任务9-1 实时日志处理系统

|  |  |
| --- | --- |
| 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 | * 熟悉实时日志处理系统的基本架构内涵； * 知晓常见的日志处理工具及相关技术要点。 |
| 能力目标 | * 可以独立设计并搭建简易实时日志处理系统； * 拥有对日志数据进行深入分析及故障排查的能力。 |
| 三、知识储备 | |
| 1. **日志数据的实时采集工具**   在 Spark 集群中，实时采集日志数据对监控和故障排查非常重要。常用的实时采集工具有 ELK Stack、Fluentd、Prometheus 和 Grafana、Apache Kafka、以及 Splunk。  ELK Stack（Elasticsearch, Logstash, Kibana）是一套强大的日志处理工具。Fluentd 是一个灵活的数据收集器，可以实时收集和转发日志数据。Prometheus 是一个开源监控系统，主要用于指标的收集和报警。Apache Kafka 是一个高吞吐量的消息队列系统，可以实时处理日志数据流。Splunk 是一个全面的日志管理和分析平台，支持实时数据采集、搜索和可视化。  这些工具和平台能够帮助你高效地采集、存储和分析 Spark 日志数据，提升系统的监控能力和故障排查效率。    **2.横向扩展技巧**  在日志系统中，实现拓展和横向扩展可以显著提高处理能力和可靠性。首先，采用分布式架构是关键，这意味着将日志系统的各个组件（如收集、存储和分析）分布到多个节点上，从而避免单点过载。日志收集工具如 Apache Kafka 和 Fluentd 能够将日志从多个源集中到一起，而存储系统如 Elasticsearch 和 HDFS 则支持分布式存储。分析工具如 Apache Spark 可以处理大规模的数据集，确保系统能够处理大量日志数据。 | |

|  |
| --- |
| 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. 日志生成和收集  假设一个公司需要实时处理和分析来自其 web 应用的日志数据，以便实时监控用户活动、检测异常行为和生成报告。公司决定使用 Apache Spark 作为实时处理平台，并将日志数据存储在 Amazon S3 中。处理的代码通过Python编写，通过PySpark完成系统。日志数据包含用户的访问记录、请求时间、请求 URL、响应时间等信息。假设 web 应用服务器生成的日志数据格式如下：   |  | | --- | | 2024-08-25 09:00:00, user123, /homepage, 200, 123ms  2024-08-25 09:01:00, user456, /login, 500, 456ms |   每条日志包含时间戳、用户 ID、请求 URL、响应状态码和响应时间。日志被发送到 Apache Kafka 的 logs 主题中。  **2. 设置 Apache Kafka**  启动 Kafka 集群，并创建一个名为 logs 的主题。配置 Kafka 生产者将日志数据发送到该主题。将日志数据按行输入到 Kafka 中。具体代码如下：   |  | | --- | | kafka-console-producer.sh --broker-list localhost:9092 --topic logs |   3.配置 Apache Spark  使启动 Apache Spark 集群，并使用 Spark Streaming 处理 Kafka 中的日志数据。具体代码如下：   |  | | --- | | from pyspark.sql import SparkSession  from pyspark.sql.functions import col, split  # 初始化 SparkSession  spark = SparkSession.builder \  .appName("RealTimeLogProcessing") \  .getOrCreate()  # 创建 Kafka 读流  logs\_df = spark.readStream \  .format("kafka") \  .option("kafka.bootstrap.servers", "localhost:9092") \  .option("subscribe", "logs") \  .load()  # 解析日志数据  logs\_df = logs\_df.selectExpr("CAST(value AS STRING)")  # 分割日志字段  logs\_df = logs\_df.select(  split(col("value"), ",").getItem(0).alias("timestamp"),  split(col("value"), ",").getItem(1).alias("user\_id"),  split(col("value"), ",").getItem(2).alias("request\_url"),  split(col("value"), ",").getItem(3).alias("status\_code"),  split(col("value"), ",").getItem(4).alias("response\_time")  )  # 数据处理示例：统计每个状态码的出现次数  status\_code\_count = logs\_df.groupBy("status\_code").count()  # 输出到控制台  query = status\_code\_count.writeStream \  .outputMode("complete") \  .format("console") \  .start()  query.awaitTermination() |   4.数据存储和展示  将处理后的数据存储到 Amazon S3 以便长期存储。Spark 允许将结果写入 S3。配置 Grafana（或其他数据可视化工具）连接到存储在 S3 中的日志数据，以生成实时监控仪表盘和报告。具体代码如下：   |  | | --- | | query = status\_code\_count.writeStream \  .outputMode("append") \  .format("parquet") \  .option("path", "s3://your-bucket/logs/") \  .option("checkpointLocation", "s3://your-bucket/checkpoints/") \  .start() | |
| 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 |
| 增强日志数据解析，支持多种日志格式解析，提高系统的灵活性。增加异常检测和告警，实时检测异常行为并触发告警，提升系统的监控能力。多维度分析和报告，实现用户行为分析和流量来源分析，生成详细报告。 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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9-1 实时日志处理系统 | | | | | | |
| 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 | ☆☆☆☆☆ | | | | | | | | | | |