C0861 “云计算导论”实践 7

准备工作

1. 传统平台下的文献管理系统。
2. 伪分布式或分布式 Hadoop 平台的安装和配置。
3. 基于 HDFS 的附件存储和基于 HBase 的结构化数据存储。

实践描述

1. 按照如下方式写入 2 万条文献记录：
   (1) 在网上找一个 2000 个英文单词的词典。
   (2) 将词典中每个词在文献标题中出现 50 次，每个单词在同一个文献标题中只出现一次，文献标题的长度在 2-10 个单词之间。
   (3) 每条文献的作者由词典中随机选择 2-3 个单词组成。
   (4) 指定 50 种会议或期刊名称，每条文献从中随机选择一种。
   (5) 年份在 1900-2014 之间随机选择一个。
2. 采用 MapReduce 对 HBase 中文献的标题、作者等分别建立倒排文件。
3. 将倒排文件载入内存，进行文献检索。

提交内容

1. 源代码。（40%）
2. 配置文档。（40%）
3. 口头报告幻灯片，报告时间为 10 分钟。（20%）
4. 组内分工，包括小组成员的学号、姓名和贡献比例（各成员的贡献比例之和为 100%）。

说明

1. 提交截止时间为 2014-03-12 23:59:59，提交方式为 TSS。
2. 不需要进行系统演示。