任务6-2 广播变量和累加器

|  |  |
| --- | --- |
| 任务书 | |
| 一、任务描述与要求 | |
| 任务描述 | 深入学习广播变量的定义、用途和工作原理。理解为什么在分布式计算环境中广播变量是重要的优化技术。 |
| 具体要求 | 1.了解广播变量在分布式计算中的作用；  2.正确创建和使用广播变量。 |
| 二、任务目标 | |
| 知识目标 | * 理解播音变量与累加量的概念； * 掌握播音变量和累加量的计算方法。 |
| 能力目标 | * 能够准确识别播音变量和累加量； * 运用播音变量和累加量解决实际问题。 |
| 三、知识储备 | |
| 1.广播变量的原理和使用场景  广播变量是Spark中一种优化大型只读数据分发的重要机制。其核心原理是将数据从Driver节点高效地分发到所有Executor节点，并在每个节点的共享内存中存储单一副本。这种方法显著减少了数据传输和存储开销，因为数据只需传输一次，而不是为每个任务单独传输。  在实现上，广播变量在传输前会被序列化，Spark会选择高效的序列化格式以减少网络传输负担。对于大型数据，还会采用分块传输的策略，提高传输效率和容错性。在Executor端，数据会被反序列化并存储在共享内存中，供该节点上的所有任务访问。     1. **累加器的类型和使用方法**   累加器是Spark提供的一种特殊变量，用于在分布式计算环境中进行安全的计数或求和操作。它的主要优势在于能够在并行任务中进行更新，而不会出现竞争条件。Spark内置了几种常用的累加器类型，包括LongAccumulator用于整数累加、DoubleAccumulator用于浮点数累加，以及CollectionAccumulator用于收集元素到列表中。除此之外，开发者还可以通过继承AccumulatorV2类来创建自定义的累加器，以满足特定的需求。 | |

|  |
| --- |
| 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. 生成模拟数据集  在Spark中，广播变量和累加器是两种常用的分布式计算工具。广播变量用于将只读变量广播到所有工作节点，以减少数据传输开销。累加器用于在分布式环境中对数据进行累加统计。本案例将通过生成一个简单的数据集，展示如何在Spark中使用广播变量和累加器。将生成一个简单的模拟数据集，包含用户ID和他们的购买金额。数据集将使用Scala编程语言创建。具体代码如下：   |  | | --- | | import org.apache.spark.sql.SparkSession  import org.apache.spark.sql.functions.\_  // 初始化SparkSession  val spark = SparkSession.builder()  .appName("Spark Broadcast and Accumulator Example")  .master("local[\*]")  .getOrCreate()  // 生成数据  val data = Seq(  (1, "user1", 100.0),  (2, "user2", 150.0),  (1, "user1", 200.0),  (3, "user3", 250.0),  (2, "user2", 300.0),  (4, "user4", 350.0),  (1, "user1", 400.0),  (3, "user3", 450.0),  (5, "user5", 500.0)  )  // 创建DataFrame  val df = spark.createDataFrame(data).toDF("transaction\_id", "user\_id", "amount")  df.show() |   2. 使用广播变量  广播变量用于将一个只读的变量（例如配置数据、查找表等）广播到所有节点上。这样可以避免在每个任务中重复发送相同的数据，减少网络开销。有一个包含优惠折扣的映射表，需要将其广播到所有节点。完整代码如下：   |  | | --- | | // 创建一个优惠折扣的映射表  val discountMap = Map("user1" -> 0.1, "user2" -> 0.15, "user3" -> 0.2)  // 广播映射表  val broadcastDiscountMap = spark.sparkContext.broadcast(discountMap)  // 使用广播变量来计算折扣后的总金额  val discountedDF = df.map(row => {  val userId = row.getString(1)  val amount = row.getDouble(2)  val discount = broadcastDiscountMap.value.getOrElse(userId, 0.0)  (userId, amount \* (1 - discount))  }).toDF("user\_id", "discounted\_amount")  discountedDF.show() |   3.使用累加器  累加器用于在分布式计算中累加数值，可以用于计算总和、计数等。累加器可以由多个任务并行更新，并在计算完成后获取最终值。创建一个累加器来统计所有交易的总金额。完整代码如下：   |  | | --- | | import org.apache.spark.util.LongAccumulator  // 创建一个累加器  val totalAmountAccumulator = spark  .sparkContext.longAccumulator("TotalAmountAccumulator")  // 使用累加器统计总金额  df.foreach(row => {  val amount = row.getDouble(2)  totalAmountAccumulator.add(amount.toLong)  })  // 获取累加器的值  println(s"总金额: ${totalAmountAccumulator.value}") | |
| 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 |
| 广播一个复杂的数据结构，例如嵌套的Map或自定义对象，研究如何有效地广播和使用这些结构。探索在广播变量需要动态更新时的策略，比如根据计算进度更新广播变量，并评估对性能的影响。 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评价标准 | | | | | | | | |
| 一、基本信息 | | | | | | | | |
| 项目名称 | | 单元六 高级数据处理 | | | | | | |
| 任务名称 | | 任务6-2 广播变量和累加器 | | | | | | |
| 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 | ☆☆☆☆☆ | | | | | | | | | | |