任务3-1 DataFrame和Dataset的创建与操作

|  |  |
| --- | --- |
| 任务书 | |
| 一、任务描述与要求 | |
| 任务描述 | 将需要处理的数据从文件（如CSV、JSON）或数据库等数据源中载入到程序中，并检查数据的格式和规模等信息。 |
| 具体要求 | 1.使用SparkSession从不同的数据源（如文件、数据库等）创建DataFrame和Dataset；  2.利用Spark SQL对DataFrame进行基本的数据筛选、排序和转换操作。 |
| 二、任务目标 | |
| 知识目标 | * 了解 DataFrame 和 Dataset 的结构特点； * 掌握不同方式创建 DataFrame 和 Dataset。 |
| 能力目标 | * 能够熟练运用多种方法创建 DataFrame 和 Dataset； * 具备使用 DataFrame 和 Dataset 进行高效数据操作的能力。 |
| 三、知识储备 | |
| 1.Spark SQL 简介  Spark SQL 是 Apache Spark 中用于处理结构化数据的模块。它为执行SQL语句和混合了SQL和复杂分析任务的编程工作流提供了支持。Spark SQL 提供了一种统一的接口来处理各种数据源，并在内部利用 Spark 的强大分布式计算能力来优化查询性能。    2. DataFrame简介  DataFrame 是 Spark SQL 中一个重要的分布式数据集合，与传统的关系数据库中的表格类似。它提供了对结构化和半结构化数据的高级抽象，并且利用 Catalyst 查询优化器和 Tungsten 执行引擎实现高效计算。  Spark SQL 中的 DataFrame 是处理结构化数据的强大工具，结合 Catalyst 查询优化器和 Tungsten 执行引擎，提供了高效、灵活的数据操作能力。它简洁易用、适应性强，适合于各种数据分析和处理任务。 | |

|  |
| --- |
| 操作指南 |
| 一、制定计划 |
| |  |  |  |  | | --- | --- | --- | --- | | **序号** | **作业项目** | **序号** | **作业项目** | | 1 |  | 5 |  | | 2 |  | 6 |  | | 3 |  | 7 |  | | 4 |  | 8 |  | | 计划审核 | 审核意见： | | | |
| 二、实施方案 |
| 1. 下载并配置Spark环境  Spark是一个开源的，基于内存计算的大数据处理引擎。在开始使用Spark SQL之前，需要先配置好Spark环境。首先需要进入官方网站下载Spark，并按照相应步骤进行配置。安装完成后，引入必要的Spark库：  from pyspark.sql import SparkSession  # 创建SparkSession  spark = SparkSession.builder.appName("Spark SQL Example").getOrCreate()  2. 获取数据并创建DataFrame  可以通过读取各种数据源（如CSV、JSON、Parquet）来获取数据，并创建DataFrame进行处理。以下示例代码展示了如何从CSV文件加载数据：  # 读取CSV文件数据  df = spark.read.csv("path/to/data.csv", header=True, inferSchema=True)  # 打印数据类型和前5行数据  df.printSchema()  df.show(5)  3.使用Spark SQL进行数据处理  使用Spark SQL可以轻松地对数据进行各种操作，例如筛选、排序、统计等，具体代码如下：  # 注册临时视图  df.createOrReplaceTempView("table")  # 数据查询示例  result = spark.sql("SELECT \* FROM table WHERE columnA > 50 ORDER BY columnB ASC")  # 显示查询结果  result.show()  4.结果导出与可视化  为了进一步分析和展示数据，可以将处理后的数据导出为不同格式，供其他工具使用。具体代码如下：  # 导出数据为CSV格式  result.write.csv("path/to/result.csv", header=True)  最终可将查找数据导出成csv文件，具体如图所示：  图片5.8 |
| 三、实施记录 |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **序号** | **关键步骤记录** |  | **序号** | **解决问题记录** | | 1 |  |  | 1 |  | | 2 |  |  | 2 |  | | 3 |  |  | 3 |  | | 4 |  |  | 4 |  | | 5 |  |  | 5 |  | | 6 |  |  | 6 |  | | 7 |  |  | 7 |  | | 8 |  |  | 8 |  | | 9 |  |  | 9 |  | |
| 四、任务拓展 |
| 该任务要求从两个不同的数据源读取数据，进行数据清洗与转换操作，然后对这些数据进行聚合和连接操作，最后将结果写入到 Parquet 文件中。具体操作步骤如下：   1. 数据读取：读取订单数据和产品数据。订单数据存储在 CSV 文件中，包含各个订单的详细信息。产品数据存储在 JSON 文件中，包含产品的详细信息。 2. 数据清理和转换：对订单数据进行清洗，去除订单 ID 和金额为空的无效记录，确保数据的完整性。 3. 数据聚合：以订单为单位进行聚合计算，计算每个订单的总金额。具体来说，是对每个订单的各个商品金额求和，得到每个订单的总金额。 4. 数据连接：将聚合后的订单数据与产品数据进行连接操作。连接键为 product\_id，即通过 product\_id 把订单数据中的商品和产品数据中的商品信息关联起来，获得每个订单中各个商品的详细信息。 5. 保存结果：将最终连接得到的数据结果保存为 Parquet 文件，路径为 path/to/output.parquet，以便进一步使用和分析。 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评价标准 | | | | | | | | |
| 一、基本信息 | | | | | | | | |
| 项目名称 | | 单元三 使用Spark SQL处理数据 | | | | | | |
| 任务名称 | | 任务3-1 DataFrame和Dataset的创建与操作 | | | | | | |
| 班级 |  | | 学号 | |  | 姓名 | |  |
| 组名 |  | | 学时 | |  | 日期 | |  |
| 组员  分工 |  | | |  | | |  | |
|  | | |  | | |  | |
| 二、任务检查评价单 | | | | | | | | |
| |  |  |  | | --- | --- | --- | | **评价内容** | **评价标准** | **得分** | | 知识运用  （20分） | 掌握相关理论知识，理解本次任务要求，制定详细计划，计划条理清晰，逻辑正确（20分） |  | | 理解相关理论知识，能根据本次任务要求、制定合理计划（15分） | | 了解相关理论知识，有制定计划（10分） | | 无制定计划（0分） | | 专业技能  （40分） | 结果验证全部满足。（40分） |  | | 结果验证只有一个功能不能实现，其它功能全部实现（30分） | | 结果验证只有一个功能实现，其它功能全部没有实现（20分） | | 结果验证功能均未实现（0分） | | 核心素养  （20分） | 具有良好的自主学习能力、分析解决问题的能力、整个任务过程中有指导他人（20分） |  | | 具有较好的学习能力和分析解决问题的能力，任务过程中无指导他人（15分） | | 能够主动学习并收集信息，有请教他人进行解决问题的能力（10分） | | 不主动学习（0分） | | 课堂纪律  （20分） | 设备无损坏、设备摆放整齐、工位区域内保持整洁、无干扰课堂秩序（20分） |  | | 设备无损坏、无干扰课堂秩序（15分） | | 无干扰课堂秩序（10分） | | 干扰课堂秩序（0分） | | 得分（满分100） | |  | | 综合评价 | ☆☆☆☆☆ | | | | | | | | | | |