**案例简介**

大数据课程实验案例：网站用户行为分析，本案例涉及数据预处理、存储、查询和可视化分析等数据处理全流程所涉及的各种典型操作，涵盖Linux、MySQL、Hadoop、HBase、Hive、Sqoop、R、Eclipse等系统和软件的安装和使用方法。

**案例目的**

1. 熟悉Linux系统、MySQL、Hadoop、HBase、Hive、Sqoop、R、Eclipse等系统和软件的安装和使用；
2. 了解大数据处理的基本流程；
3. 熟悉数据预处理方法；
4. 熟悉在不同类型数据库之间进行数据相互导入导出；
5. 熟悉使用R语言进行可视化分析；（不做要求）
6. 熟悉使用Elipse编写Java程序操作HBase数据库。

**预备知识**

需要案例使用者，已经学习过大数据相关课程，了解大数据相关技术的基本概念与原理，了解Windows操作系统、Linux操作系统、大数据处理架构Hadoop的关键技术及其基本原理、列族数据库HBase概念及其原理、数据仓库概念与原理、关系型数据库概念与原理、R语言概念与应用（不做要求）。

**环境要求**

本案例要求在伪分布式下完成或在集群下完成。

**软件工具**

本案例所涉及的系统及软件

1. Linux系统（Ubuntu16.04或14.04或18.04）
2. MySQL（版本无要求）
3. Hadoop（2.7.1或2.7.3，不能用3.0及以上版本，因为Sqoop工具无法支持Hadoop3.0以上版本）
4. HBase（1.1.2或1.1.5，HBase版本需要和Hadoop版本兼容）
5. Hive（1.2.1，Hive需要和Hadoop版本兼容，不要安装Hive3.0以上版本）
6. Sqoop（必须用1.4.6，注意，Sqoop无法支持Hadoop3.0以上版本）
7. R（版本无要求）（不做要求）
8. Eclipse（版本无要求）

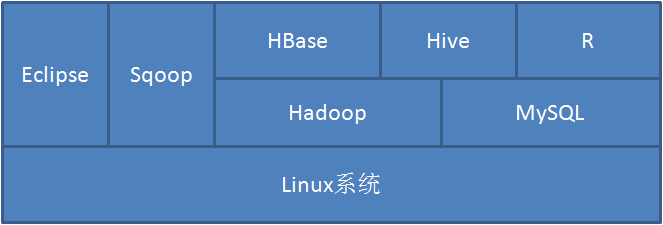


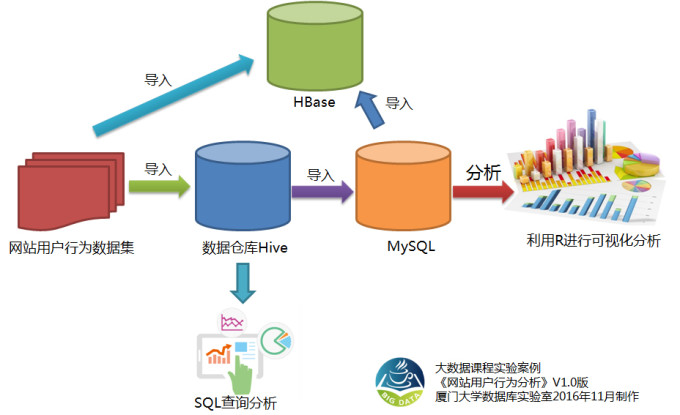
图 案例所涉及软件总体概览图

**数据集**

网站用户购物行为数据集2000万条记录。

**案例任务**

1. 安装Linux操作系统
2. 安装关系型数据库MySQL
3. 安装大数据处理框架Hadoop
4. 安装列族数据库HBase
5. 安装数据仓库Hive
6. 安装Sqoop
7. 安装R（不做要求）
8. 安装Eclipse
9. 对文本文件形式的原始数据集进行预处理
10. 把文本文件的数据集导入到数据仓库Hive中
11. 对数据仓库Hive中的数据进行查询分析
12. 使用Sqoop将数据从Hive导入MySQL
13. 使用Sqoop将数据从MySQL导入HBase
14. 使用HBase Java API把数据从本地导入到HBase中
15. 使用R对MySQL中的数据进行可视化分析（不做要求）



**实验步骤**

步骤零：实验环境准备

步骤一：本地数据集上传到数据仓库Hive

步骤二：Hive数据分析

步骤三：Hive、MySQL、HBase数据互导

步骤四：利用R进行数据可视化分析

每个实验步骤所需要的知识储备、训练技能和任务清单如下：

**步骤零：实验环境准备**

|  |  |
| --- | --- |
| 所需知识储备 | Windows操作系统、Linux操作系统、大数据处理架构Hadoop的关键技术及其基本原理、列族数据库HBase概念及其原理、数据仓库概念与原理、关系型数据库概念与原理 |
| 训练技能 | 双操作系统安装、虚拟机安装、Linux基本操作、Hadoop安装、HBase安装、Sqoop安装、Eclipse安装 |
| 任务清单 | 1. 安装Linux系统；2. 安装Hadoop；3. 安装MySQL；4. 安装HBase；5. 安装Hive；6. 安装Sqoop；7. 安装R；8. 安装Eclipse |

**步骤一：本地数据集上传到数据仓库Hive**

|  |  |
| --- | --- |
| 所需知识储备 | Linux系统基本命令、Hadoop项目结构、分布式文件系统HDFS概念及其基本原理、数据仓库概念及其基本原理、数据仓库Hive概念及其基本原理 |
| 训练技能 | Hadoop的安装与基本操作、HDFS的基本操作、Linux的安装与基本操作、数据仓库Hive的安装与基本操作、基本的数据预处理方法 |
| 任务清单 | 1. 安装Linux系统；2. 数据集下载与查看；3. 数据集预处理；4. 把数据集导入分布式文件系统HDFS中；5. 在数据仓库Hive上创建数据库 |

**步骤二：Hive数据分析**

|  |  |
| --- | --- |
| 所需知识储备 | 数据仓库Hive概念及其基本原理、SQL语句、数据库查询分析 |
| 训练技能 | 数据仓库Hive基本操作、创建数据库和表、使用SQL语句进行查询分析 |
| 任务清单 | 1. 启动Hadoop和Hive；2. 创建数据库和表；3. 简单查询分析；4. 查询条数统计分析；5. 关键字条件查询分析；6. 根据用户行为分析；7. 用户实时查询分析 |

**步骤三：Hive、MySQL、HBase数据互导**

|  |  |
| --- | --- |
| 所需知识储备 | 数据仓库Hive概念与基本原理、关系数据库概念与基本原理、SQL语句、列族数据库HBase概念与基本原理 |
| 训练技能 | 数据仓库Hive的基本操作、关系数据库MySQL的基本操作、Sqoop工具的使用方法、HBase API的Java编程、Eclipse开发工具使用方法 |
| 任务清单 | 1. Hive预操作；2. 使用Sqoop将数据从Hive导入MySQL；3. 使用Sqoop将数据从MySQL导入HBase；4. 使用HBase Java API把数据从本地导入到HBase中 |

**步骤四：利用R进行数据可视化分析（不做要求）**

|  |  |
| --- | --- |
| 所需知识储备 | 数据可视化、R语言 |
| 训练技能 | 利用R语言对MySQL数据库中的数据进行数据可视化分析、R的安装、相关可视化依赖包的安装与使用、各种可视化图表生成方法 |
| 任务清单 | 安装R语言包、安装可视化依赖包、柱状图可视化分析、散点图可视化分析、地图可视化分析 |