**实验六：Hive编程实践**

一、实验目的

能够在本机上安装和部署Hive，并熟练使用常用的HiveQL语句，解决实际大数据分析场景中的问题。

二、实验平台

（1）虚拟机软件：推荐使用的开源虚拟机软件为VirtualBox，可以在本地个人机器的Windows操作系统或MacOS上安装VirtualBox软件，然后在VirtualBox上安装并且运行Linux操作系统。Linux系统推荐使用Ubuntu操作系统，如Ubuntu18.04（或Ubuntu16.04），其他Linux操作系统发行版也可；

（2）Hadoop：3.1.3；

（3）JDK：1.8；

（4）Hive：3.1.3；

（5）编程语言：HiveQL。

三、实验内容和要求

本题任务是使用Hive对一个Google图书的n-grams数据集执行一些简单的分析。“n-gram”[[1]](#footnote-0)是一个包含n个单词的短语。该数据集列出了books.google.com网站上书籍中存在的所有n-gram和一些统计数据。在本题中，仅可使用Google图书二元组bigram (也称1-grams)，对应两个数据文件在dataset文件夹（已经分享到北航云盘[[2]](#footnote-1)）中。这两个文件中的每一行都具有以下格式（字段之间用制表符tab进行分隔）：



文件中一个bigram的示例如下：



该示例表示在1978（1979）年，单词“circumvallate”总共出现了335（261）次，并且来源于91（95）本不同的书籍。

**问题1**：在自己本机之前部署的Hadoop伪分布式集群之上安装和部署Hive，并对本机安装和部署过程以图文形式进行说明。

**问题2**：将两个数据文件上传到 HDFS并将它们合并存储到一张Hive表中。

**问题3**：对于每个独特的bigram，计算其每年出现的平均次数，输出需要保存到HDFS或Hive表中。以上面的数据示例为例，结果应该是：



注：分母是该bigram出现的年数。假设数据集包含过去100年的所有的bigram，并且上述示例的两行记录是单词“circumvallate”的唯二记录。那么平均值应为：



而不是：



**问题4**：输出每年平均出现次数最高的20个bigram（按平均值降序排列），输出需要保存到HDFS或Hive表中。如果多个bigram具有相同的平均值，可以随机选择任何一个排在前面。

**提示**：课程教材《大数据技术原理与应用（第3版）》中9.6节中有一些关于HiveQL的用例说明，可在编写本作业的HiveQL语句时参考。在使用交互式Shell（或其他方式）时，应该首先测试一小部分数据而不是整个数据集。一旦Hive命令/脚本按预期工作，就可以在完整的数据集上运行它们。此外，如果本机性能不足以支撑对整个数据集进行分析，可以采用抽样方式对数据进行采样，在采样的小数据集上进行分析，但需要说明采样的方式。

**提交要求**：针对上述问题完成作业，特别要求针对上述问题中需要编程的任务编写HiveQL语句来进行实现并将输出保存到HDFS或Hive的表中。提交物中应包含实验报告PDF和HiveQL语句（可放在实验报告中）。可参考本文档最后的“附录：课程实验报告模版”进行报告撰写，其中“实验内容与完成情况”部分应给出Hive安装和部署过程、实验过程、源码编译/运行所需的各类条件说明、输入/输出的各类截图/文字说明等。

附录：课程实验报告模版

|  |  |  |
| --- | --- | --- |
| “大数据工程”课程实验报告 | | |
| 题目： | 学号姓名： | 日期： |
| 实验环境： | | |
| 实验内容与完成情况： | | |
| 出现的问题： | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | |

注：报告篇幅可根据实际题目情况进行调整。

1. https://baike.baidu.com/item/n%E5%85%83%E8%AF%AD%E6%B3%95/19133139 [↑](#footnote-ref-0)
2. https://bhpan.buaa.edu.cn/link/AA899727C1871B41CBA72B85065D04BC5F [↑](#footnote-ref-1)