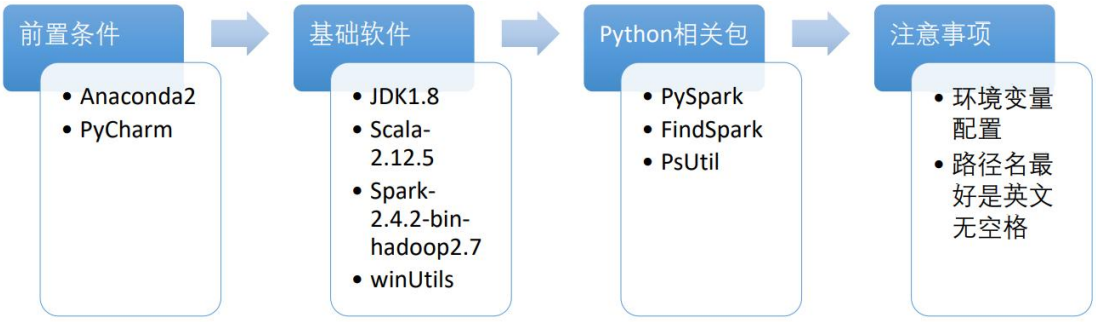
实习日志

5.23 星期一

- **今日任务**：搭建 windows 环境下的 Spark 环境
- **操作流程**：主要是按照老师给的流程进行操作和安装



- **遇到的问题**：在搭建环境的时候遇到了许多问题，在老师的帮助下和自己的探索下最终解决了问题（耗费了一整天的时间，特此记录）

由于最开始创建的 conda 环境是 3.9 的 python 版本，而 jdk、scala、hadoop 都是按照老师给的版本下载，导致出现了版本不兼容的问题，于是我新建了 python 版本为 3.6.13 的虚拟环境 BG，在此尝试。

但是此时我遇到了本次任务最大的难题，在 vscode 中运行测试代码时，遇到了这样的报错【Java not found and JAVA_HOME environment variable is not set.】，最开始我们认为是环境变量的设置问题，于是在检查之后我发现环境变量中没有加入 ClassPath、JRE_HOME 环境变量。但是很遗憾，加入之后仍然出现相同的报错。然后尝试在命令行环境下运行测试代码，惊奇的发现居然能够运行成功。这里我们断定认为是编辑器的问题。于是我尝试下载 pycharm，在 pycharm 中运行测试代码，但是很遗憾仍然出现相同问题。最后在网上绝望的搜寻的时候，看到一篇博客，按照它的说法，我在系统变量的 path 中添加了如下内容（之前仅仅加了 jdk 的）

```
%JAVA_HOME%\jre\bin
```

加入之后再到 vscode 中运行，之前的错误就不见了。但是出现了新的错误：

```
py4j.protocol.Py4JError: org.apache.spark.api.python.PythonUtils.isEncryptionEnabled does not exist in the JVM
```

在代码开头加上如下两句，初始化找到本机的 spark 环境就可以了。

```
import findspark
findspark.init()
```

```
+-------+
|sum(id)|
+-------+
|   45.0|
+-------+


(BG) D:\本科\时空数据处理与组织\实习\code>成功: 已终止 PID 13652 (属于 PID 15940 子进程)的进程。
成功: 已终止 PID 15940 (属于 PID 16192 子进程)的进程。
成功: 已终止 PID 16192 (属于 PID 16220 子进程)的进程。
```

- **感想**：经过此次磨难，我深刻的体会到只要有解决 Bug 的决心，就一定能成功。

当然在解决此类问题的时候要学会控制变量，一步一步的排除。

5.24 星期二

•今日任务：根据给定的啤酒销售数据和去年同期销量数据，把 xlsx 文件转换为 TXT 文件，针对 11 月份啤酒销售数据，并通过编程进行数据处理和计算。
去除整月销量为 0 的数据。

1) 去除整月销量为 0 的数据。
2) 转换数值格式，把销量数据中的引号、逗号等处理掉，并转换为数值。（基于提供的”实习 2 啤酒销量数据.tx”数据文件）
3) 有多少类型的啤酒？
4) 哪 5 种啤酒卖得最好？（销量最高）
5) 在去年销量大于 500 的区域中，哪个销售区域销售的啤酒同比去年增长最快？（按照增长率计算）
6) 统计每种啤酒的 11 月份前 3 周的销量。
7) 统计啤酒卖得最好的前三个区域的 11 月份前 3 周销量。

•操作流程：主要是基于 spark 进行 RDD 的一些基础编程，详细可见实验 2 的操作说明

•遇到的问题：

①RDD 对象的数据格式：在对 RDD 对象不太了解的情况下，有时会把其当当做 DataFrame 类型进行操作进而报错，实际上解决此类问题的方法我认为是--输出中间结果。事实上无论任何一个新的结构方法你面前，都要试着写一行代码输出一行，这样可以快速帮助你熟悉其距离细节。

②代码的整体美观性：最开始不太注意代码美观，导致很多错误出现，还增加了错误的寻找时间。事实上，代码美观带来的好处就是代码结构清晰、逻辑性强，这样给原本以缩进来控制的 python 语言纠正错误带来了很好的便利。

•感悟：

本次实验的最主要收获即为熟悉了 spark 环境以及 RDD 对象的操作。

从最开始对于大数据环境的一无所知，到现在能够使用其解决一些实际问题，这无疑是收获巨大的，也同时告诉我们其实大数据没有想象的那么难操作，只要有学习的决心，实际上其和普通的编程没有什么区别。

5.25 星期三

•今日任务：主要包括三个内容：SparkSQL 的基本操作、RDD 到 DataFrame 的转换、使用 DataFrame 读写 mySQL 数据库。（具体内容见实验文档）

•操作流程：具体的实验步骤和代码详见实习报告 3

•遇到的问题：

①安装 mySQL 的时候配置文件出现了错误，在搜索资料之后，将其中注释删除就可以了

②在安装 mysql 的驱动时，发现仅仅在 spark 的安装目录下存放是不够的，还需要在相应的 conda 环境的 pyspark 的 jars 文件夹下存放，才可运行。

·感悟：

本次实验主要是熟悉了 sparkSQL 的操作以及 RDD 和 DataFrame 的转换等，对于大数据的相关操作有了更加深刻的理解。

个人感觉这种循序渐进的教学模式是必需的也是效果良好的，从最开始的基础操作出发，再到后面去解决一些问题，能够更好的帮助我掌握这些技能。

5.27 星期五

·今日任务：

1. 搭建 Kafka，配置对应的 Spark，并验证是否正确运行。
2. 编写车辆位置数据模拟生成程序。
   从车辆坐标文件中，获取车辆轨迹信息，然后定时将数据发送到指定 Kafka 的消息队列的车辆位置 topic 中。
3. 基于 Spark Structured Streaming，编写车辆轨迹处理程序，实时计算车辆进行速度。
   假设出发是车速为 0，每收到一条对应车辆的坐标信息，就根据收到的坐标点和上一次的坐标点计算之间距离，然后距离除以时间差，作为当前车速。

·遇到的问题：

在进行数据接收实验的时候，我遇到了【py3j.protocol.Py4JJavaError:An error occurred while calling o43.load】的问题，后来经过查询资料得到了解决找到 kafka-clients-0.10.0.1.jar 文件，把这个文件复制到 spark 的 jar 下便运行成功。

·感悟：

通过实验四的实习，我一定程度上熟悉了 Spark 连接 Kafka 的 Structured Streaming 的基本操作，掌握了生产者和消费者的基本使用，并由此对课程中所讲述的有关流计算的概 念有了更加深刻地认识。

5.28 星期六

·今日任务：

1. 从文件中导入数据，并转化为 DataFrame。
2. 训练决策树模型，用于预测居民收入是否超过 50K；
3. 对 Test 数据集进行验证，输出模型的准确率。

·操作流程：具体的实验步骤和代码详见实习报告 5

·遇到的问题：

①数据导入的时候测试集行末有一个‘.’没有发现，导致后面 精度评定的时候出错，通过 strip 函数导入测试集去除即可。

②数据导入时分割符最开始认为是‘,’，但实际上是‘, ’，这会导致后面字符特征数值化的时候出错。

·感悟：

本次实验熟悉了 spark 环境下的 MI 编程实现机器学习的算法流程。最为直观的感受就是机器学习虽然有着比较复杂的运算逻辑，但是经过 MI 库包装之后变成了非常结构化、简洁化、公式化的流程。事实上，只要我们将数据处理成标

准格式，后面我们只需要按部就班地创建决策树模型，进行训练，进行预测，精度评定即可。

5.29 星期天

- **今日任务**：完成车辆轨迹的分析（速度计算、停留点分析、运动状态分析）
- **操作流程**：具体的实验步骤和代码详见综合实习报告
- **遇到的问题**：

本次实验遇到了一些问题，具体如下：

1. 格式问题：由于使用了窗口函数，导致操作对象全部为 column 对象，导致出现了很多问题。

（1）例如在调用计算速度函数时，由于第一行数据没有前一行，如果在函数中仍然调用前一行数据则会报错。解决方法就是在加入判断，对第一行进行特殊处理。

（2）在计算距离函数中使用 sin、cos 等一系列的数学公式时，我开始调用的是 pyspark.sql 的 functions 中的函数，但由于其操作对象是 column 而报错（当时就是说操作对象不为 column）。解决方法是调用 math 中的相关函数。

（3）最开始使用 withColumn 增加列时，会出现“返回类型不匹配”的错误，解决方法是利用 udf 将运算函数限定类型，再在 withColumn 中调用。如下：

```
GET_SPEED=udf(get_speed,FloatType())
```

2. 特殊数值处理：比如在计算速度的时候要设置第一行和最后一行的速度为 0m/s。当时因为 dataframe 格式没有办法单独修改某个值而花费了大量的时间。最后使用 zipWithIndex 函数增加了序号列，根据序号进行判断即可特殊设置。（但是这里的问题和上面类型的问题一起遇到，导致花费了大量时间解决，但也收获颇丰）。

- **感悟**：

通过本次实验，主要是复习巩固了 spark 的基础知识，将所学内容运用到实际当中。

对于本次实习的学习方式我是比较适应的，由浅入深，循序渐进，在一个一个的小练习中掌握基础知识，在最后的大实验中融会贯通。对于 spark 的 RDD 编程，我认为其更加离散化，虽然也主要是逐行操作，但是通过一些自定义的函数也可以实现数据的自由。而 spark 的 DatFrame 则更加结构化严谨化，其主要是行操作甚至窗口操作。至于 sparksql 则是一个实用额的数据增删改查的工具，将 spark 和 sqsl 语句结合从而提供更加多样化的功能。sparkML 主要是关于机器学习领域，其流水线思的操作步骤，大大简化了机器学习编程的难度。

本次实验主要是分了车辆轨迹的速度、停留点、运动状态等。事实上在事件

问题中我们会碰到各种细节，为了使编程模型更加贴近现实，我们必须处理好这些细节，这也有助于我们提高思维能力和编程习惯。尽管在实验中遇到很多困难，但是在老师的耐心帮助下和自己的细心探索之下，完美的完成了实验任务。

本次实验让我初步认识了大数据处理框架，初步熟悉了大数据处理流程，为以后的学习工作奠定了良好的基础。最后感谢老师的教导，致此。