任务9-3 数据湖分析与处理

|  |  |
| --- | --- |
| 任务书 | |
| 一、任务描述与要求 | |
| 任务描述 | Catalyst是Spark SQL中的查询优化器，利用基于规则和成本的优化框架来提高查询执行效率。了解Catalyst的工作流程，包括解析、生成和优化逻辑计划及物理计划的不同阶段。 |
| 具体要求 | 1.了解Catalyst优化器的工作原理；  2.实现谓词下推和投影下推。 |
| 二、任务目标 | |
| 知识目标 | * 领会数据湖的存储架构要点； * 了解数据湖分析的常用工具。 |
| 能力目标 | * 可以通过工具进行数据湖分析操作； * 拥有提升数据湖处理流程效率的能力。 |
| 三、知识储备 | |
| 1. **数据湖的概念和架构设计**   数据湖是一种用于存储和管理大规模数据的体系结构，其设计目的是容纳各种类型的数据，包括结构化数据（如关系数据库表）、半结构化数据（如JSON、XML）和非结构化数据（如文本、视频、图片）。数据湖的核心特征在于，它允许以原始格式存储数据，并不对数据进行预处理或转换，这使得用户可以在需要时灵活地对数据进行分析和处理。   1. **数据质量控制和数据治理**   数据质量控制专注于确保数据在存储和使用过程中具备高质量标准。这涉及多个方面：  数据验证：数据验证的过程包括检查数据是否符合设定的规则和标准。比如，确保数据格式正确（如日期格式应为YYYY-MM-DD），数据值在合理范围内（例如年龄不能为负数），以及数据的一致性（例如相同字段在不同记录中应保持一致）。  数据清洗：数据清洗是指对数据进行修正和优化的过程。它包括识别并处理数据中的错误、重复和不一致项。数据清洗可能涉及到填补缺失值、标准化数据格式（如将所有地址转化为统一格式）以及删除重复记录等操作。  数据监控：数据监控的目的是持续跟踪数据质量，及时发现并解决潜在的问题。这通常通过数据质量仪表板和报告工具来实现，这些工具提供了数据质量指标和趋势分析，帮助管理人员识别数据质量问题的源头和影响。 | |

|  |
| --- |
| 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. 数据生成  在现代企业中，数据湖（Data Lake）是存储和处理大规模数据的关键基础设施。假设有一个电子商务公司，名为“ShopMaster”，它从不同的渠道（如网站、移动应用和社交媒体）收集了大量用户行为数据、交易数据和产品数据。为了更好地分析这些数据以获取商业洞察，ShopMaster决定使用数据湖来存储原始数据，并利用Apache Spark进行数据处理和分析。为了进行分析，首先生成一些模拟数据。这些数据将被存储在Amazon S3中，数据格式包括CSV和JSON。用户行为数据：   |  | | --- | | user\_behaviors.csv  user\_id,timestamp,action  1,2024-08-01 10:00:00,view  2,2024-08-01 10:05:00,click  1,2024-08-01 10:10:00,view  3,2024-08-01 10:15:00,view  2,2024-08-01 10:20:00,click  4,2024-08-01 10:25:00,view  1,2024-08-01 10:30:00,click  3,2024-08-01 10:35:00,view  2,2024-08-01 10:40:00,view  4,2024-08-01 10:45:00,click |   交易数据：   |  | | --- | | order\_id,user\_id,product\_id,quantity,amount  1001,1,101,2,29.98  1002,2,102,1,19.99  1003,1,103,3,59.97  1004,4,104,1,39.99  1005,2,101,1,14.99 |   产品数据：   |  | | --- | | products.json  [  {"product\_id": 101, "name": "Widget A", "category": "Widgets", "price": 14.99},  {"product\_id": 102, "name": "Widget B", "category": "Widgets", "price": 19.99},  {"product\_id": 103, "name": "Gadget A", "category": "Gadgets", "price": 19.99},  {"product\_id": 104, "name": "Gadget B", "category": "Gadgets", "price": 39.99}  ] |   **2.数据加载**  接下来进行数据加载。具体代码如下：   |  | | --- | | from pyspark.sql import SparkSession  # 创建SparkSession  spark = SparkSession.builder \  .appName("ShopMasterDataProcessing") \  .getOrCreate()  # 读取数据  user\_behaviors\_df = spark.read.csv("s3://shopmaster-datalake/user\_behaviors.csv", header=True, inferSchema=True)  transactions\_df = spark.read.csv("s3://shopmaster-datalake/transactions.csv", header=True, inferSchema=True)  products\_df = spark.read.json("s3://shopmaster-datalake/products.json") |   3.创数据预处理  数据加载后，进行基本的数据预处理。具体代码如下：   |  | | --- | | # 数据清理和转换  from pyspark.sql.functions import to\_date  user\_behaviors\_df = user\_behaviors\_df.dropna() # 删除缺失值  transactions\_df = transactions\_df.withColumn("amount", transactions\_df["amount"].cast("float")) # 转换金额列类型  user\_behaviors\_df = user\_behaviors\_df.withColumn("date", to\_date(user\_behaviors\_df["timestamp"], "yyyy-MM-dd HH:mm:ss")) |   4.数据分析  进行两个主要的分析。分析行为具体代码如下：   |  | | --- | | user\_activity\_df = user\_behaviors\_df.groupBy("date").count().withColumnRenamed("count", "active\_users")  user\_activity\_df.show() |   5.分析产品销售情况  分析产品具体代码如下：   |  | | --- | | product\_sales\_df = transactions\_df.groupBy("product\_id").sum("amount").withColumnRenamed("sum(amount)", "total\_sales")  product\_sales\_df = product\_sales\_df.join(products\_df, "product\_id")  product\_sales\_df.show() | |
| 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 |
| “ShopMaster”是一家电子商务公司，通过数据湖（Data Lake）存储用户行为、交易和产品数据。为了提高业务洞察和决策支持，公司需要通过Apache Spark处理这些数据。该任务包括数据生成、加载、预处理、分析以及结果存储。请尝试完成以下任务拓展：  生成和存储数据：创建模拟的用户行为数据、交易数据和产品数据，并将这些数据存储在Amazon S3数据湖中。  数据加载和预处理：使用Apache Spark从数据湖中加载数据，并进行必要的数据清理和转换。  数据分析：对用户活跃度和产品销售情况进行分析，生成相关业务报告。  结果存储与共享：将分析结果存储回数据湖，以便于后续使用和决策支持。 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评价标准 | | | | | | | | |
| 一、基本信息 | | | | | | | | |
| 项目名称 | | 单元九 Spark项目实战 | | | | | | |
| 任务名称 | | 任务9-3 数据湖分析与处理 | | | | | | |
| 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 | ☆☆☆☆☆ | | | | | | | | | | |