C0861 “云计算导论”实践 8

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. 扩展字典，按照实践 7 的要求，产生 10 万条文献记录，并写入 HBase。在每条记录中添加如下信息：
   (1) 文献的录入者：生成 100 个用户的姓名，每个用户录入了 1000 条文献。
   (2) 录入时间：必须在文献出版年份与当前时间之间，随机生成，精确到天。
   (3) 文献的评价：每篇文献随机生成 2-20 条评价，每条评价包含 1 名评价者和评价时间。评价者从用户姓名中随机选取；评价时间必须在录入时间与当前时间之间，随机生成，精确到天。
2. 采用 MapReduce 进行以下统计：
   (1) 每个用户在过去一周、一个月、半年、一年、所有时间中录入的文献数量。
   (2) 每个用户在过去一周、一个月、半年、一年、所有时间中评价的文献数量。
   (3) 将文献按照评价次数排序，如果评价次数相同，平均评价时间越迟的排序越高。
3. 在文献记录中随机选择 100 条记录，将每条记录重复插入 HBase 中 2-5 遍（次数随机）。采用 MapReduce 查找其中的重复记录。

提交内容

1. 源代码。（60%）
2. 口头报告幻灯片，报告时间为 10 分钟。（40%）
3. 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 2014-03-19 23:59:59，提交方式为 TSS。
2. 不需要进行系统演示。