

**大数据处理实验报告**

实验四：大数据实时分析

专业班级： CS2002

学 号： U202015324

姓 名： 屈绍博

指导教师： 石宣化

报告日期： 2022.4.18

**计算机科学与技术学院**

**《大数据处理》课程实验报告**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实验地点 | | 南一楼804 | 课程名称 | | 大数据处理 | | |
| 实验题目 | | 大数据实时分析 | 成绩 |  | | 指导教师 |  |
| 教师评价 | □ 实验过程正确； □ 源程序/实验内容提交； □ 程序结构/实验步骤合理；  □ 实验结果正确； □ 语法、语义/命令正确； □ 报告规范；  其他： | | | | | | |
| **一、实验目的**   1. 了解大数据实时分析的用途 2. 掌握大数据实时分析的基本命令   **二、实验内容**   1. 实验环境配置 2. Python脚本生成测试数据 (20’) 3. 配置Kafka (10’) 4. 安装Flume客户端 (10’) 5. 配置Flume采集数据 (20’) 6. MySQL中准备结果表与维度表数据 (10’) 7. 使用DLI中的Flink作业进行数据分析 (20’) 8. 资源释放 9. 实验总结 （10’）   **三、实验环境**  （1）集群  版本：MRS 1.9.2  类型：流式集群  组件选择：Storm 1.2.1, Kafka 1.1.0, KafkaManager 1.3.3.1 and Flume 1.6.0  （2）CPU架构  Master节点  鲲鹏通用计算增强型： 4 vCPUs | 16 GB | kc1.xlarge.4  系统盘：通用型SSD 100 GB x 1  数据盘：通用型SSD 200 GB x 1  分析Core节点  鲲鹏通用计算增强型： 4 vCPUs | 16 GB | kc1.xlarge.4  系统盘： 通用型SSD 100 GB x 1  数据盘： 通用型SSD 100 GB x 1  （3）数据库  引擎：MySQL  版本：5.7  实力类型：单机  存储类型：SSD云盘  性能规格：鲲鹏通用增强型 | 2 vCPUs | 4 GB, 最大连接数: 1200,  TPS/QPS: 340 | 7100  （4）队列  类型：通用队列  AZ策略：单AZ  CPU架构：鲲鹏  规格：16 CUs  **四、实验过程或步骤（源程序）** 1．Python 脚本生成测试数据  * 1. **利用putty登录**   打开 Putty，输入前面为 MRS 的 master 节点绑定的公网 IP    **图1.1 利用putty登录公网IP**  输入用户名root和密码后即可登入    **图1.2 登入成功**   * 1. **进入/opt/client/目录，使用 vim 命令编写 Python 脚本：autodatagen.py**     **图1.3 操作结果**    **图1.4 添加脚本**  并在autodatagen.py中输入python指令    **图1.5 写入脚本**   * 1. **创建存放测试数据的目录**   使用 mkdir 命令在/tmp 下创建目录 flume\_spooldir    **图1.6 创建目录**   * 1. **利用脚本生成数据并进行查看**   执行 Python 命令，测试生成 100 条数据    **图1.7 运行脚本**  使用 more 命令查看生成的数据    **图1.8 查看数据** 2.配置 Kafka  1. **设置环境变量**   使用 source 命令进行环境变量的设置    **图2.1 环境变量设置**   1. **在 kafka 中创建 topic**   利用mrs manager查找zookeeper对应的IP，并利用下方指令创建topic    **图2.2 创建topic**   1. **查看topic信息**     **图2.3 查看topic信息** 3. 安装 Flume 客户端  1. **校验下载的客户端文件包**   按照步骤安装客户端后，我们需要对客户端文件包进行验证  进入/tmp/MRS-client 目录并调用ll观察文件目录    **图3.1 校验压缩包**   1. **解压压缩包获取校验文件与客户端配置包**   执行tar -xvf MRS\_Flume\_Client.tar来解压压缩包获取客户端配置包    **图3.2 操作结果**   1. **校验文件包**   执行sha256sum -c MRS\_Flume\_ClientConfig.tar.sha256校验文件包    **图3.3 操作结果**   1. **安装 Flume 运行环境**   首先解压“MRS\_Flume\_ClientConfig.tar”文件，    **图3.4 操作结果**  并安装客户端运行环境到目录“/opt/Flume\_env”    **图3.5 操作结果**  显示complete，安装成功   1. **执行命令配置环境变量。**     **图3.6 操作结果**   1. **安装 Flume 客户端**   执行命令解压 Flume 客户端文件。    **图3.7 操作结果**  安装 Flume 到目录“/opt/FlumeClient”    **图3.8 操作结果**   1. **重启 Flume 服务**   利用restart指令重启flume指令    **图3.9 操作结果** 4. 配置 Flume 采集数据  1. **在 conf 目录下编辑文件 properties.properties为如下内容**     **图4.1 文件编辑结果**  按照自己的kafka IP修改其中的内容   1. **创建消费者消费 kafka 中的数据**   执行创建指令    **图4.2 操作结果**  并新建窗口    **图4.3 打开二窗口**  调用python脚本随机生成数据    **图4.4 随机生成数据** 5. MySQL 中准备结果表与维度表数据  1. **创建数据库**     **图5.1 操作结果**   1. **进入 SQL 界面**     **图5.2 操作结果**   1. **创建维度表并插入数据**     **图5.3 操作结果**   1. **创建flink作业的结果表**     **图5.4 操作结果**  更新语句，执行后可以看到已经创建的表 6. 使用 DLI 中的 Flink 作业进行数据分析  1. **创建 Flink 作业**     **图6.1 操作结果**   1. **编辑 Flink 作业的 SQL 脚本**   按照自己的IP地址编辑脚本，启动作业    **图6.2 操作结果**   1. **查看作业运行详情**   可以在页面中检测运行日志    **图6.3 操作结果**   1. **验证数据分析**   使用 PuTTY 登录到 MRS 的 master 节点后，进入 Python 脚本所在目录，再次执 行 Python 脚本生成测试数据    **图6.4 监视窗口**  可以观察到表中有数据进入   1. **修改脚本使生成测试数据的 Python 脚本每隔 10 秒钟运行 一次**   我们可以在main函数中加入循环，并让每次循环十秒钟运行一次，这样便可以实现每十秒钟写入一次数据。    **图6.5 修改脚本**  在监视中也可以观察到数据不断被写入    **图6.6 监视窗口**  **五、出现的问题与解决方案**  此次实验遇到的问题较多，主要是由于操作过多，一点点小的错误即使不会对下一步操作造成很大的影响，也可能为之后的操作埋下隐患，在问题出现很久之后才暴露出来。  比如我在建立了所有的集群之后开始利用putty进行操作，一开始的登入，脚本的编写都十分顺利，但当我尝试登入zookeeper时系统总是提示我没有权限，在我多次检查IP和安全组配置无误后还是会出现这个错误，后来请教助教后才发现是我的第一个集群错误勾选了Kerberos认证，导致登入失败，由于Kerberos配置无法中途修改，我只能删掉这个集群重新开始，但是后面建立的所有集群和数据湖等都利用了这个集群的安全组，绑定了这个集群的IP，串行工作的第一步出现了错误，我只能将所有工作推倒重来，重新建立了所有集群。才让后面的工作可以正常进行。  在后面的DLI中的Flink作业任务中，在测试网络连通性时我又遇到了问题，我的网络总是无法被打通，我仔细检查了我的MySQL脚本，发现在脚本中存在一些错误，修改之后再次对连通性进行测试发现还是失败，我便在安全组中添加了对应端口，终于解决了这个问题。  **六、实验总结**  这是大数据处理的最后一次实验，我们这一次将python和MySQL结合了进来，也利用到了我们之前学过的一些知识，不过总体的任务还是根据任务书对这个系统进行摸索和熟悉，总体较为简单，但由于操作较多的问题，一点点小的错误即使不会对下一步操作造成很大的影响，也可能为之后的操作埋下隐患，在问题出现很久之后才暴露出来，这也提示我们写程序要注意细节，犯错误很容易，但找出错误并解决错误有时会付出很大的代价。 | | | | | | | |