发送ajax请求返回json

2．数据内容包括商家网址、商家名、商家所在商圈、商家所在城市、商家所属类型（例如川菜、烧烤等类别）、评分、人均价格

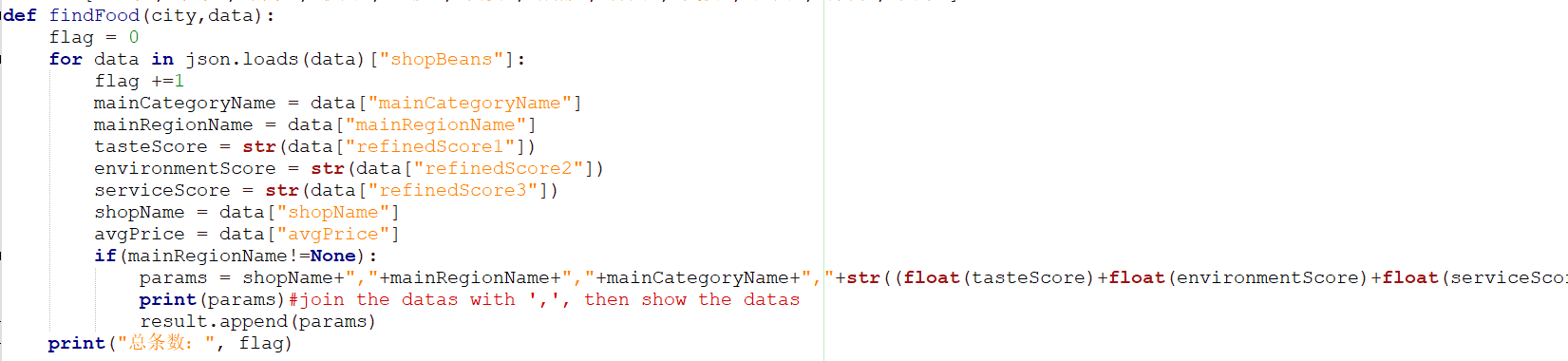
3．数据存储方式：将json以文本形式保存在hdfs文件系统里，简单明了，不做阐述。

4 . 数据预处理：主要用于数据动态化展示

必要的代码解释在代码注释中均有给出



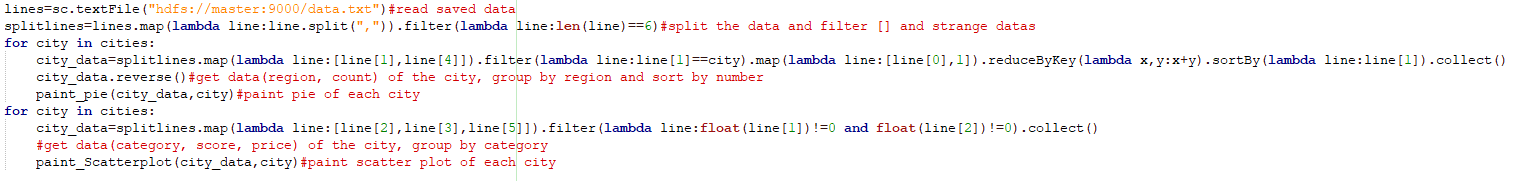
**图表 2.1.1 数据获取**



**图表 2.1.2 数据格式**



**图表 2.1.3 存储方式（也可以不存）**

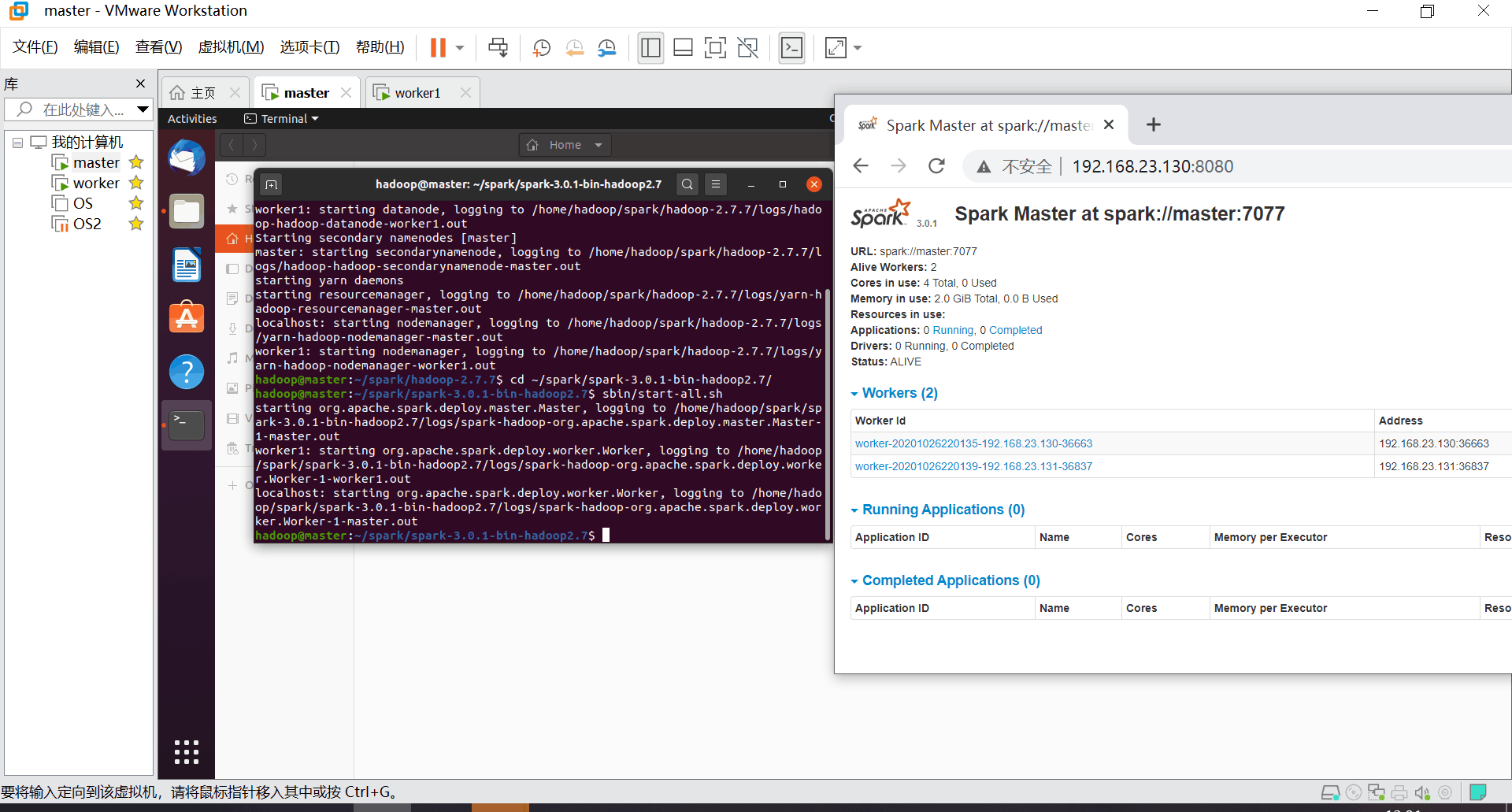


**图表 2.1.4 数据预处理过程**

**三、流计算总体架构**

必要的代码解释在注释中均有给出

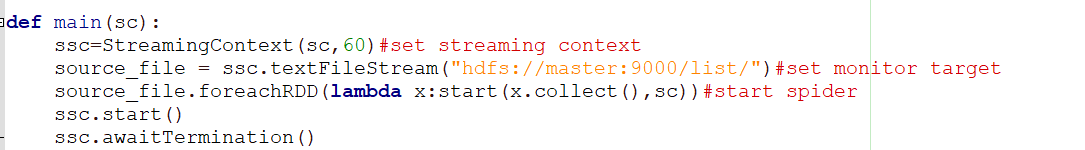
1 . 启动集群



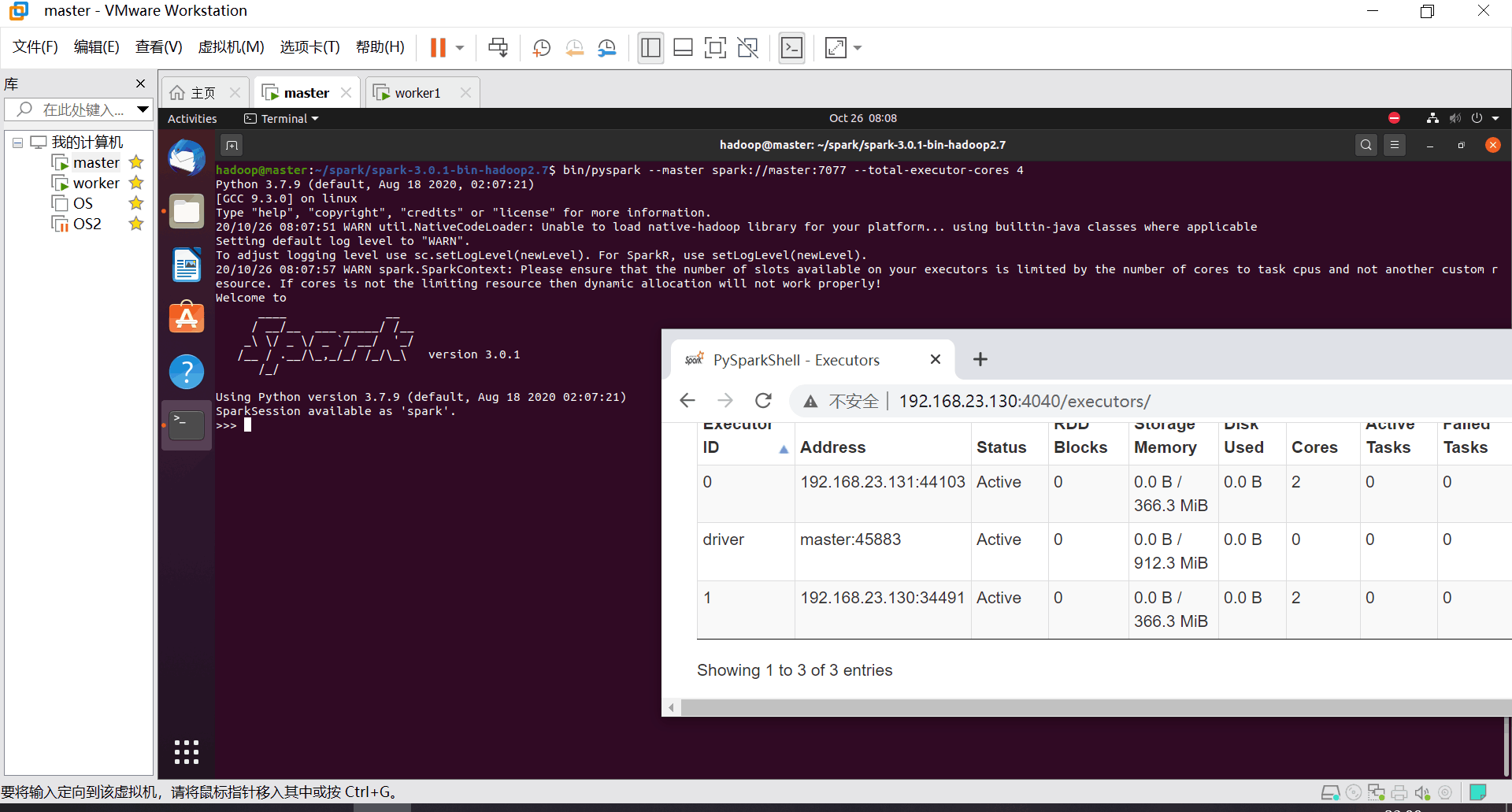
图表 3.1流监听启动

2 . 流准备和监听

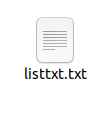
第一行设置流文本，再设置监听目标，然后启动spider

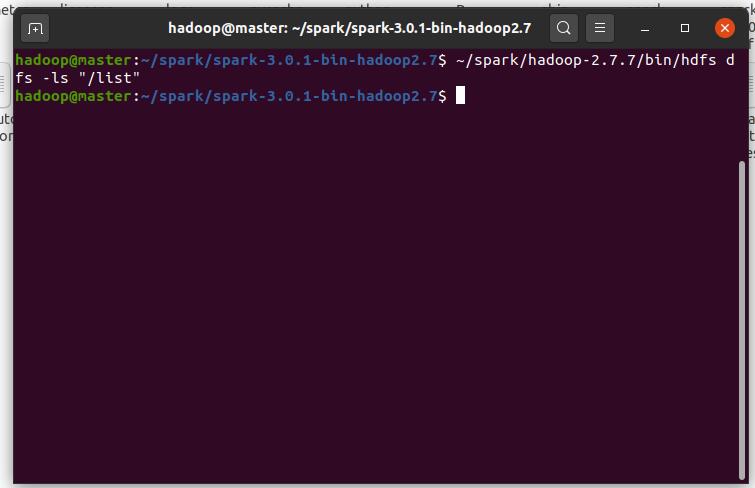


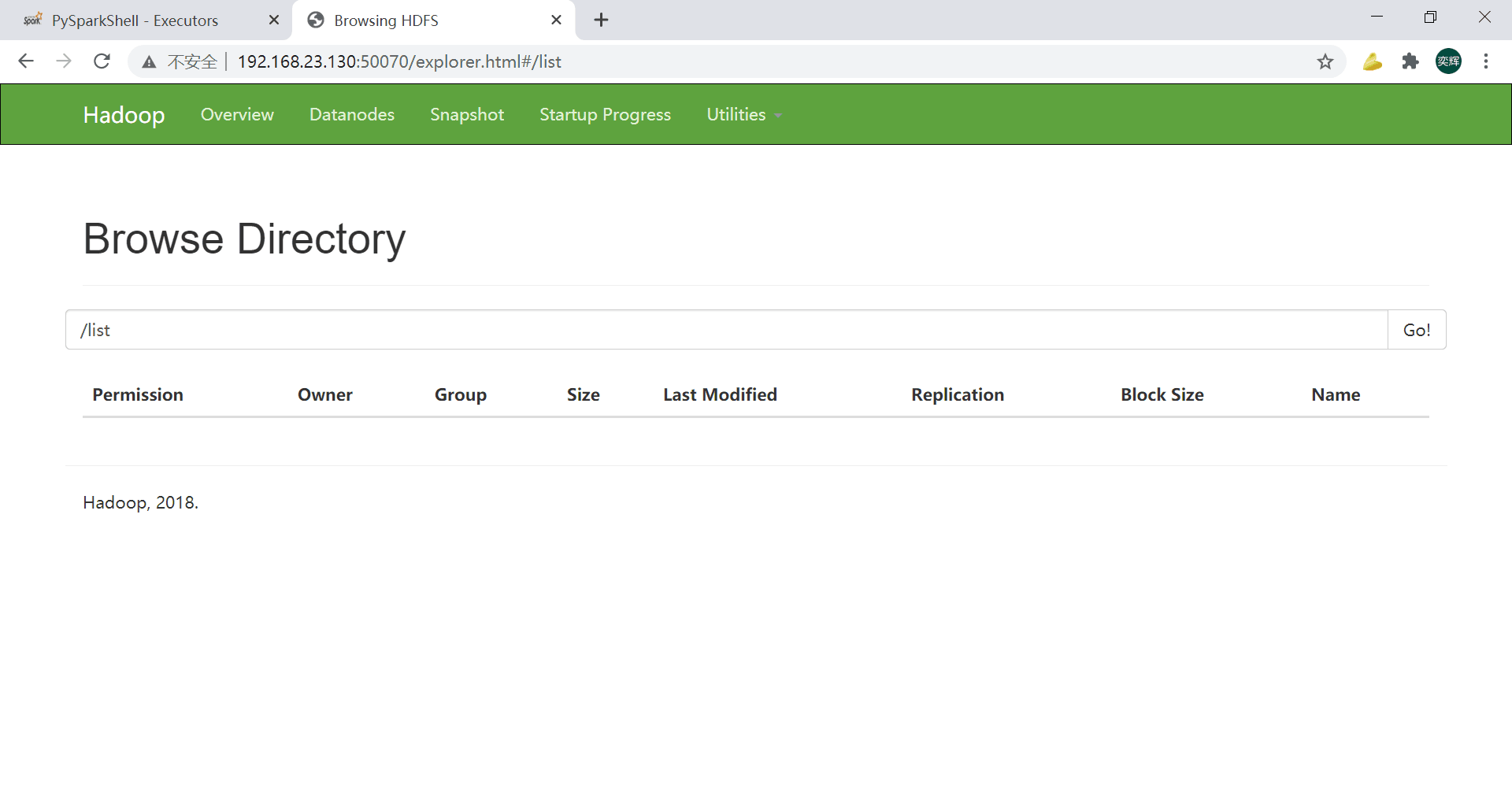
**图表 3.2.1 流准备和监听代码**



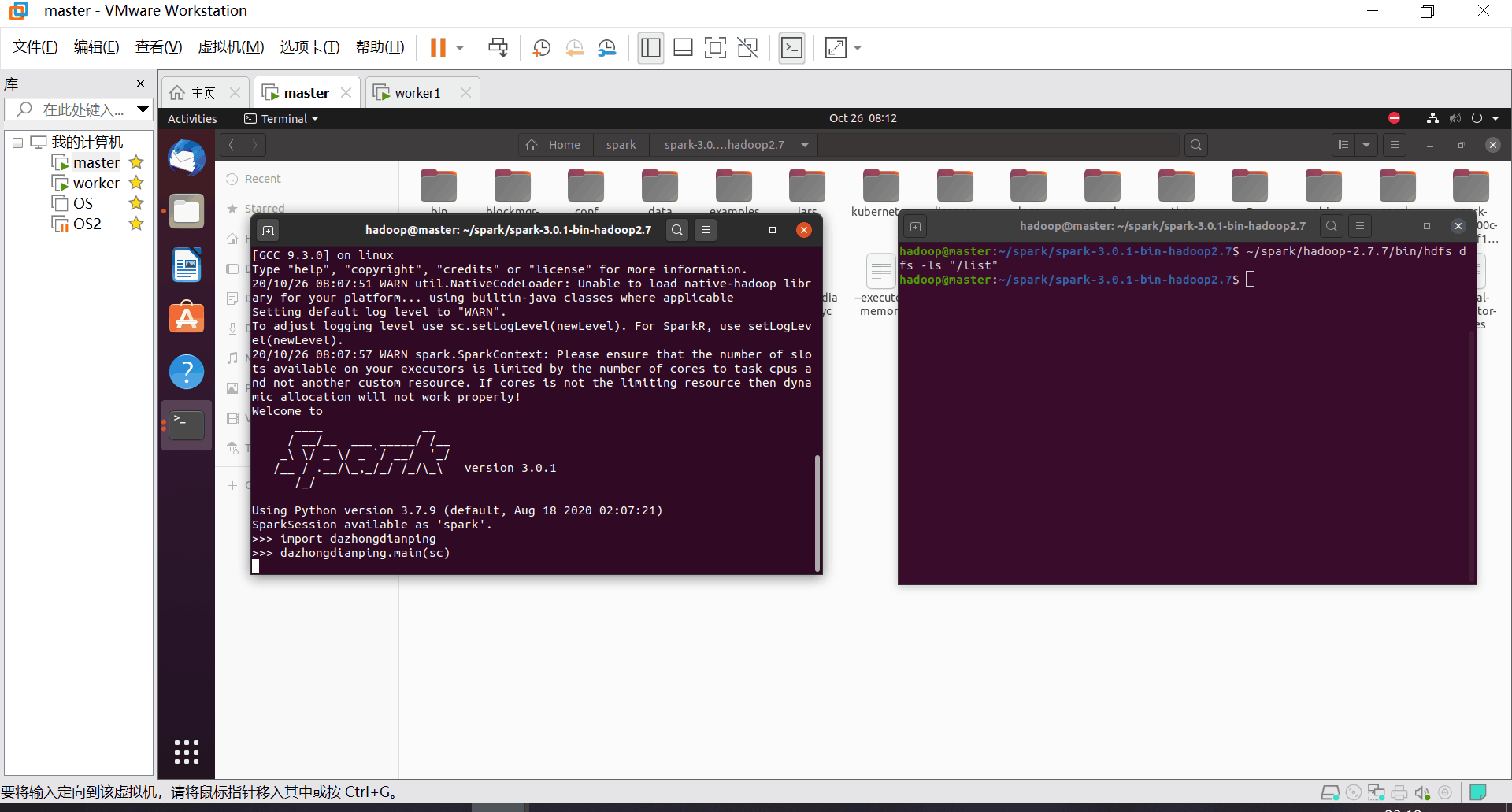
**图表 3.2.2 启动pyspark，查看参与者信息**

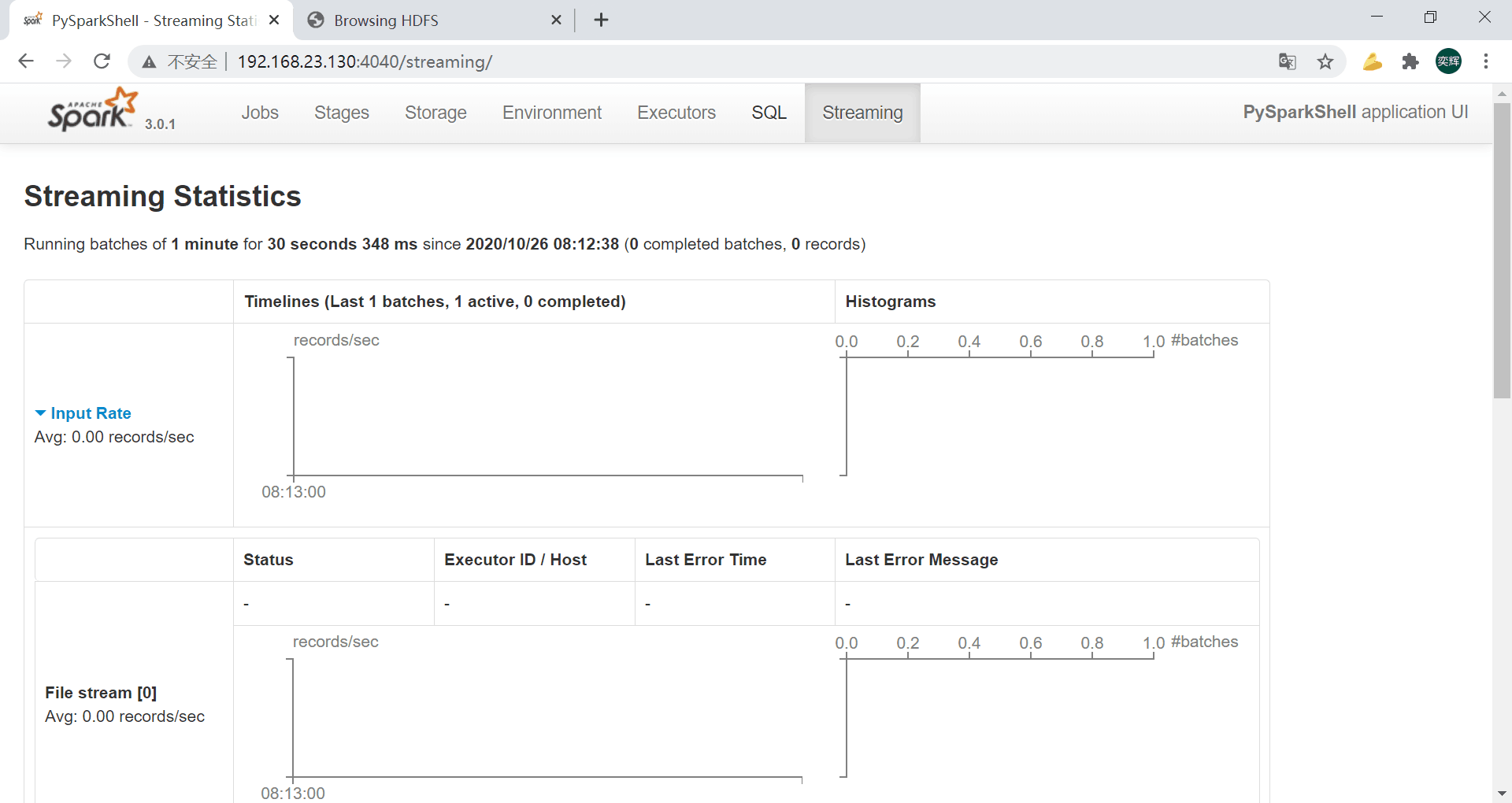
**图表 3.2.3 这是准备作为输入流的文件**

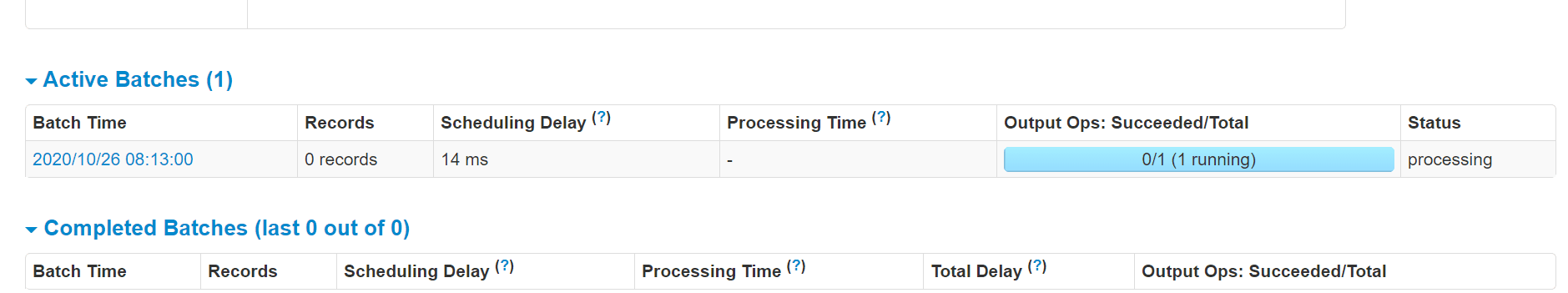




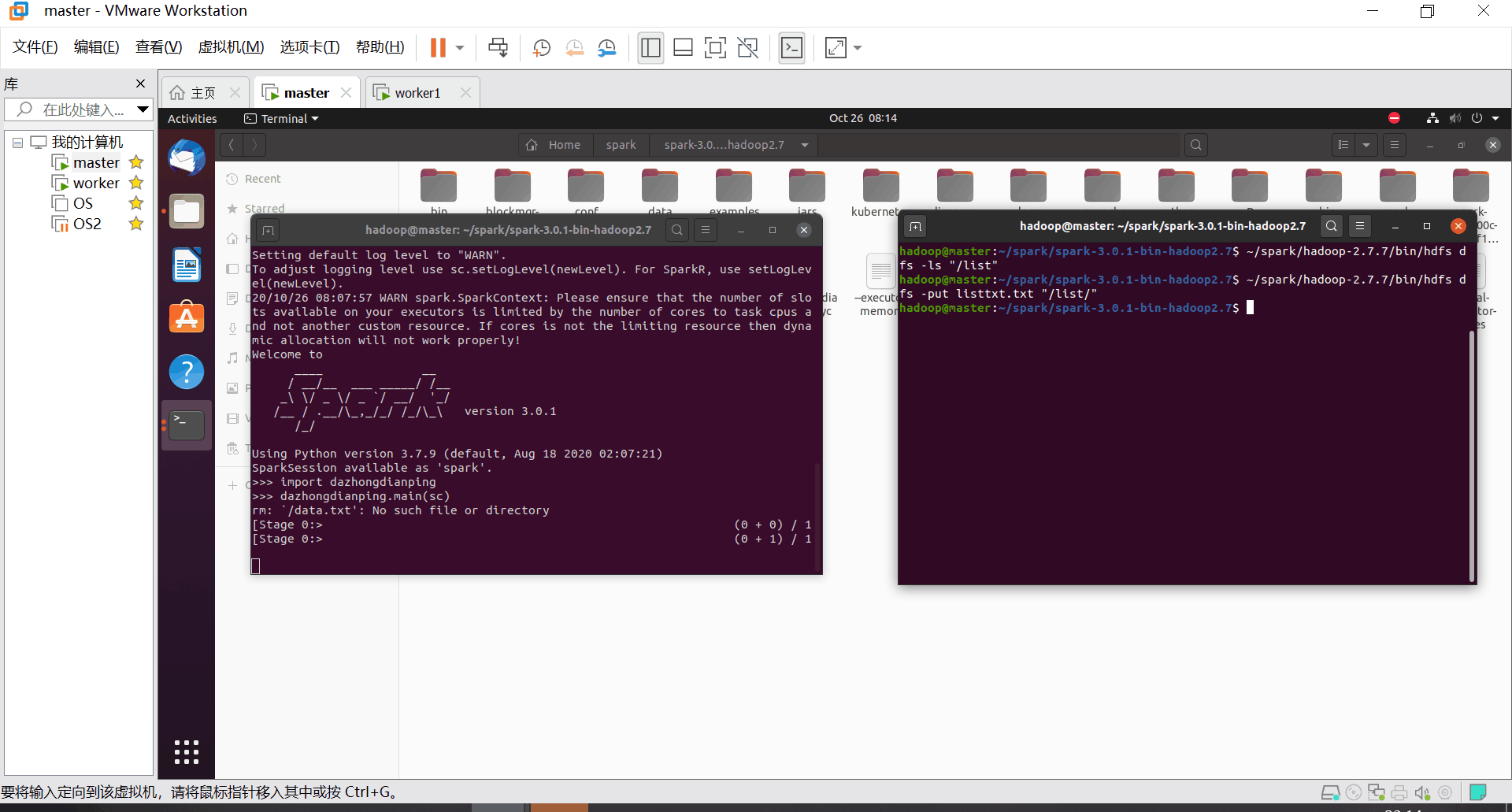
**图表 3.2.4 查看hadoop当前所监听的list文件夹下文件为空**

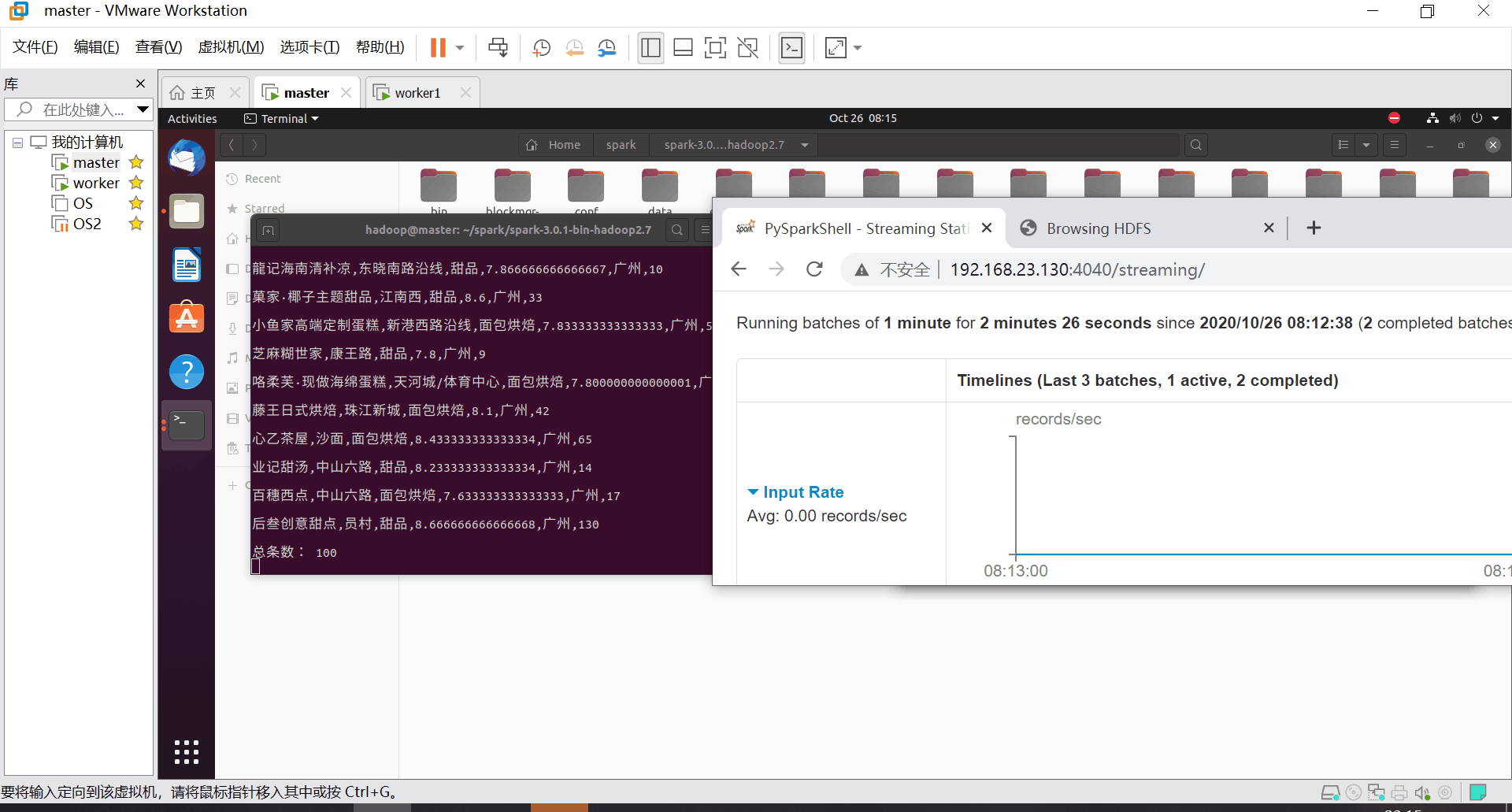


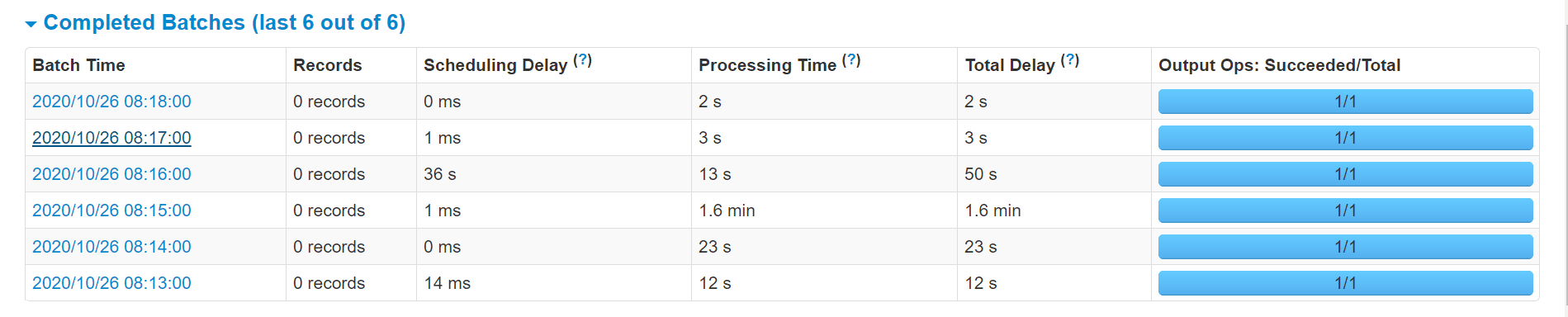




**图表 3.2.5 启动spark streaming，没分钟进行一次监听**

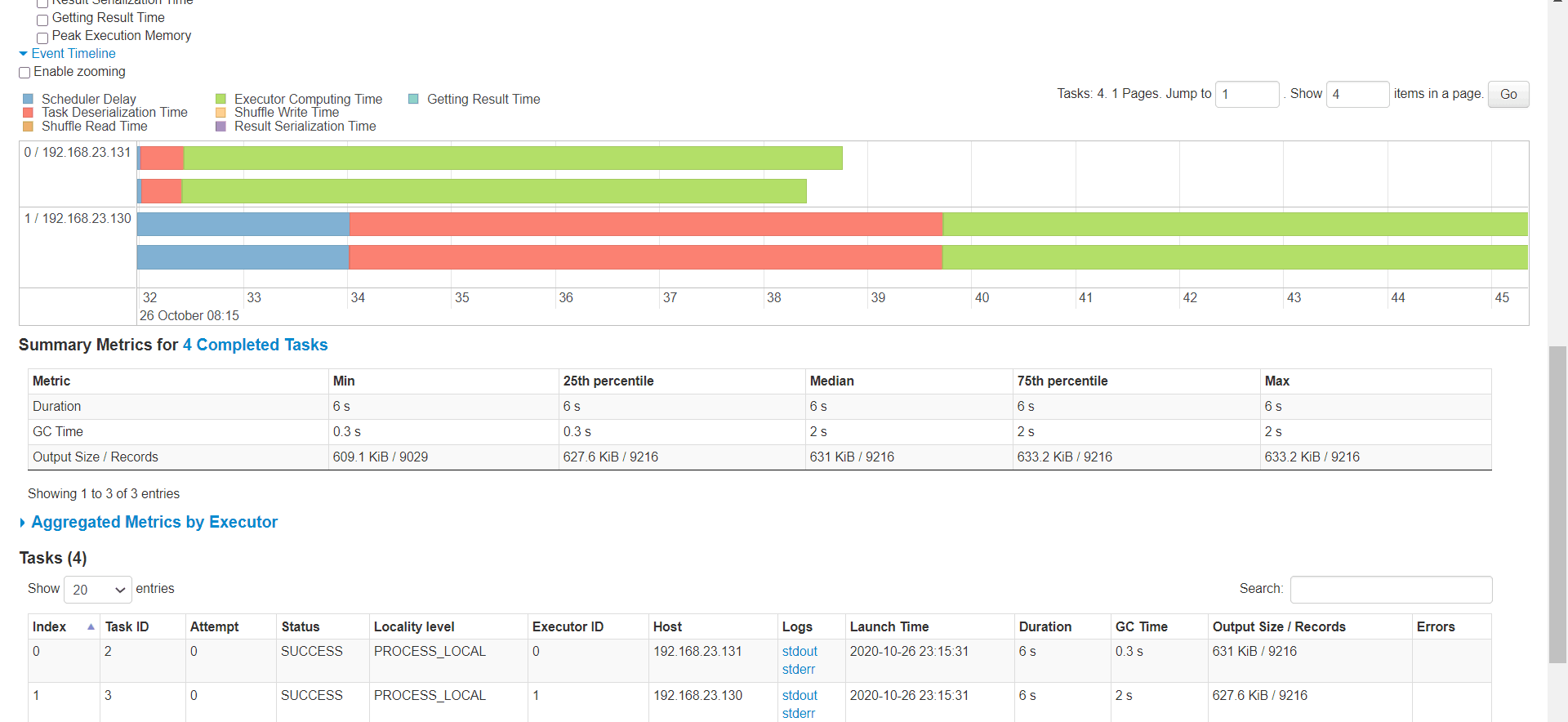
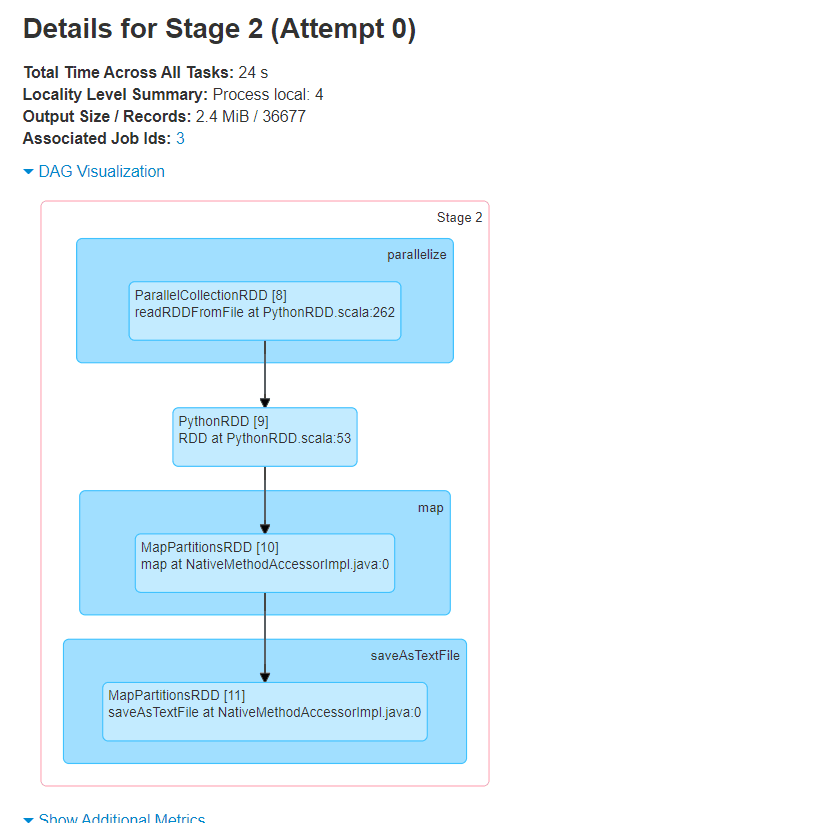


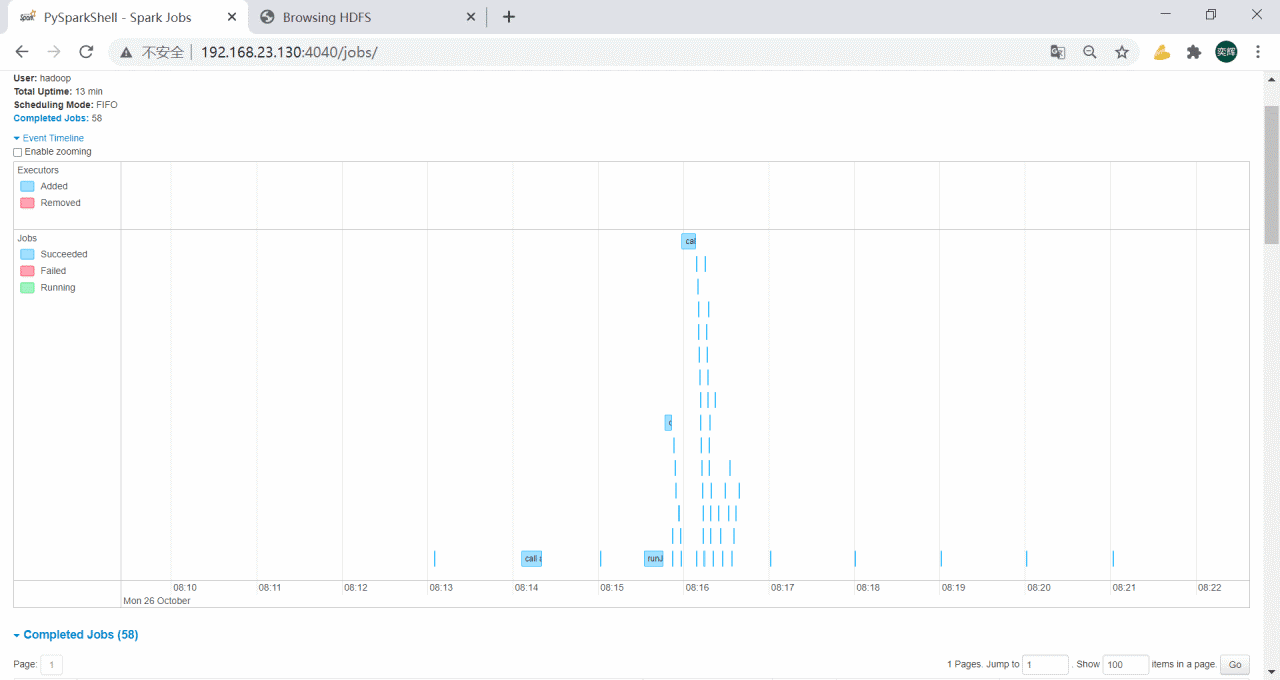




**图表 3.2.6 向监听文件夹写入文件，监听到文件写入开始进行计算**

3 . 流计算



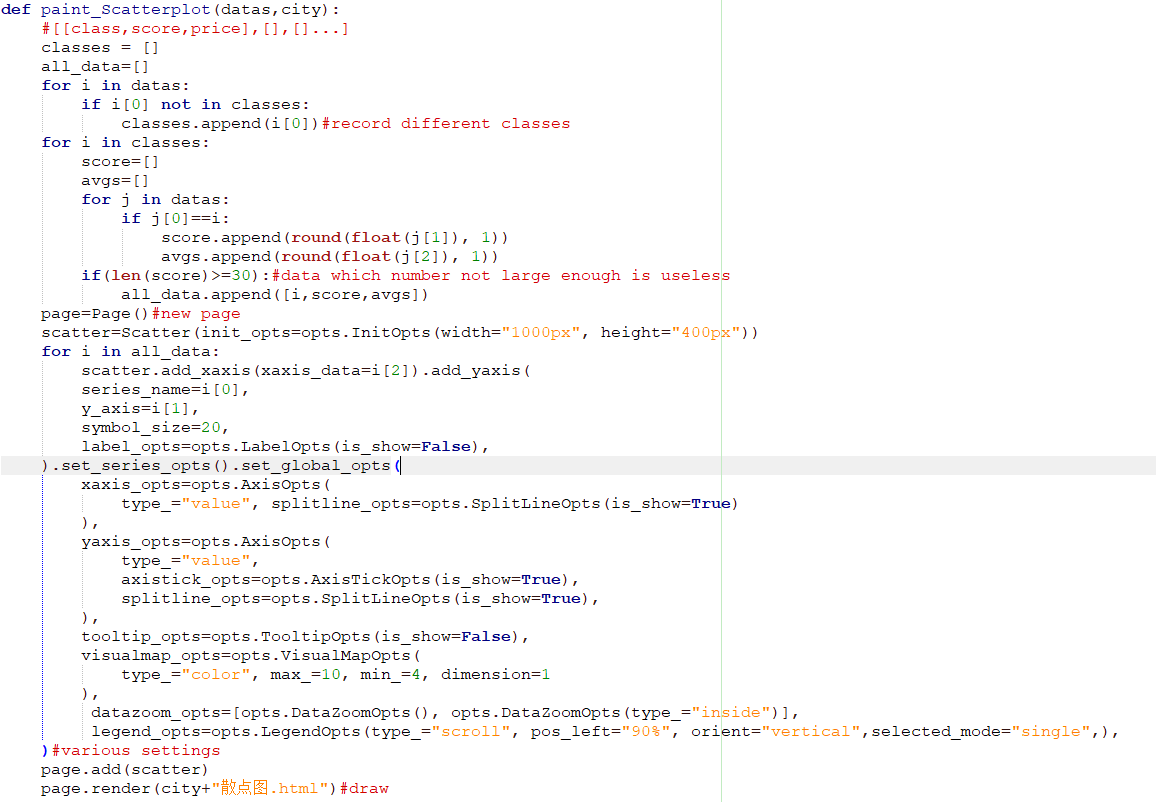
图表 3.3 查看任务运行的具体情况

图表 3.4 查看进行时的任务

4 . 计算结果动态展示

利用pyecharts组件绘制了图表

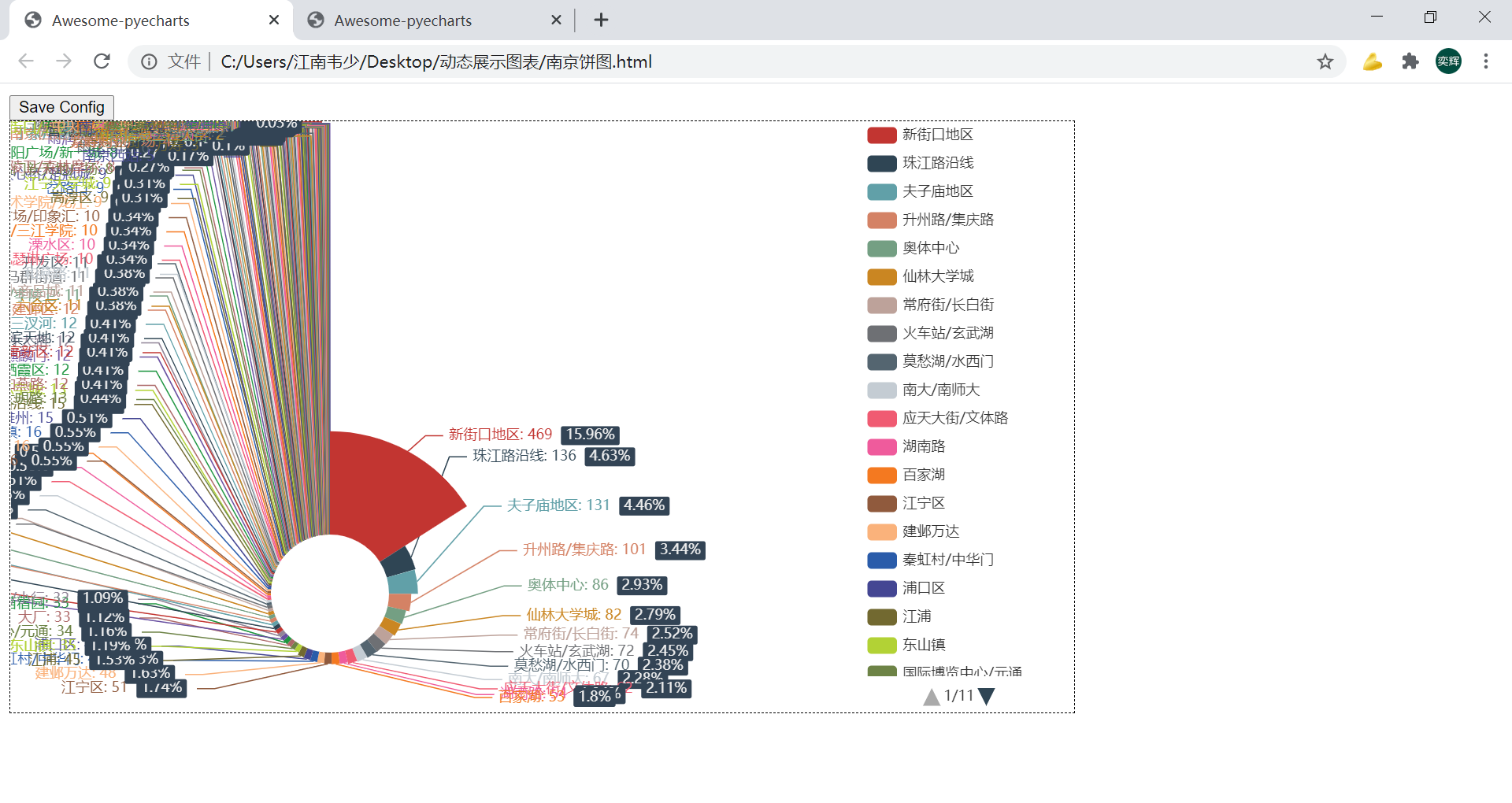




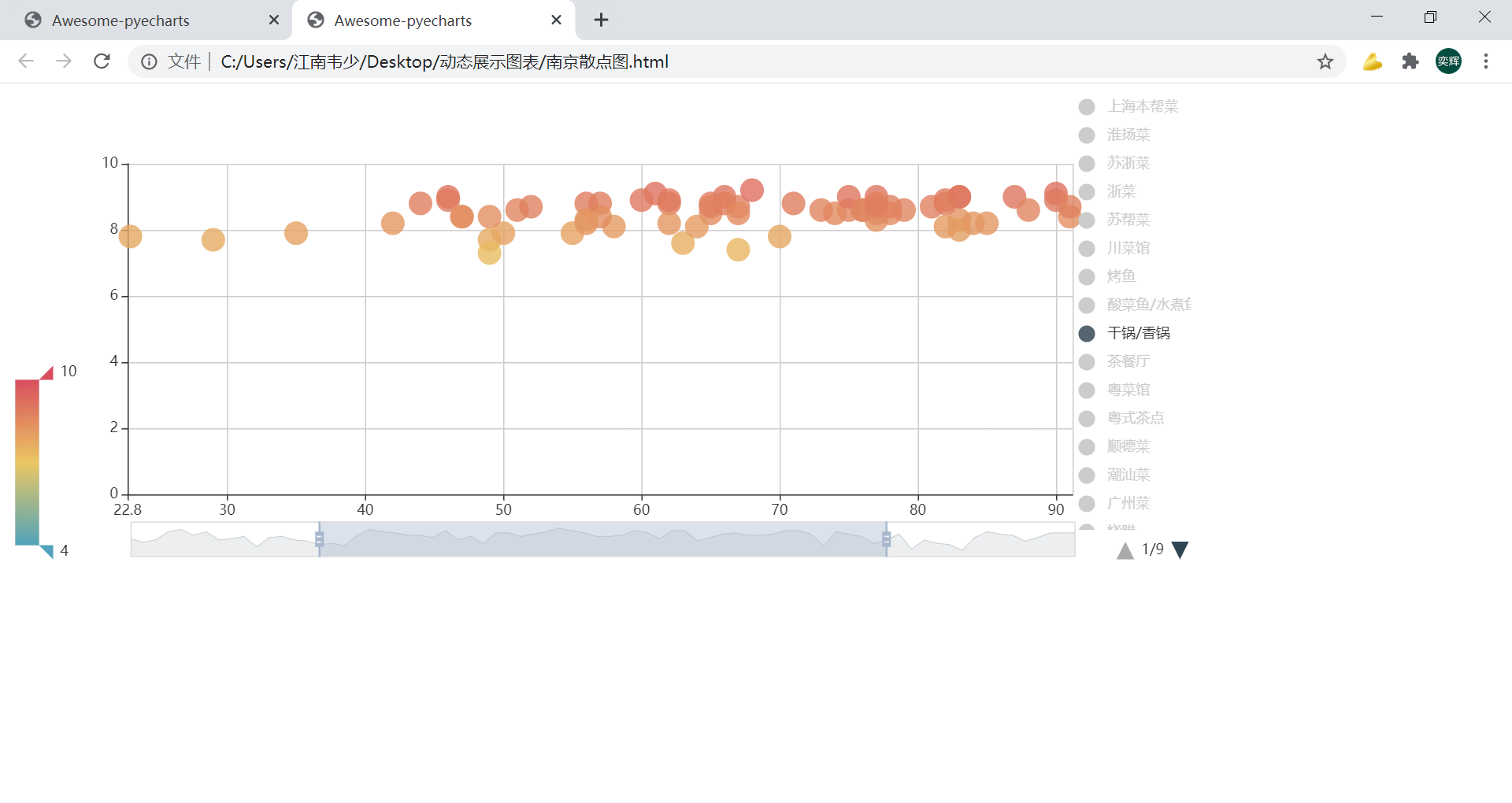
**图表 3.5.1 利用pyecharts绘制动态图表**

结果文件包含在提交文件夹的“动态展示图表下”。由于数据量原因，只绘制了12个 主要城市的相关图表。

其中包括饼图与散点图，分别直观展现出此城市不同商圈上榜的餐厅数和此城市不 同菜系中上榜餐厅的价格情况。以下为南京市相关的饼图和散点图的静态示例，动态图 参见文件夹。



**图表 3.5.2 饼图示例（静态）**

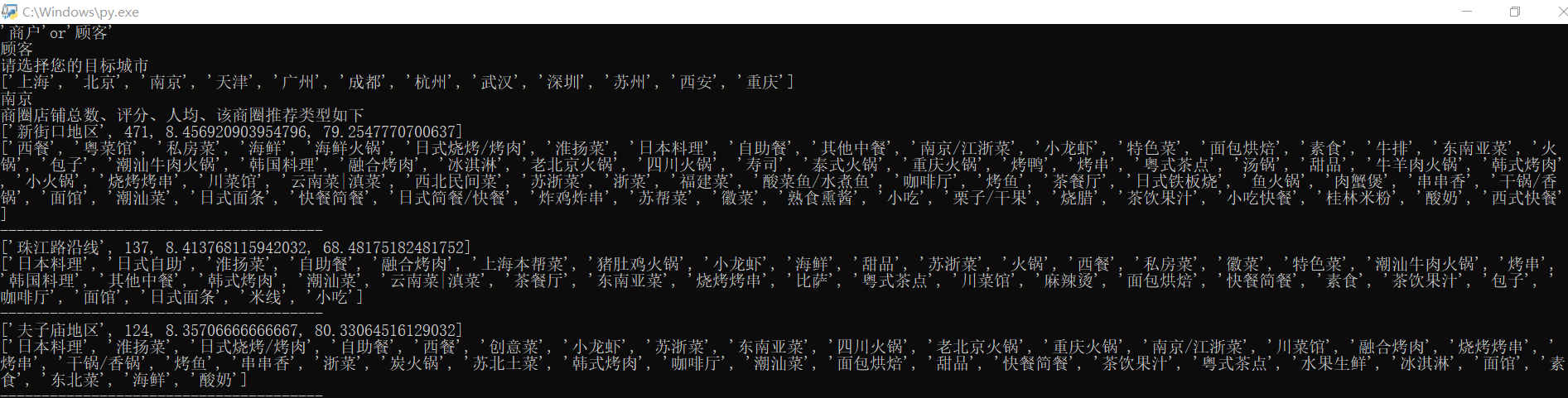


**图表 3.5.3 散点图示例（静态）**

**四、业务需求实现情况**

利用这些数据，我们制作了面向顾客和商家的推荐系统。

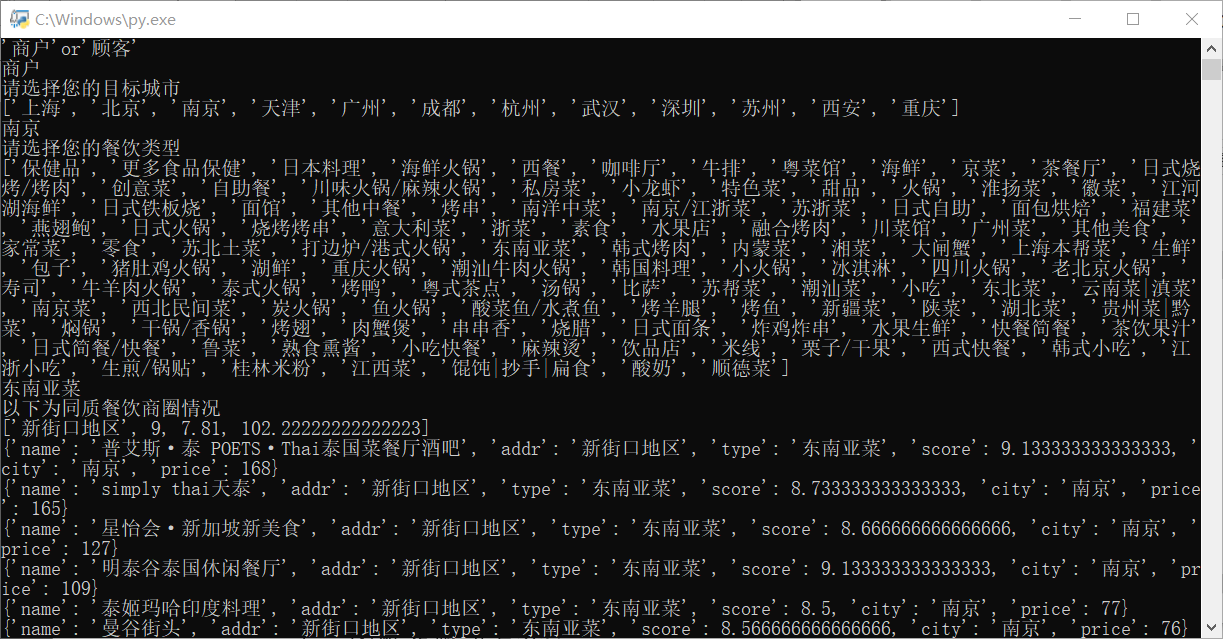
1. 面向顾客给出商圈推荐



图表 4.1.1

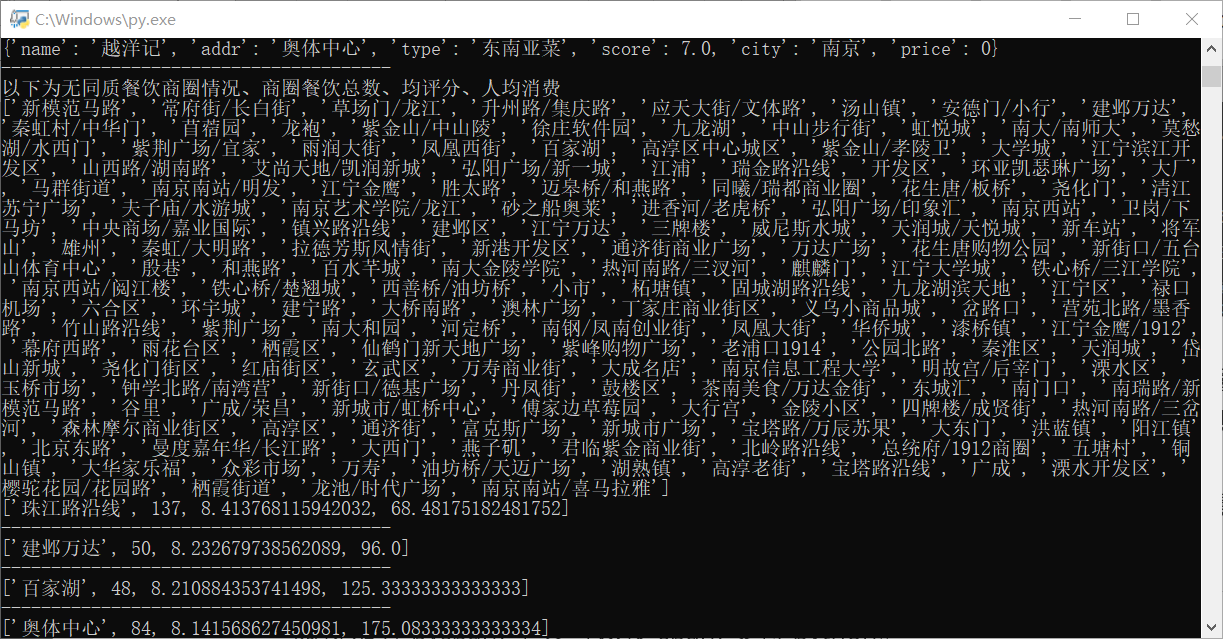
给出了目标城市的评分最高商圈以及该商圈的推荐餐饮类型

1. 面向商户综合给出商铺最优化选址；



图表 4.2.1

首先选择目标城市和目标餐饮类型，然后输出该城市所有同质餐饮的商圈分布情况，以及无同质餐饮的商圈情况，包括人均消费、评价均分、餐饮店总数等，供商户参考。



图表 4.2.2

**五、困难及解决方案**

1．原先计划在阿里云服务器上搭建spark环境，但是由于阿里云服务器的问题，导致要在其上启动spark集群需要使用自己的私网地址，而这个私网地址无法被其他服务器访问，需要将私网地址转为和公网地址一致。但这需要按流量计费开通网络服务，故由于资金不足而放弃，只能在虚拟机上进行实验，导致运行速度较慢，但是在可接受范围之内。

2． 不熟悉scala语法，于是选择了python，调用python的库，并改用pyspark代替spark-shell作为spark运行的窗口。

3．对数据进行可视化时，需要调用pyechart2.0，但是 pyspark默认的python版本为python2，不支持python3。折腾了半天，修改pyspark默认调用的python版本为python3，成功用上pyechart。