****

**实验二：实时分析**

**实验报告**

|  |  |
| --- | --- |
| **学 期：** | 2023-2024学年 第二学期 |
| **专 业：** | 计算机科学与技术（实验班） |
| **学号姓名：** | 21071003 高立扬 |
| **完成日期：** | 2024.05 |

### 问题

1. 场景说明

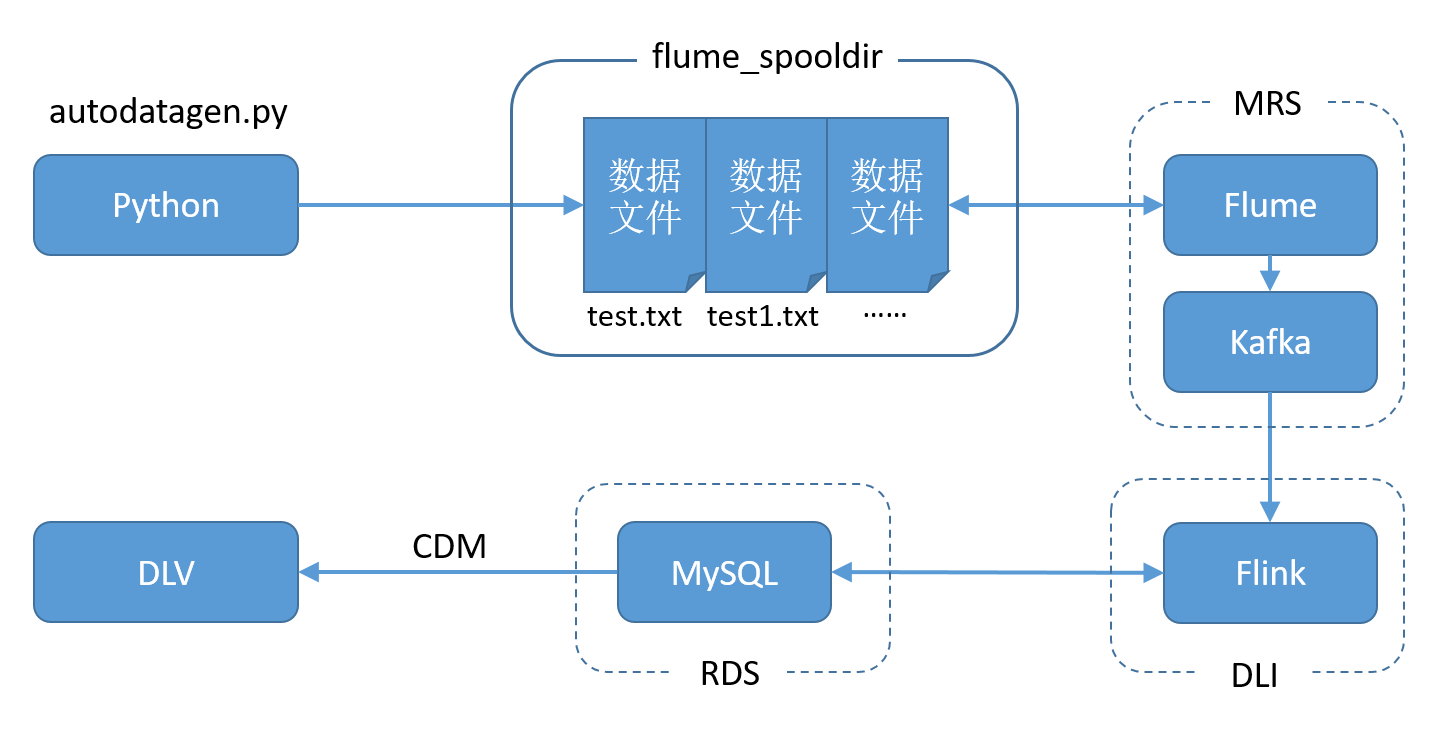
实时分析是大数据五大应用场景中比较重要的场景，通常要求在极短的时间内对产生的海量数据进行有效分析并给出响应。处理时间要求极高，一般都是毫秒级。处理数据量巨大，一般每秒产生数据在百兆以上。像购物、新闻、视频等信息流推荐，电商销售数据实时统计都是实时分析的典型应用。

1. 实验介绍

本实验通过对某电商网站双十一模拟的销售数据进行实时分析（商品销售总额和销售总额排名前五的店铺），掌握实时分析的基本流程和技术架构。

本实验通过编写的Python脚本模拟生产商品的实时销售数据，然后由Flume+Kafka完成数据采集，通过DLI中的Flink作业进行数据分析，分析结果存入RDS中，最后DLV读取RDS中的数据通过大屏完成可视化。

数据处理流程示意：



1. 相关云服务说明

·MRS (MapReduce Service)

MapReduce服务是一个在华为云上部署和管理Hadoop系统的服务，一键即可部署Hadoop集群。MRS提供租户完全可控的一站式企业级大数据集群云服务，完全兼容开源接口，结合华为云计算、存储优势及大数据行业经验，为客户提供高性能、低成本、灵活易用的全栈大数据平台，轻松运行Hadoop、Spark、HBase、Kafka、Storm等大数据组件，并具备在后续根据业务需要进行定制开发的能力。

·DLI (Data Lake Insight)

数据湖探索是完全兼容Apache Spark、Apache Flink、openLooKeng（基于Apache Presto）生态，提供一站式的流处理、批处理、交互式分析的Serverless融合处理分析服务。用户不需要管理任何服务器，即开即用。支持标准SQL/Spark SQL/Flink SQL，支持多种接入方式，并兼容主流数据格式。数据无需复杂的抽取、转换、加载，使用SQL或程序就可以对云上CloudTable、RDS、DWS、CSS、OBS、ECS自建数据库以及线下数据库的异构数据进行探索。

·RDS (Relational Database Service)

华为云关系型数据库是一种基于云计算平台的即开即用、稳定可靠、弹性伸缩、便捷管理的在线关系型数据库服务，支持单机和主备部署模式，支持MySQL、PostgreSQL、SQL Server等主流的关系型数据库引擎。云数据库RDS服务提供了专业的数据库管理平台， 让用户能够在云上轻松的进行设置和扩展云数据库，用户无需编程就可以执行所有必需任务，简化运营流程，减少日常运维工作量，从而专注于开发应用和业务发展。

·CDM (Cloud Data Migration)

云数据迁移是提供同构/异构数据源之间批量数据迁移服务，帮助客户实现数据自由流动。支持文件系统，关系数据库，数据仓库，NoSQL，大数据云服务和对象存储等数据源，无论是客户自建还是公有云上的数据源。

·DLV (Data Lake Visualization)

数据可视化是一站式数据可视化开发平台，提供丰富多样的2D、3D可视化组件，采用拖拽式自由布局，旨在帮助用户快速定制和应用属于自己的数据大屏。适配云上云下多种数据源，无缝集成华为云数据仓库服务、数据湖探索、关系型数据库、对象存储服务等，支持本地CSV、在线API及企业内部私有云数据。

### 解决方法

1. 华为云服务器配置

进入控制台后，鼠标移动到左侧三根横线处将“服务列表”展开，并在列表中点击“网络”分类里面的“弹性公网IP EIP”。步骤2点击右上角的“购买弹性公网IP”计费模式选择“按需计费”，区域选择“华北-北京四”，线路选择“全动态BGP”，公网带宽选择“按流量计费”，带宽大小选择“50”，购买量选择“2”。返回到控制台后，点击刷新按钮可以看到已购买的2个弹性IP。

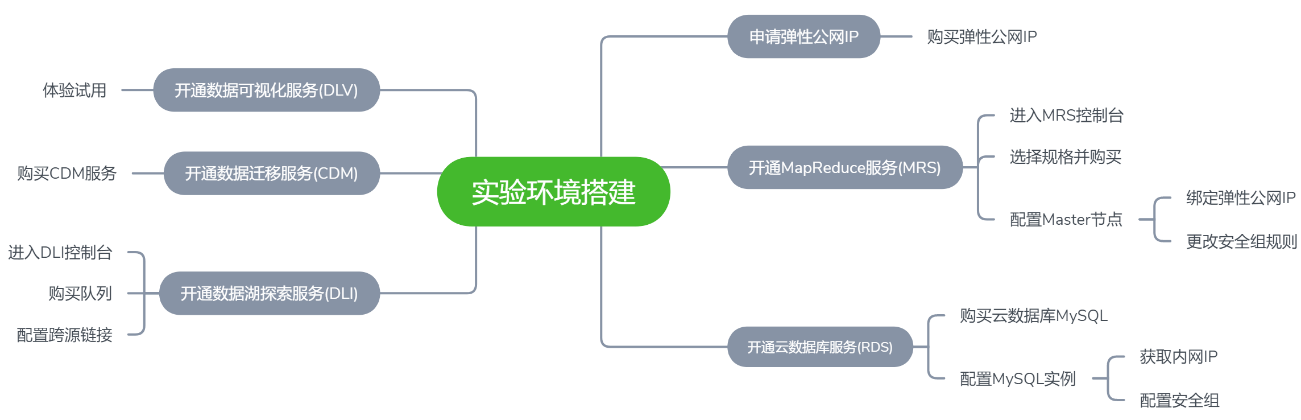
在控制台展开“服务列表”，在列表中点击“EI企业智能”分类里面的“MapReduce服务”。在MRS控制台页面点击右上角的“购买集群”。选择“自定义购买”进入软件配置界面。区域选“华北-北京四”。 集群名称“mrs\_rta”（可以自定义），集群版本选则“MRS 1.9.2”，集群类型选择“流式集群”并勾选所有组件。进入硬件配置界面，计费模式选择“按需计费”。可用区默认，选择虚拟私有云（如果还没有虚拟私有云，点击“查看虚拟私有云”进入虚拟私有云控制台，参照附录5.1进行创建，创建后回到MRS控制台并点击虚拟私有云后面的刷新按钮），子网会自动选定，安全组“自动创建”，弹性公网IP处点击下拉选择一个前面购买的弹性公网IP（此处IP是MRS集群管理界面的IP）。CPU架构选择“鲲鹏计算”，关闭Master节点实例数量中的“集群高可用”，流式Core节点的实例数量改为“1”（注：若没有当前规格选择其他规格也可以）。标签、弹性伸缩、引导操作保持默认。告警选择“关闭”，委托、数据盘加密默认即可。关闭“Kerberos认证”，按规则设置密码（该密码用于登录集群管理页面）。登录方式选择“密码”，输入密码（此处为登录MRS集群节点ECS机器的面）。通信安全授权，勾选“确认授权”复选框，点击右下角的“立即购买”。在现有集群列表中点击集群名称，进入集群页面。在集群页面点击“节点管理”，点击下箭头展开master\_node\_default\_group节点组，点击master的节点名称进入master节点（华为云只在master节点上安装了组件的客户端，登录其他节点无法操作组件）。选择“弹性公网IP”，点击“绑定弹性公网IP”（弹性IP在前序步骤中已经购买）。选择一个弹性公网IP后点击“确定”。弹性公网IP绑定完毕，后面可以使用PuTTY等远程工具通过此IP登录集群的master节点服务器进行相关的操作。在master节点界面选择“安全组”，点击下箭头展开安全组，点击“更改安全组规则”（配置好IP后需要修改安全组规则，否则无法登陆到master节点服务器）。优先级为“1”（优先级范围1~100，数字越小级别越高），策略“允许”，协议端口处选择“TCP”>“基本协议”>“全部放通”，然后点击确定MRS服务到此购买并配置完成。

在控制台展开“服务列表”，在列表中点击“数据库”分类里面的“云数据库RDS”。在实例管理中，点击右上角的“购买数据库实例”。计费模式选择“按需计费”，区域选择“华北-北京四”。实例名称“rds-rta”（可以自定义），数据库引擎“MySQL”，数据库版本“5.7”，实例类型“单机”，存储类型、可用区、时区默认。性能规格选择“鲲鹏通用增强型”>“2 vCPUs | 4 GB”，存储空间、磁盘加密默认。选择虚拟私有云和子网（使用与MRS相同的虚拟私有云和子网），安全组、端口默认即可。在实例管理界面，点击刚创建的实例名称（如果浏览器关闭了，可以通过控制台-服务列表-数据库-中再找到云数据库RDS进入）。进入RDS实例页面，点击左侧的“连接管理”，选择“内网连接”可以查看到MySQL的内网IP地址，此IP地址会在后面步骤中用到（点后面的方形按钮可以复制）。在内网连接界面向下滚动页面，在“安全组规则”中，选择“入方向规则”，然后点击“一键添加”。

在控制台展开“服务列表”，在列表中点击“EI企业智能”分类里面的“数据湖探索DLI”。计费模式选择“按需计费”，区域选择“华北-北京四”。 队列名称“queue\_flink”（可以自定义），队列类型选择“通用队列”，勾选专属资源模式，AZ策略选择“单AZ”，CPU架构“鲲鹏”，队列规格“16CUs”，高级配置、标签默认。在“全局配置”中选择“服务授权”，勾选“VPC Administrator”（创建跨源连接需要）。在DLI控制台页面，点击“跨源连接”菜单，在“增强型跨源”页签下点击 “创建”按钮。输入连接名称“DLI\_RDS”(可自定义)，绑定队列选择上一步创建的DLI队列，选择虚拟私有云和子网（均与MRS服务相同），主机信息、标签默认，点击“确定”。点击跨源连接的连接名称，进入到连接的详情页面，可以看到连接状态为“CREATING”，表示对等连接正在创建，创建过程需要几分钟的时间。

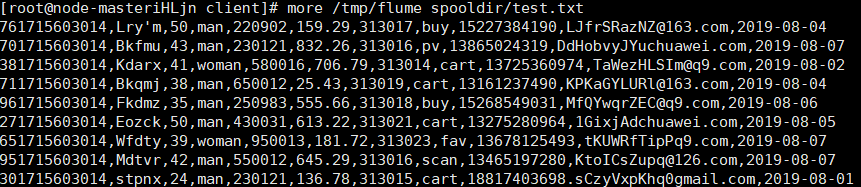
进入华为云主页，登录后点击“产品”>“大数据”>“数据可视化”>“数据可视化DLV”进入可视化DLV服务（也可以在“搜索产品”填写“DLV”快速找到服务）。点击“体验试用”（每个用户有一次体验试用的机会，如果已经没有体验试用机会，点击“进入控制台”后购买“基础版”）。

1. 工作流程



### 实验结果

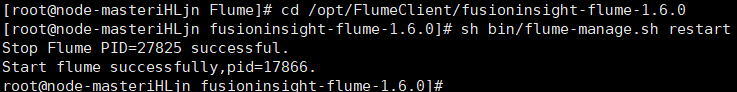
1. Python脚本生成测试数据



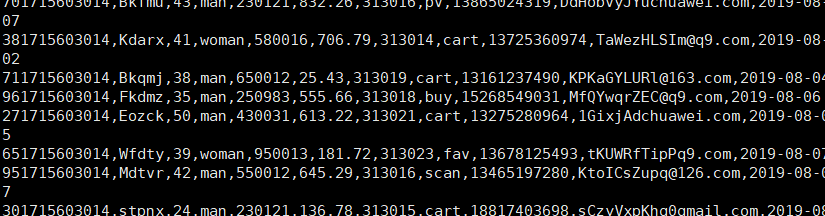
1. 配置Kafka



1. 安装Flume客户端



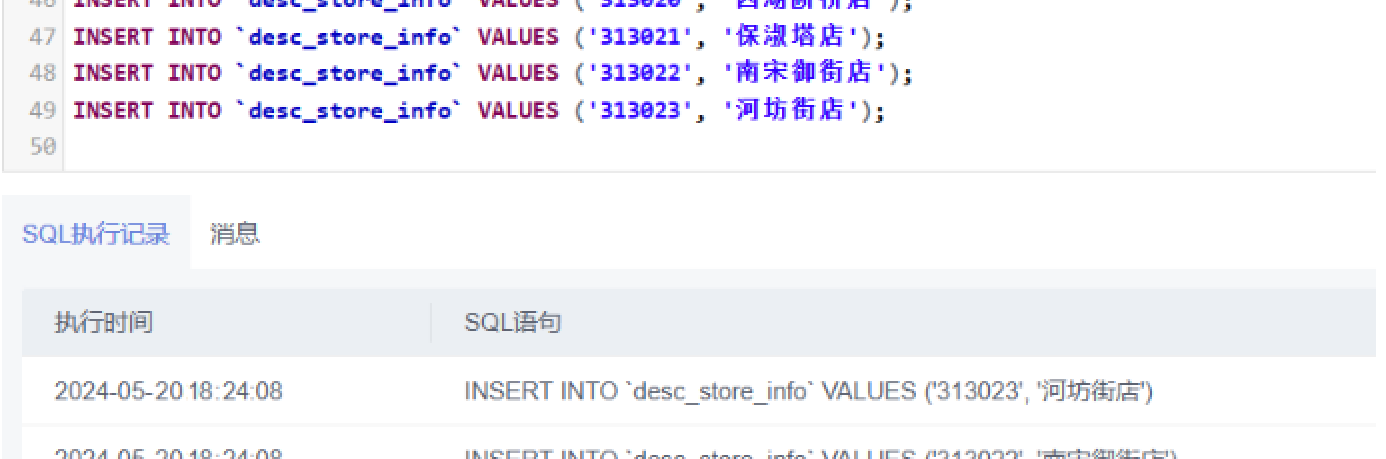
1. 配置Flume采集数据（测试Flume到Kafka是否打通）



1. MySQL中准备结果表与维度表数据



1. 使用DLI中的Flink作业进行数据分析



1. DLV数据可视化

