学号 专业 姓名

实验日期 教师签字 成绩

实验报告

【实验名称】 基于Hive的开源软件社区数据分析

【实验目的】

掌握Hive架构原理及安装方法

掌握HiveQL常用语法

掌握Sqoop2架构原理及安装方法

掌握Sqoop2 任务创建方法

使用Hive进行开源软件社区数据分析

【实验原理】

见操作步骤文档

【实验内容】

**题目**：安装配置Hive和Sqoop2，并基于Hive完成开源软件社区数据分析。

**要求**：1.安装、配置Hive，使用Beeline连接Hive Server2；

2.安装配置Sqoop2；

3.创建Sqoop2任务，将实验1中爬取的SegmentFault数据导入到Hive数据表；

4.导入GHTorrent数据集，创建相关Hive表（外部表）：users、projects、commits、watchers和issues；

5.融合GHTorrent数据集和爬取的SegmentFault数据集，构造用户宽表和项目宽表，抽取尽可能多的属性。