**实验七：Spark编程实践**

一、实验目的

能够在本机上安装和部署Spark；能够针对实际场景下的大数据分析问题，使用Spark Shell和Java语言编写程序，实现数据的统计分析与计算。

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。注：本作业不要求使用希冀大数据平台进行实验。

（2）编程工具或语言：Spark Shell和Java。

三、实验内容和要求

**数据集描述**：

在这一部分中，你将需要实现一个简单的Spark应用程序。我们在北航云盘中[[1]](#footnote-0)提供了一个数据文件，包含了一些IOT（物联网）设备收集的样本数据。

**任务要求**：

1. 要求首先使用Hadoop平台的Shell命令，将上述文件加载到本机的HDFS的某个文件目录中（依赖于课程前面实验环节在本机部署的Hadoop平台）。提示：可参考课程配套指南<https://dblab.xmu.edu.cn/blog/2460/>中的“一、利用Shell命令与HDFS进行交互”。

2. 要求在本机安装和部署Spark平台。提示：可参考课程配套实验指南<https://dblab.xmu.edu.cn/blog/2501/>中“一、安装 Spark2.4.0”进行Spark的安装和部署。

3. 要求在Spark Shell中读取HDFS上的上述文件，然后，统计出文件的行数。提示：可参考课程配套实验指南<https://dblab.xmu.edu.cn/blog/2501/>中“二、使用Spark Shell编写代码”进行Spark Shell代码编写和运行。

4. 要求编写基于Java语言的Spark应用程序，读取HDFS中的上述文件，然后，统计出文件的行数；随后要求使用Maven对Java独立应用程序进行编译打包，并将生成的jar包通过spark-submit提交到Spark中进行运行。提示：可参考课程配套实验指南<https://dblab.xmu.edu.cn/blog/2501/>中“（二）使用Maven对Java独立应用程序进行编译打包”进行Java应用程序的编写和运行。

**提交要求**：按上述要求进行程序编写、编译、提交和运行。作业提交物中应包含Java源代码文件夹和实验报告PDF（可包含Spark Shell代码、Hadoop Shell代码等其他代码文件）。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”部分可参考课程PPT、课程配套实验指南<https://dblab.xmu.edu.cn/blog/2501/>，给出Spark的软件安装和部署过程、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://bhpan.buaa.edu.cn/link/AA3F41728563AA4FF08F016276D15AB229 [↑](#footnote-ref-0)