

## 主观题 1：数字技术平台部署（30 分）

本任务需要在 Linux 中完成 Flume 的安裝配置、Kafka 集群和 Flink 集群的搭建，请使用 root 用户完成相关配置，具体要求如下。

### 任务 1. Flume 安裝配置

（1）将 master 节点的/data 目录下的 Flume 安装包解压到/opt/software 目录下（若目录不存在需自行创建/opt/software 目录）。

（2）进入 Flume 安装目录的 conf 目录，将 flume-env.sh.template 重命名为 flume-env.sh，并将/etc/profile 文件中的 Java 安装目录（JAVA\_HOME）添加至 flume-env.sh 文件末尾。

（3）删除 Flume 安装目录的 lib 目录下的 guava-11.0.2.jar 包，之后查看/etc/profile 文件中的 Hadoop 安装目录（HADOOP\_HOME），将 Hadoop 安装目录的/share/hadoop/common/lib/目录中的 guava-27.0-jre.jar 复制至 Flume 安装目录的 lib 目录。

（4）在 master 节点的/etc/profile 文件中配置 Flume 环境变量 FLUME\_HOME 和 PATH 的值，并使配置文件立即生效，之后查看 Flume 版本，检测 Flume 是否安装成功。

### 任务 2. Kafka 集群搭建

（1）将 master 节点的/data 目录下的 Kafka 安装包解压到/opt/software 目录下（若目录不存在需自行创建/opt/software 目录）。

（2）进入 Kafka 安装目录的 config 目录修改 server.properties 配置文件，将“broker.id”改为“0”，“log.dirs”改为“/opt/logs/kafka-logs”，“zookeeper.connect”改为“master:2181,slave1:2181,slave2:2181”。

（3）将 master 节点配置好的 Kafka 文件远程发送至 slave1、slave2 节点相同目录下，并将 slave1、slave2 节点的 server.properties 配置文件中的 broker.id 分别修改为 1、2。

（4）在 master 节点的/etc/profile 文件中配置 Kafka 环境变量 KAFKA\_HOME 和 PATH 的值，并使配置文件立即生效。再将 master 节点配置好的/etc/profile 文件远程发送至 slave1、slave2 节点，同样使配置文件立即生效。

（5）分别在各节点启动 ZooKeeper 集群，确保 ZooKeeper 集群启动后再在各节点启动 Kafka 集群，并查看各节点进程。

### 任务 3. Flink 集群搭建：

（1）在 master 节点将/data 目录下的 flink-1.14.0-bin-scala\_2.12.tgz 安装包解压到/opt/software

目录下（若目录不存在需自行创建/opt/software 目录）。

（2）进入 Flink 安装目录的 conf 目录，修改 workers 文件，注释原文件内容并添加 slave1 和 slave2。

（3）进入 Flink 安装目录的 conf 目录，修改 masters 文件，注释原文件内容并添加 master:8081。

（4）进入 Flink 安装目录的 conf 目录，按照下表修改或添加 flink-conf.yaml 文件中参数（参数值中 master 需替换为/etc/hosts 文件中的具体 IP）。

表 1 flink-conf.yaml 文件参数

参数名称	参数值
jobmanager.rpc.address	master
jobmanager.heap.size	512m
taskmanager.memory.flink.size	1024m
taskmanager.numberOfTaskSlots	3
parallelism.default	1
rest.address	master
rest.bind-address	master
io.tmp.dirs	/opt/logs/flink/tmp

（5）在每个节点上创建日志文件目录/opt/logs/flink/tmp。

（6）将 master 节点配置好的 Flink 文件远程发送至 slave1、slave2 节点相同目录下。

（7）在 master 节点修改/etc/profile 文件，设置 Flink 环境变量 FLINK\_HOME 和 PATH 的值，并使配置文件立即生效。

（8）启动 Flink 集群，并查看各节点的进程。

【说明】

（1）进入环境后需先在 Linux 终端执行命令“initnetwork”，或者双击桌面上名称为“初始化网络”的图标，初始化实训平台网络。

（2）若想切换至 slave1 或 slave2 节点，可以打开新的 Linux 终端窗口，然后输入“ssh slave1”或“ssh slave2”即可切换到对应的节点。

（3）安装包获取需要在 Linux 终端使用 wget 命令获取：

```
“                                wget                                -P                                /data/
http://house.tipdm.com/SZ-Competition/software/apache-flume-1.11.0-bin.tar.gz”
“wget -P /data/ http://house.tipdm.com/SZ-Competition/software/kafka_2.12-2.4.1.tgz”
“                                wget                                -P                                /data/
http://house.tipdm.com/SZ-Competition/software/flink-1.14.0-bin-scala_2.12.tgz”
```

## 主观题 2：大数据技术应用 （40 分）

### 任务 4：数据清洗与存储

财务数据包含大量有用的原始信息，在用于财务分析前需进行一系列的数据预处理工作，确保数据的准确性和一致性。本题的“BalanceSheet.xlsx”是某企业 2019 年至 2022 年的资产负债表，记录了不同报告期资产负债表的财务指标，请依据题目要求运用 Python 语言，对财务分析数据进行数据探索、预处理、整合等操作。

（1）读取“BalanceSheet.xlsx”2019 年至 2022 年工作表内的数据，纵向合并多份数据，并设定“报告期”列为数据框索引（index），调用 index 和 shape 属性（例如，使用 df.index 或 df.shape）查看合并后数据框的基本信息，以确认数据处理的正确性。

（2）基于任务（1）处理后的数据，对数据进行探索（注：探索过程不对数据做更改）。

①数据存在缺失值，统计并输出不同报告期中缺失值的数量。

②若财务指标的数值均缺失，将失去分析价值，统计并输出无分析价值的财务指标的个数。

③通常财务指标的单位为“万元”，但仍有部分单位为“元”，分别统计并输出不同单位的财务指标的个数。

（3）基于任务（1）处理后的数据，对数据进行预处理。

①删除数据中无分析价值的财务指标，而对于其他指标的缺失值，使用数值 0 填充。

②为了保障财务数值单位一致，将财务指标单位为“元”的数值金额换算为“万元”，数值金额保留 2 位小数（四舍五入）。

③鉴于金额单位已完成换算，需剔除财务指标名称中的“#”、“(万元)”和“(元)”等字符，规范化财务指标名称。

④根据“流动比率=流动资产合计/流动负债合计”计算公式，计算不同报告期的流动比率，计算结果保留 2 位小数（四舍五入），并将最终值存储于“流动

比率”列。

⑤对处理完成后的数据存储至“result3.xlsx”结果文件。

#### 【数据获取】

下载题目附件中的数据，上传到实训平台中

#### 【文件读取路径】

“/data/BalanceSheet.xlsx”

### 任务 5：数据分析与可视化

企业盈利能力分析是财务分析环节之一，其分析结果能协助企业投资者将资金投向盈利能力强的企业，而影响盈利能力的因素包括有营业收入、营业成本、营业额等。本题提供了某企业的商品销售表（SaleData.csv，采用 gbk 编码）与财务利润表（IncomeStatement.xlsx，金额单位：万元），请依据题目要求运用 Python 语言，对数据进行统计与图形绘制。

（1）读取商品销售表，根据商品类别字段，统计各商品类别的销售总金额、销售总数量，并完整展示统计结果。

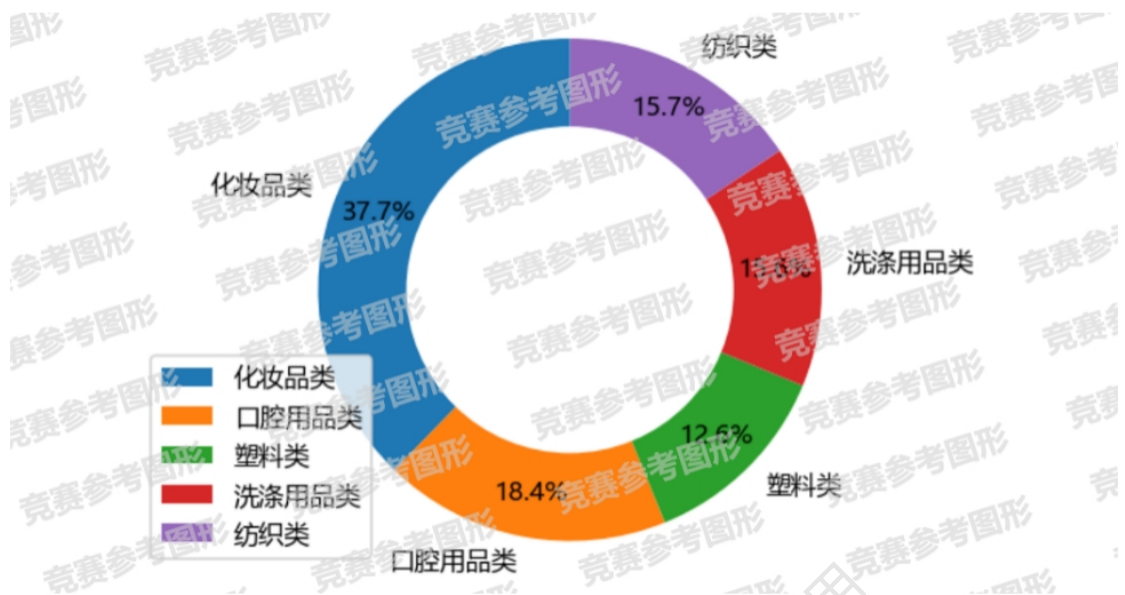
（2）读取财务利润表，根据计算公式，计算“所得税率”、“净利润率”和“毛利率”三种利润趋势分析指标，并完整展示计算结果。

所得税率 = 所得税费用 / 利润总额

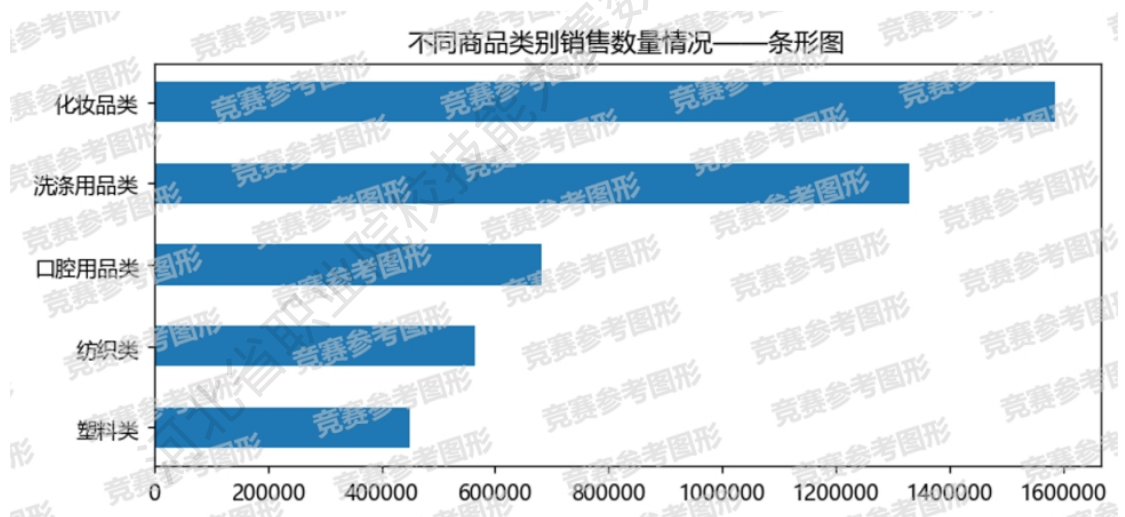
净利润率 = 净利润 / 营业总收入

毛利率 = (营业总收入 - 营业成本) / 营业总收入

（3）使用任务（1）的统计结果，对销售额降序排序，并绘制不同商品类别销售额占比环形图。依据参考图所示，设定环形图的标签、百分比（保留 1 位小数）与图例。



(4) 使用任务(1)的统计结果,对销售数量升序排序,绘制不同商品类别销售数量条形图。依据参考图所示,设定图的标题、轴刻度。



#### 【数据获取】

下载题目附件中的数据,上传到实训平台中

#### 【文件读取路径】

“/data/SaleData.csv”。

“/data/IncomeStatement.xlsx”。

### 主观题 3: 人工智能技术应用 (30 分)



## 任务6 人工智能技术应用

基于 recruitment.xlsx 的招聘数据，按照题目要求使用 Python 完成下列任务。

- (1) 读取 recruitment.xlsx 数据，检查数据是否存在重复值，若存在重复值，统计重复数量并删除重复值。
- (2) 删除“职位描述”列为空的数据，统计表格的行列大小。
- (3) 对“职位描述”列进行清洗特殊符号、分词、去除停用词等操作，将结果作为新列保存到数据中，列名记为“职位描述分词”，展示该列前五行的数据。
- (4) 基于“职位描述分词”构建文档-词频语料库。
- (5) 根据文档-词频语料库构建 LDA 主题模型，设置主题数为 5，迭代次数为 20，并打印出关键词前 6 的主题分布。
- (6) 预测每个职位的主题，将结果作为新列保存到数据中，列名记为“职位主题”。将整个数据保存为 Excel 文件，保存路径设定为“/data/result.xlsx”。
- (7) 评估模型，输出模型的困惑度 (Perplexity) 和一致性 (Coherence)。

### 【数据获取】

下载题目附件中的数据，上传到实训平台中

### 【文件读取路径】

“/data/recruitment.xlsx”