实验 3 HBase、Hive 的安装与使用

更新日期：2015-10-25

1. 实验要求

本次实验所有任务都在小组第一次实验安装的**自己的本地 Hadoop 环境**进行。

实验任务

1. 在自己本地电脑上正确安装和运行 HBase 和 Hive。
2. 在 HBase 中创建一张表 Wuxia，用于保存下一步的输出结果。
3. 修改第 2 次实验的 MapReduce 程序，在 Reduce 阶段将倒排索引的信息通过文件输出，而每个词语及其对应的“平均出现次数”信息写入到 HBase 的表“Wuxia”中。
4. 编写 Java 程序，遍历上一步中保存在 HBase 中的表，并把表格的内容（词语以及平均出现次数）保存到本地文件中。
5. Hive 安装完成后，在 Hive Shell 命令行操作创建表(表名：Wuxia(word STRING, count DOUBLE))、导入平均出现次数的数据、查询(出现次数大于 300 的词语)和前 100 个出现次数最多的词。

输出格式

1. 请在实验报告中附上用于展示上述每一步操作结果的屏幕截图（例如 HBase Shell 中 scan ‘Wuxia’的屏幕截图和 Hive Shell 相关操作的屏幕截图）。
2. 第 3 步倒排索引的输出格式同第 2 次实验（“平均出现次数”不用输出）。
3. 第 4 步的输出格式为：“词语 \TAB 平均出现次数”。请节取部分输出内容附在报告中。

选做内容

该部分内容不做要求，供学有余力的同学尝试练习。

1.本次实验额外提供了一份停词表。先将停词表导入到 HBase 中。在进行实验任务的第 3 步时，在 Map 阶段查询 Hbase 中的停词表，并对停用词进行过滤（不对停用词进行统计）。

2. 实验数据

本次实验提供了金庸、梁羽生等五位小说家的作品全集作为第 3 步的输入数据。每部小说对应一个文本文件。

文本文件均使用 UTF-8 字符编码，并且已分词，两个汉语单词之间使用空格分隔。

停词表也为 UTF-8 编码，每个词语占用一行。

全部数据集：全部数据集请从集群的 HDFS 上下载，并导入到自己的本地环境中。