

# 《大数据综合实训》练习04

## • 说明

- 在D盘创建考生文件夹：命名为“《大数据综合实训练习2班》”
- 在IDEA的项目中，创建Spark任务，名称为 `test04`
- 答题结束后，需要提交以下资料
  - 1.需在IDEA中导出 `test04.scala` 文件为 `HTML` 文件，并保存在考生文件夹
  - 2.程序运行结束后，需对整个IDEA的界面进行截图【截图界面包括运行的结果和项目名称】，命名“`idea.jpg`”保存在考生文件夹，截图需要能看到“项目、任务名称+运行结果”
  - 3.将考生文件夹打包提交。

## 一. 启动考试环境

```
# 1.1
能启动VMware和FinalShell连接虚拟机即可得分

# 1.2 启动hadoop集群
start-dfs.sh
start-yarn.sh

# 1.3 启动Hive服务
nohup hive --service metastore &
nohup hive --service hiveserver2 &
```

## 二. 加载数据到HDFS

```
-- 2.1 在HDFS的根目录下创建名为 `exam_data04` 的文件夹

-- 2.2 加载examdata.csv数据文件到刚才创建的 `exam_data04` 的文件夹中
```

## 三. 加载数据到Hive

接收Jar包后，到集群运行加载数据

```
[root@master ~]# spark-submit --master yarn --deploy-mode client \
--class org.example.LoadData Spark2024-1.0.jar
```

## 四. 数据清洗和指标运算

```
package org.example

// 导入必要的Spark SQL库
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._

object test04 {
  // main函数是程序的入口点
  def main(args: Array[String]): Unit = {
    // 初始化Spark会话
    val spark = SparkSession.builder()
      .appName("DWD Data Processor") // 设置应用程序名称
      .master("local[*]") // 设置运行模式为本地模式，使用所有可用的核心
      .enableHiveSupport() // 启用对Hive的支持，允许与Hive交互
      .config("hive.metastore.uris", "thrift://192.168.36.100:9083") // 设置Hive元
      数据存储的Thrift服务地址
      .config("spark.sql.warehouse.dir",
        "hdfs://192.168.36.100:9000/user/hive/warehouse") // 设置Hive仓库在HDFS上的位置
      .getOrCreate() // 创建Spark会话，如果存在则获取现有的

    // 调用自定义的cleanDataAndStoreToDWD函数，处理并存储数据到DWD层
    cleanData(spark)

    // 执行完毕后关闭Spark会话
    // spark.stop()
  }

  // 定义一个处理数据并将其存储到DWD层的函数
  def cleanData(spark: SparkSession): Unit = {
    // 从ODS层读取用户和订单数据
    val odsSalesDF = spark.table("ods.ods_sales")

    // 对用户数据进行清洗：删除空值、重复值
    val cleanedSalesDF = odsSalesDF
      .na.drop() // 删除所有列中的空值
      .dropDuplicates() // 删除重复值
    println("完成数据进行清洗：删除空值、重复值.....")

    // 计算销售总额（总金额 = 单价 * 销售数量 * (1 - 折扣)），并添加为新列total_sales
    val dwdSalesDF = cleanedSalesDF.withColumn(
      "total_sales", col("price") * col("quantity") * (lit(1) - col("discount"))
    )

    println("dwdSalesDF数据表第1行的内容")
    dwdSalesDF.show(1)

    println("程序运行结果：")
    // 1. 计算所有订单的平均销售额
    val avgTotalSales =
      dwdSalesDF.agg(avg("total_sales").alias("average_total_sales")).first().getDouble(
        0)
  }
}
```

```
println(f"1.所有订单的平均销售额: $avgTotalSales%.2f 元")

// 2.计算2024-04月份的销售总额
val marchSalesDF = dwdSalesDF.filter(col("sale_date").startsWith("2024-04"))
val marchTotalSalesDF =
marchSalesDF.agg(sum("total_sales").alias("total_sales_amount"))
val marchTotalSalesAmount = marchTotalSalesDF.first().getDouble(0)
println(s"2.2024-04月份的销售总额: $marchTotalSalesAmount")

// 3. 计算 "价格大于 10000 元" 的商品数量
val expensiveProductsCount = dwdSalesDF.filter(col("price") > 10000).count()
println(s"3.价格大于 10000 元的商品数量: $expensiveProductsCount")

// 4.查询 "折扣小于10%" 的订单数量
val discountThreshold = 0.10
val highDiscountOrderCount = dwdSalesDF.filter(col("discount") <
discountThreshold).count()
println(s"4.折扣大于 10% 的订单数量: $highDiscountOrderCount")

}
}
```