个人周工作报告

|  |  |  |
| --- | --- | --- |
| 姓名：徐贝澄 | 项目名称：大数据实训第一周 | 填报时间：2020/11/21 |

1.本周工作内容

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目名称** | **项目阶段** | **起始时间** | **本周工作简述** | **任务是否完成** |
| 大数据实训 | 第一周 | 11月16日 | 安装环境，学习hadoop | 是 |
| **具体工作内容** | | | | |
| 1、，第一次课，老师先介绍了分布式存储的概念。以及重点的hadoop存储和读取数据的实现。接着就进入Hadoop 集群环境搭建的环节。拷贝并安装了两台虚拟机，在VirtualBox中运行。由于最原始的界面操作极不友好，所以又安装了SmarTTY获得更加便捷的命令行操作。 | | | | 是 |
| 2、第二次课，重点是hadoop的MapReduce并行计算。学习了：  1. MapReduce的设计思想是分而治之、先分后合。  2. MapReduce的架构是：Map`阶段拆分处理，Reduce阶段汇总得结果。  3. 通过最基础的用MapReduce实现单词频次统计的样例来上手实操，增强理解。 | | | | 是 |
| 3、 第三次课，老师进一步讲了更多样例，体会了Mapper和Reducer的更多变化和用法。主要收获如下：  1. 在不同的样例中主要学习了如何确定 Map 阶段的 kay 和 value，以及如何根据业务需求来设计 Reducer  2. 介绍了 setup 和 cleanup 函数。它们总共只执行一次，分别在Reducer执行前进行资源初始化以及在Reducer执行后全局处理reduce中所有数据map对象或进行释放工作。  3. 补充了Java基础，重点讲了容器类的使用，包括List、Map 等，还有排序等使用的函数。最后，以一个WordCountTop样例训练了Java容器工具类在 MapReduce 中的运用。 | | | | 是 |

注：检核内容如未完成，请在上面说明未完成原因。

|  |  |
| --- | --- |
| **存在问题** | **改进措施及预期结果** |
| 1、在搭建环境后，VirtualBox 中在命令行输入java -version 有正确的输出结果，说明环境变量配置正确。而在SmarTTY中输入同样的指令却终是报错，而查看环境变量文件的时候，发现也已经配置了。 | 这个问题至今还没有解决，所以虽然SmarTTY很方便，但是可能并不能使用它。 |
| 2、在Eclipse工程中导入hadoop包后，无法import，显示......is not accessiable， | 问题在从头再来了两次后，消失了，一切正常。 |
| 3、编码问题，hadoop处理的是数据，而不同编码格式就会导致乱码的出现。主要遇到了以下两种情况：  1. 打开老师上传的代码时，中文注释变成乱码。  2. 在要处理的数据种包含中文时，用`hadoop`读取的数据为乱码。 | 1. 原因：老师可能使用的是GBK编码，而起初IDE中读取文件使用的是UTF-8格式。  解决方案：将IDE中的编码格式修改为GBK方可正常显示。  2. 原因：Hadoop在涉及编码时都是UTF-8，如果文件编码格式是其它类型（如GBK)，则会出现乱码。而我们的txt文件刚好是采用的GBK格式存储，读取后输出自然就出现了乱码。  解决方案：将预处理数据文件编码格式设置为UTF-8 |

2.下周工作安排

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目名称** | **项目状态** | **起始时间** | **本周工作简述** | **任务检核内容** |
| 大数据实训 | 第二周 | 11月23日 | 进一步学习 | 紧跟进度 |
| **其它工作内容** | | | | |
| 1、确定好项目，并着手开始补充欠缺的技术能力。 | | | | 项目报名 |
| 2、在 github 和 CSDN 上寻找一些使用 hadoop的样例，并努力消化。 | | | | 看懂并修改 |
| 3、学习Linux操作系统的基本操作。 | | | | 实操检测 |