

大数据实训第一周周报

一、本周学习实践进展

本周一共三次课，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。

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。

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 `key` 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。

二、遇到后已解决和仍存在的问题

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`，但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。

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. 在要处理的数据种包含中文时，用 `hadoop` 读取的数据为乱码。

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`

三、下周学习实践计划安排

1. 确定好项目，并着手开始补充欠缺的技术能力。

2. 在 `github` 和 `CSDN` 上寻找一些使用 `hadoop` 的样例，并努力消化。
3. 学习 `Linux` 操作系统的基本操作。