|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 6 月6日－2015 年6月8日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 |  | |  | | 1 |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 删除疑似表中前三个月未处理和非暴恐的信息 | | 吴大衍 | 1 | Client短期内频繁连接服务器导致无法绑定端口 | 修改linux内核配置参数，快速回收端口号 | | | | |
| 撰写论文实验部分 | |
| 星期二 | | 编写fastdfs自动备份和删除程序 | | 吴大衍 | 2 |  |  | | | | |
| 撰写论文实验部分 | |
| 星期三 | |  | | 吴大衍 | 3 |  |  | | | | |
| 星期四 | |  | | 吴大衍 | 4 |  |  | | | | |
|  | |  |  | | | | |
| 星期五 | |  | | 吴大衍 | 5 |  |  | | | | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 5 月30日－2015 年6月3日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 算法描述及其理论分析部分的撰写（包括BFSS 和SBFSS两种算法） | | 完成 | | 1 | 论文实验部分加结尾部分撰写 | | 完成 | 周五 | 王老师 | 吴大衍 |
|  |  | |  | |  |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | BFSS 算法时间复杂度和空间复杂度分析部分撰写，包括算法参数选择分析 | | 吴大衍 | 1 |  |  | | | | |
| 星期二 | | BFSS 存在问题及其SBFSS算法描述部分撰写 | | 吴大衍 | 2 |  |  | | | | |
| 星期三 | | SBFSS算法描述及其false positive理论分析部分撰写 | | 吴大衍 | 3 |  |  | | | | |
| 星期四 | | 阅读SBF算法论文，总结SBF参数选择策略 | | 吴大衍 | 4 |  |  | | | | |
| SBFSS算法false negative部分撰写，包括符合选择参数等 | |
| 星期五 | | SBFSS算法时间复杂度和空间复杂度理论分析部分撰写 | | 吴大衍 | 5 |  |  | | | | |
| 修改论文已完成部分，主要包括用词和语法 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 5 月23日－2015 年5月27日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 线上系统fastdfs的备份以及过期数据的删除 | | 完成 | | 1 | 实现线上系统的实时备份 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 撰写论文理论分析部分 | | 基本完成 | | 2 | 完成论文算法理论分析部分以及部分实验的撰写 | | 完成 | 周五 | 王老师 | 吴大衍 |
|  |  | |  | |  |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 安装分布式fastdfs（支持多盘多group） | | 华宇  吴大衍 | 1 | 单client多线程上传下载文件会报错 | 多client实现并行上传下载 | | | | |
| 编写存储过程定时调用任务给华宇生成数据 | |
| 星期二 | | 生成映射表（包括近一个月的疑似数据以及样本数据） | | 吴大衍 | 2 | Fastdfs启动一次可能无法成功 | 每次启动需要观测每台storage server的状态，只有active才证明启动正确，否则需要重启 | | | | |
| 备份线上fastdfs数据 | |
| 星期三 | | 调试无法upload的bug | | 魏美茹  吴大衍 | 3 | 调用api下载fastdfs上端口为8888的文件，无法通过返回值判断是否下载成功，下载失败返回值不一定为-1 | 需要针对8888端口的数据多次下载 | | | | |
| 远程协助魏美茹调试缩略图生成与上传部分bug | |
| 星期四 | | 完成fastdfs的数据备份 | | 组内成员  吴大衍 | 4 | 加载端调用ffmpeg无法生成图片 | 加载从ftp拉取数据的方式不对导致视频文件在传输过程中损坏，改为二进制传输即可 | | | | |
| 和钟老师讨论研究近况 | |
| 讨论班讨论 | |
| 星期五 | | 安装nginx，支持http访问fastdfs数据 | | 王老师  吴大衍 | 5 | 部分数据通过http无法访问 | 1. 对于盘符超过9的盘，fastdfs会用16进制表示而不是10进制，所以需要将超过9的盘符改成16进制。 2. Fastdfs有延时同步机制，即在同一个group内的数据需要延迟一段时间才能同步，导致访问其中一个的url不能拿到另一台server上的数据 | | | | |
| 向王老师汇报研究进展 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 5 月16日－2015 年5月20日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 完成缩略图程序的开发和部署 | | 完成 | | 1 | 线上系统fastdfs的备份 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 调研fastdfs的工作原理（包括安装、配置等） | | 完成 | | 2 | 给华宇生成展示所需数据 | | 完成 | 周二 | 钟老师 | 吴大衍 |
| 3 | 论文完成研究现状以及部分算法介绍的撰写 | | 完成 | | 3 | 论文完成算法部分的撰写 | | 完成 | 周五 | 王老师 | 吴大衍 |
|  |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 党课学习 | | 吴大衍 |  |  |  | | | | |
| 学习fastdfs工作原理 | |
| 完成论文概述部分撰写 | |
| 星期二 | | 部署分布式fastdfs | | 吴大衍 |  |  |  | | | | |
| 撰写论文研究现状部分 | |
| 星期三 | | 按照华宇展示需求修改存储过程 | | 王蒙  吴大衍 |  |  |  | | | | |
| 完成论文研究现状部分的撰写 | |
| 星期四 | | 配合魏美茹部署新版加载程序（加入缩略图的生成及上传功能） | | 魏美茹  吴大衍 |  |  |  | | | | |
| 撰写论文算法部分 | |
| 星期五 | | 撰写论文算法部分 | | 贾德宾  王卓  吴大衍 |  |  |  | | | | |
| 参加定向越野比赛 | |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 5 月9日－2015 年5月13日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 图像检索部分V3的撰写 | | 完成 | | 1 | 调研fastdfs的基本原理，测试其鲁棒性 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 声纹识别以及移动终端安全防护v2和v3的撰写 | | 完成 | | 2 | 论文概述和研究现状部分的撰写 | | 完成 | 周五 | 王老师 | 吴大衍 |
| 3 | 和名扬协同开发视频缩略图的生成以及上传部分代码 | | 完成 | |  |  | |  |  |  |  |
| 4 | 论文摘要和概述部分撰写 | | 基本完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 图像检索部分V3撰写 | | 钟老师  吴大衍 |  |  |  | | | | |
| 协调落实fileicon字段的填充过程 | |
| 星期二 | | 完成2-pass算法和mysql的对比实验 | | 王老师  吴大衍 |  |  |  | | | | |
| 和王老师讨论并确认下一步工作 | |
| 星期三 | | 硕士答辩记录 | | 组内老师  吴大衍 |  |  |  | | | | |
| 开发视频缩略图生成部分代码 | |
| 星期四 | | 开会安排文档中支撑组件分布的撰写分工 | | 小组成员  吴大衍 |  |  |  | | | | |
| 和魏美茹协同开发 | |
| 组内讨论班 | |
| 星期五 | | 撰写声纹识别和移动终端安全防护v2v3 | | 吴大衍 |  |  |  | | | | |
| 撰写论文摘要和概述部分 | |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 5 月3日－2015 年5月6日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 基于BFSS 设计 two pass 算法并和传统数据库进行实验对比 | | 基本完成 | | 1 | 完成two pass算法和传统数据库的对比实验并向王老师汇报 | | 完成 | 周五 | 王老师 | 吴大衍 |
| 2 | 查找图像检索系统的相关材料，撰写V1 | | 完成 | | 2 | 完成V2撰写 | | 完成 | 周二 | 李 老师 | 吴大衍 |
| 3 |  | |  | | 3 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期二 | | 阅读提出冰山查询的论文和RICHARD M.CARP的two pass算法解决冰山查询的论文 | | 吴大衍 | 1 |  |  | | | | |
| 星期三 | | 基于BFSS 算法设计two pass算法查找精确低频项 | | 吴大衍 | 2 |  |  | | | | |
| 星期四 | | 查找图像检索系统的相关资料，阅读QBIC系统实现的论文 | | 吴大衍 | 3 |  |  | | | | |
| 生成two pass算法和传统数据库对比的实验数据 | |
| 星期五 | | 撰写V1 | | 吴大衍 | 4 |  |  | | | | |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 4 月25日－2015 年4月29日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 完成BFSS算法在合成数据集上的实验 | | 完成 | | 1 | 验证BFSS算法在传统数据仓库中近似查询的有效性 | | 完成 | 周三 | 王老师 | 吴大衍 |
| 2 | 撰写材料 | | 完成 | | 2 | 完成BFSS算法在真实数据集上的实验 | | 完成 | 周五 | 王 老师 | 吴大 衍 |
| 3 |  | |  | | 3 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 改进实验代码 | | 吴大衍 | 1 | 低频项输出算法需要遍历值域 | 1. 实验评估算法这部分代价 2. 验证BFSS在传统数据仓库中近似查询的有效性 3. 改进这部分算法 | | | | |
| 学习java进程资源监控工具jprofile的用法 | |
| 星期二 | | 在合成数据集上进行实验并记录实验结果 | | 吴大衍 | 2 | 算法需要真实数据做支撑验证其研究意义 | 寻找真实数据，在真实数据上做实验，验证算法的研究意义 | | | | |
| 星期三 | | 完成BFSS算法在合成数据上的实验 | | 吴大衍 | 3 |  |  | | | | |
| 按照分工撰写材料 | |
| 星期四 | | 阅读bloom filter改进算法stable bloom filter算法论文 | | 王老师  吴大衍 | 4 |  |  | | | | |
| 梳理低频项研究进展并制作ppt | |
| 向王老师汇报工作进展并就相关问题以及研究后续工作向王老师请教 | |
| 星期五 | | 和钟老师讨论工作进展 | | 李老师  钟老师  吴大衍 |  |  |  | | | | |
| 按照分工撰写材料 | |  |
| 按照李老师的修改意见修改材料 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 4 月18日－2015 年4月22日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 详细阅读ES search阶段代码 | | 完成 | | 1 | 完成BSS算法在zipf分布上的实验并向王老师汇报 | | 完成 | 周三 | 王老师 | 吴大衍 |
| 2 | 阅读Impala 2.4代码，了解其多线程scan实现原理 | | 完成 | | 2 |  | |  |  |  |  |
| 3 | 设计低频项挖掘算法BSS并做实验 | | 基本完成 | | 3 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 阅读ES 2.1.0search阶段代码，主要回答了其是否包括流水线作业以及底层lucene语法支持等 | | 吴大衍 | 1 | 如何为BSS算法找到合适背景 | 低频项对数据流的信息熵贡献很大，说明了低频项包含的信息量很大，可以利用低频项挖掘很多有意义的信息，比如用户的喜好厌恶等。 | | | | |
| 星期二 | | 阅读impala2.4代码，了解其scan部分多线程实现原理 | | 吴大衍 | 2 | 真实的实验数据来源如何获取。很多论文中没有提真实数据如何获取，即使有也无法下载 | 先在zipf数据上做实验，汇报的时候向王老师请教。 | | | | |
| 星期三 | | 设计低频项挖掘算法BSS | | 王老师  吴大衍 | 3 |  |  | | | | |
| 和王老师讨论低频项挖掘工作进展 | |
| 星期四 | | 编写BSS算法实验代码 | | 钟老师  吴大衍 | 4 |  |  | | | | |
| 和钟老师讨论低频项研究进展 | |
| 讨论班学习 | |
| 星期五 | | 生成BSS算法实验数据（zipf分布，参数是0.5,1,1.5,2,2.5,3） | | 吴大衍 |  |  |  | | | | |
|  | |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 3 月14日－2015 年3月18日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 讨论班报告 | | 完成 | | 1 | 合并3.2代码 | | 完成 | 周一 | 钟老师 | 吴大衍 |
| 2 | 完善3.2全文检索模块 | | 完成 | | 2 | 调研多磁盘并发写技术 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 3 |  | |  | | 3 | 系统学习AQP中sampling相关技术 | | 完成 | 周五 | 吴大衍 | 吴大衍 |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 完善3.2全文检索模块，使得3.2支持select \_score | | 吴大衍 | 1 | AQP的研究时间长，方向很广，不好切入和深入 | 找到AQP方向的综述论文或者书籍，系统学习相关算法并跟踪最近的研究动向，找到具体的研究兴趣点 | | | | |
| 查找AQP相关文献 | |
| 星期二 | | 阅读蓄水池抽样算法、concise sample算法、count sample算法相关论文 | | 吴大衍 | 2 |  |  | | | | |
| 星期三 | | 阅读等高直方图算法、等宽直方图算法、v-optimal算法相关论文 | | 吴大衍 | 3 |  |  | | | | |
| 星期四 | | 阅读哈尔小波变换算法相关论文 | | 组内成员  吴大衍 | 4 |  |  | | | | |
| 制作汇报ppt | |
| 讨论班汇报 | |
| 星期五 | | 整理论文材料、总结论文算法、制定研究计划 | | 吴大衍 |  |  |  | | | | |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 3 月7日－2015 年3月11日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 调试3.2代码 | | 完成 | | 1 | 准备讨论班报告 | | 完成 | 周三 | 吴大衍 | 吴大衍 |
| 2 |  | |  | | 2 | 整理3.2代码 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 3 |  | |  | | 3 | 调研中间件相关技术 | |  |  | 钟老师 | 吴大衍 |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 部署3.2系统 | | 吴大衍 | 1 | Seald启动失败 | 1. 环境变量设置错误 2. 元数据和配置文件读取有差别，修改fe端代码后重新编译。 | | | | |
| 星期二 | | 调试3.2代码 | | 吴大衍 | 2 | Fe端无法识别match | 3.2中fe的Expr增加了抽象函数需要match实现 | | | | |
| 星期三 | | 调试3.2fe端代码 | | 吴大衍 | 3 | 3.2中注册slot的方法和3.1不同 | 重新实现\_score列的注册 | | | | |
| 星期四 | | 3.2fe端代码调试通过 | | 吴大衍 | 4 | \_score列传到be端错位 | 手动设置\_score列的位置 | | | | |
| 星期五 | | 3.2整体调试通过 | | 吴大衍 |  |  | | | | | |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 2 月29 日－2015 年3月4日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 3.1全文检索代码移植到3.2上 | | 完成 | | 1 | 内部测试3.2全文检索模块查询性能并优化 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 编译3.2代码 | | 完成 | |  |  | |  |  |  |  |
| 3 |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 整理数据流、AQP相关论文；全文检索代码加上注释 | | 吴大衍 | 1 | 由于前段sql支持的语法较多，导致生成的sql-parser文件中的函数超过65535bytes的限制 | 精简sql-parser.y | | | | |
| 星期二 | | 测试DBroker-3.1全文检索模块功能实现以及查询性能 | | 吴大衍 | 2 | 3.2在3.1的基础上做了修改，比如有些函数名发生变化，导致直接patch会失败 | 手动patch | | | | |
| 星期三 | | 基于DBroker-3.1制作补丁并给be,se端打上补丁 | | 吴大衍 | 3 | 直接co下来的代码make失败 | 拷贝缺少的文件并修改环境变量 | | | | |
| 星期四 | | 合并thrift部分代码、解决冲突并编译 | | 钟老师  林蝉  吴大衍 |  |  |  | | | | |
| 搭建DBroker-3.2fe端调试环境 | |
| 组会讨论中间件开发问题 | |  |  |  |  | | | | |
| 星期五 | | 合并fe端代码 | | 吴大衍 |  |  | | | | | |
| 编译DBrokerSE-3.2 | |
|  | |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 1 月25 日－2015 年1月29日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 全文检索多域单查询开发（支持best\_fields和most\_fields两种搜索类型） | | 完成 | | 1 |  | |  |  |  |  |
| 2 | 理论推导基于count-min的低频项挖掘算法的空间复杂度 | | 完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 组会 | | 李老师  钟老师  吴大衍 |  | 由于最低频项本身频率很小，为保证精度，空间复杂度为nlog(n)（n为数据项的取值值域），而这个空间复杂度已经退化到了最原始的直接对每个数进行计数 | 虽然空间复杂度较高，使用count-min算法的优势在于更新时间，所以可以考虑推广到滑动窗口这类对更新时间要求较高的模型上 | | | | |
| 阅读elasticsearch源码并做汇报 | | 1 |
| 星期二 | | 全文检索多域单查询开发 | | 吴大衍 |  |  |  | | | | |
| 星期三 | | 全文检索多域单查询开发 | | 吴大衍 |  |  |  | | | | |
| 室年会 | |
| 星期四 | | 理论推导基于count-min算法的低频项挖掘空间复杂度 | | 吴大衍 |  |  |  | | | | |
| 陪王卓去医院 | |
| 星期五 | | 阅读PLA算法相关论文 | | 吴大衍 |  |  |  | | | | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 1 月18 日－2015 年1月22日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 全文检索单field单query开发 | | 完成 | | 1 | 全文检索加入boost | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 整理Elasticsearch相关资料并制作汇报ppt | | 完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 全文检索开发，查询可以返回结果，但是查询结果中没有\_score列 | | 吴大衍 | 1 | Local idf 和 global idf不统一导致打分不准确 | 可以考虑引入随机算法，利用sketch记录每个term的文档数 | | | | |
| 星期二 | | 全文检索开发，当查询有order by \_score时，向从se中读取的列中加入\_score列（表中没有\_score列，查询过程中视sql语句动态添加） | | 吴大衍 | 2 | 如何定义elasticsearch复杂全文检索的语法 | 借鉴crate并结合DBroker自身的语法解析来确定 | | | | |
| 星期三 | | 全文检索开发，当查询有order by \_score时，返回结果返回每条记录的\_score | | 吴大衍 |  |  |  | | | | |
| 星期四 | | 全文检索开发，当查询select选项中有\_score选项时，返回结果返回每条记录的\_score；解决select count(\*) 无法显示结果的问题 | | 组内成员  吴大衍 |  |  |  | | | | |
| 讨论班 | |
| 星期五 | | 整理elasticsearch相关资料并制作汇报ppt | | 吴大衍 |  |  |  | | | | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 1 月11 日－2015 年1月15日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 阅读crate以及elasticsearch全文搜索部分源码 | | 完成 | | 1 | 全文检索功能开发 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 基于DBroker3.1进行全文检索功能的开发 | | 未完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 调研elasticsearch基本概念 | | 钟老师  吴大衍 | 1 | Crate的match语法较为复杂，直观上不太好移植到DBroker3.1中 | 首先实现最基本的match语法，整体流程跑通后再加入复杂语法 | | | | |
| 和钟老师讨论全文检索设计方案 | |
| 星期二 | | 部署elasticsearch调试环境 | | 吴大衍 | 2 | Sql中同时存在match和search时该如何处理 | 先和各位老师讨论是否有这种需求，如果有再讨论具体方案 | | | | |
| 阅读elasticsearch源码，主要包括代码流程以及boost如何对查询打分产生影响 | |
|  | |
| 星期三 | | 阅读crate源码，主要了解crate如何将sql转化成elasticsearch支持的json格式查询 | | 吴大衍 |  |  |  | | | | |
|  | |  | |
| 星期四 | | 搭建fe端调试环境并跟踪代码了解fe端对sql的解析流程以及如何取得谓词 | | 小组成员  吴大衍 |  |  |  | | | | |
| 组会 | |
| 星期五 | | 问林蝉语法文件修改方法 | | 林蝉  吴大衍 |  |  |  | | | | |
| 代码开发使得系统支持最基本的match语法，并使得field以及query值传到se端 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 1 月 4 日－2015 年1月8日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 非时间分区列排序的开发与调试 | | 完成 | | 1 | 全文检索调研与设计 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 搭建DBroker3.1测试环境 | | 完成 | |  |  | |  |  |  |  |
| 3 | 调研ElasticSearch打分机制 | | 未完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 代码开发 | | 吴大衍 | 1 | 查询中途cancel，系统不再支持下一条查询 | Fetch部分没有正确cancel，cancel时需要同时cancel掉search和fetch部分 | | | | |
| 星期二 | | 代码开发 | | 吴大衍 | 2 | 对于limit数目较大的查询无法返回结果 | Limit较大时，可能单个节点上的总记录数小于limit数，此时需要做特殊处理 | | | | |
| 调试代码使得查询有结果出现 | |
|  | |
| 星期三 | | 组会部署下一步工作 | | 钟老师 | 3 | 对于limit的较大的查询，代码search部分（主要包括lucene排序）用时不多，但是fetch部分用时较多导致查询总用时增多 | 目前找到几个原因：fetch部分用时fetch的调度策略有问题；fetch本身用时较多。具体原因还需要跟踪调试 | | | | |
|  | | 调试代码使得查询支持limit数目较大的情况 | | 吴大衍 |
| 星期四 | | 调试代码使得查询在cancel掉后可以继续支持下一条查询 | | 吴大衍 |  |  |  | | | | |
| 优化代码fetch部分的调度策略 | |
|  | |
| 星期五 | | 搭建DBroker3.1测试环境 | | 吴大衍 |  |  |  | | | | |
| 调研ElasticSearch打分机制 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 12 月 28 日－2015 年12月31日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 非分区列排序代码开发 | | 基本完成 | | 1 | 完成非分区列排序代码开发 | | 完成 | 周三 | 钟老师 | 吴大衍 |
|  |  | |  | | 2 | 调试+完善开发文档 | | 基本完成 |  | 钟老师 | 吴大衍 |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 益园志愿者活动 | | 志愿者小组  吴大衍 | 1 | fetch部分worker的逻辑和原来的querytaskworker不一样 | 保留原来的调度逻辑，重新实现每个worker的工作逻辑 | | | | |
| 星期二 | | 完成非分区列排序中search部分的开发 | | 吴大衍 | 2 | 排序分为search和fetch两部分，而且是异步实现 | 重新实现一个broker，作为中转站连接两部分 | | | | |
|  | |
|  | |
| 星期三 | | 完成非分区列排序中broker部分的开发 | | 吴大衍 | 3 |  |  | | | | |
|  | |  | |  |  |  |  | | | | |
| 星期四 | | 完成非分区列排序中fetch部分的开发 | | 吴大衍 | 4 |  |  | | | | |
|  | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 12 月 21 日－2015 年12月25日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 完成以图搜图代码调试工作并整理上传gitlab | | 完成 | | 1 | 非分区列排序代码开发 | | 基本完成 | 周四 | 钟老师 | 吴大衍 |
| 2 | 阅读se端按照时间分区列排序的代码 | | 完成 | |  |  | |  |  |  |  |
| 3 | 理解非分区列排序方案并搭建代码框架 | | 未完成 | |  |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 整理以图搜图代码并上传gitlab | | 吴大衍 | 1 |  |  | | | | |
| 星期二 | | 在结果表中加入score属性并测试按照score排序的性能 | | 钟老师  吴大衍 | 2 |  |  | | | | |
| 钟老师讲解非分区列排序方案 | |
|  | |
| 星期三 | | 阅读时间分区列排序代码 | | 志愿者小组  吴大衍 | 3 |  |  | | | | |
| 益园开会听取志愿者活动安排 | |
|  | | 修改prezi | |  |  |  |  | | | | |
| 星期四 | | 阅读时间分区列排序代码 | | 钟老师  吴大衍 | 4 |  |  | | | | |
| 理解非分区列排序方案并和钟老师讨论 | |
| 搭建代码框架 | |
|  | |  |
| 星期五 | | 逐网-2015活动志愿者 | | 志愿者小组  吴大衍 | 5 |  |  | | | | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 12 月 14 日－2015 年12月18日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | fe+be+se代码内部调试 | | 完成 | | 1 | 整理代码+上传gitlab | | 完成 | 周一 | 吴大衍 | 吴大衍 |
| 2 | 代码冲突合并+上传gitlab | | 基本完成 | | 2 | 修改视频 | |  |  | 李宇哲 | 吴大衍 |
| 3 | 制作汇报视频第一版 | | 完成 | |  |  | |  |  |  |  |
|  |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 整体调试以图搜图代码 | | 吴大衍 | 1 | 由于前期开发没有及时更新到gitlab上导致和gitlab上的代码冲突太多，一次性合并冲突费时费力还容易出现各种bug | 以后在开发过程中及时更新gitlab上的代码，缩短冲突合并周期 | | | | |
| 整理代码 | |
| 代码冲突合并 | |
| 星期二 | | 代码冲突合并 | | 吴大衍 | 2 |  |  | | | | |
| 重新编译调试冲突合并后的代码 | |
|  | |
|  | |
| 星期三 | | 代码上传gtilab | | 视频制作小组 | 3 |  |  | | | | |
| 确定视频制作分工+时间节点 | |
| 熟悉汇报文档 | |
|  | | 小组安装破解prezi+中文字体 | |  |  |  |  | | | | |
| 星期四 | | 讲解prezi基本用法 | | 视频制作小组 | 4 |  |  | | | | |
| 确定prezi模板 | |
| 寻找prezi素材 | |
| 制作prezi整体框架+汇报收尾部分 | |  |
| 星期五 | | 合并prezi各部分制作 | | 王老师  岳老师  视频制作小组 | 5 |  |  | | | | |
| 根据王老师和岳老师意见修改prezi，录制prezi视频，配合录音生成第一版汇报视频 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 12 月 07 日－2015 年12月11日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 阅读fe端代码 | | 完成 | | 1 | 整体调试 | | 完成 | 周二 | 吴大衍 | 吴大衍 |
| 2 | 修改fe端和be端代码 | | 完成 | | 2 | 准备和一室联调 | |  |  | 钟老师 | 吴大衍 |
| 3 | 调试se端代码 | | 完成 | |  |  | |  |  |  |  |
| 4 | 整体调试 | | 基本完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 阅读fe端代码 | | 吴大衍 | 1 | Catalog在load时将表的某一属性写到了thrift结构中，但是在catalog中却看不到这一项 | Be端代码版本不对，替换be端代码，重新编译后运行即可 | | | | |
| 了解配置文件的具体格式 | |
|  | |
| 星期二 | | 调试se端代码 | | 吴大衍 | 2 | Se端查询时，buffer中已经有数据，但是查询时报超时错误 | 缓冲区出现死锁，job完成后统一释放全部锁即可 | | | | |
| 构思fe端代码修改档案 | |
| 搭建fe端代码调试环境 | |
|  | |
| 星期三 | | 搭建fe端代码调试环境 | | 吴大衍 | 3 | 如何使得查询发往全部节点 | 配置文件中将每个表lucene索引文件的分区信息置为空即可 | | | | |
| 修改fe端代码 | |
| 了解cmake以及thrift工作原理 | |
| 星期四 | | 修改thrift并重新编译项目 | | 吴大衍 | 4 |  |  | | | | |
| 修改fe端和be端代码 | |
|  | |
| 星期五 | | 听取2015年大数据技术大会报告 | | 李老师  钟老师  吴大衍 | 5 |  |  | | | | |
| 整体调试 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 11 月 30 日－2015 年12月04日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 阅读se端代码 | | 完成 | | 1 | 阅读fe端代码并设计修改方案 | | 完成 | 周三 | 钟老师 | 吴大衍 |
| 2 | 设计se端接口调用方案并修改代码 | | 完成 | | 2 | 修改fe端代码 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 3 | 搭建mpp-engine | | 完成 | | 3 | 开始整体调试 | | 开始 | 下周 | 钟老师 | 吴大衍 |
| 4 | 阅读fe端代码 | | 未完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 开组会 | | 小组人员  吴大衍 | 1 | 第二版方案不需要存储图片元数据，和se端查询lucene索引的功能不太一样 | 复用se端代码的整体逻辑，按照整体逻辑加入图像检索功能。原来需要查询lucene索引部分的代码换成调用以图搜图接口，一个queryworker的基本单位换成queryjob，最后结果上传部分直接从接口返回结果中读取等。 | | | | |
| 阅读se端代码查询解析部分代码 | |
|  | |
| 星期二 | | 设计第一版修改方案 | | 吴大衍 | 2 | se端有些参数的具体含义不明 | 搭建测试环境跟踪测试 | | | | |
| 按照修改方案修改se端代码 | |
|  | |
|  | |
| 星期三 | | 调试se端修改代码 | | 吴大衍 | 3 | mpp-engine首次搭建，对搭建流程不明 | 理清启动进程以及启动方法并结合相应的日志定位问题 | | | | |
|  | |
|  | |
| 星期四 | | 和钟老师讨论修改方案 | | 钟老师  吴大衍 | 4 |  |  | | | | |
| 设计第二版修改方案 | |
| 根据修改方案整理se端实现思路 | |
| 星期五 | | 根据第二版修改方案修改se端代码 | | 高开放  吴大衍 | 5 |  |  | | | | |
| 搭建mpp-engine | |
|  | | 阅读fe端代码 | |  |  |  |  | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 11 月 23 日－2015 年11月27日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 和一室讨论图像接口事宜并准备相关材料 | | 完成 | | 1 | 阅读SE端代码并设计以图搜图接口调用方案 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 搭建分布式的rocketmq环境并做消息收发实验 | | 完成 | |  |  | |  |  |  |  |
| 3 | 阅读SE端代码并构思以图搜图接口调用方案 | | 未完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 开组会 | | 钟老师  吴大衍 | 1 | Rocketmq中的消息如何不被多次消费并且如何保证每个consumer消费的消息尽可能平均 | 将Consumer设置在同一个group中；尽量保证同一个topic的消息队列消息数量尽可能平均 | | | | |
| 与一室讨论图像场景和以图搜图的接口对接问题 | |
| 写测试问题报告 | |
| 星期二 | | 阅读rocketmq官方文档，了解rocketmq工作原理 | | 钟老师  吴大衍 | 2 | 如何将以图搜图部分嵌入到现有的SE框架中，因为以图搜图不需要我们自己管理索引，只需要调用相应的服务即可 | 定义新的workerpool以及相应的worker，并让两种worker同时运行，根据job的不同选择将task分配给对应的worker | | | | |
| 阅读数据加载以及消息队列消费部分的代码样例 | |
| 整理一室需要的有关mq的资料 | |
|  | |
| 星期三 | | 搭建分布式的rocketmq环境 | | 吴大衍 | 3 |  |  | | | | |
| 基于rocketmq做消息收发实验 | |
| 修改parquet调用接口以支持parquet块大小的修改 | |
| 星期四 | | 阅读SE端代码 | | 张金超  吴大衍 | 4 |  |  | | | | |
| 听取金超师兄的讨论会报告 | |
| 学习使用gitlab | |
| 星期五 | | 阅读SE端代码 | | 吴大衍 | 5 |  |  | | | | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 11 月 16 日－2015 年11月20日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 数据仓库测试 | | 完成 | | 1 | 钟老师安排工作 | |  |  | 钟老师 | 吴大衍 |
| 2 | 阅读2015年SIGMOD关于数据流算法Persistent Sketch的论文 | | 基本完成 | |  | 完成论文Persistent Sketch的阅读 | | 完成 | 周五 | 吴大衍 | 吴大衍 |
| 3 | 根据数据仓库测试情况进一步阅读impala关于内存使用部分的代码 | | 基本完成 | |  | 根据测试情况进一步阅读impala中关于内存和回写磁盘部分的代码 | | 完成 | 周五 | 吴大衍 | 吴大衍 |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 根据重新生成的zb\_test表，重新测试前十条用例在扩展数据上的查询时间 | | 钟老师  吴大衍 | 1 | 测试用例中有一条测试在扩展数据集上的测试时间超过5小时，经观察该查询的中间结果很大，时间大部分都在中间结果的磁盘IO上 | 可以考虑多线程并发读写中间结果或者将中间结果回写hdfs(目前impala中间结果的回写貌似不支持hdfs，具体仍需调研) | | | | |
| 增加中间结果回写磁盘目录，重新执行首轮查询报错的测试用例 | |
| 完善测试文档，重新测试结果不全的测试用例 | |
| 星期二 | | 测试并发查询性能（5条查询，每条查询并发执行一次以及每条查询并发执行三次） | | 钟老师  吴大衍 | 2 | 并发查询会报内存溢出错误，虽然高版本的impala支持内存溢出时的磁盘回写，但是在并发查询中如果不限制每条查询的内存使用就会导致多条查询内存溢出从而报错 | 将内存使用较大的查询（特别是中间结果很大的查询）限制内存使用，这样会保证查询的顺利执行，但是查询速度会下降，因为原本不需要回写磁盘的查询现在需要进行磁盘的回写 | | | | |
| 完善测试文档 | |
|  | |
|  | |
| 星期三 | | 容错测试（杀进程，杀namenode，umount磁盘） | | 钟老师  吴大衍 | 3 | 查询过程中杀死某个host的impalad进程会报连接错误，因为此时查询计划已经下发到每个host上，此时只需要重新执行一遍相同的查询即可；查询过程中杀死namenode进程会报找不到hdfs块，解决办法可以考虑双机热备；因为磁盘被datanode进程占用，所以即使hdfs对磁盘没有读写操作，umount时还是会提示设备忙碌，测试时的方法是直接杀掉一个点的datanode进程。 |  | | | | |
| 整理测试文档并上传 | |
| 测试文档脱密 | |
| 星期四 | | 阅读2015年sigmod关于数据流算法Persistent Sketch的论文 | | 吴大衍 | 4 |  |  | | | | |
|  | |
|  | |
| 星期五 | | 根据测试结果进一步阅读原生impala中关于内存使用部分的代码 | | 吴大衍 | 5 |  |  | | | | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 11 月 9 日－2015 年11月13日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 测试数据加载完成 | | 完成 | | 1 | 测试并发查询以及系统容错性能 | | 完成 | 周二 | 钟老师 | 吴大衍 |
| 2 | 测试用例查询时间 | | 基本完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 完成一二号全量数据生成并上传hdfs | | 钟老师  王琳  吴大衍 | 1 | csv哈希映射代码会在文件某一行停住，CPU还在运行，但是没有写磁盘操作，经问题定位是csvreader代码有问题 | 用bufferedreader替换csvreader | | | | |
| 并发生成3-31号的全量数据 | |
| 生成各个小表数据以及comm\_pkg表中一小时数据 | |
| 星期二 | | 完成初步测试 | | 钟老师  王琳  吴大衍 | 2 | 当查询的中间结果太大时，impala会将部分中间结果回写磁盘（/tmp/impala-scratch目录），如果这个目录写满查询会报错 | 可以考虑将中间结果写入多个磁盘，一来可以并发读写中间结果，二来可以扩容 | | | | |
| 中心汇报测试进展 | |
| 修改comm\_pkg表结构修改parquet数据生成代码 | |
| 分发comm\_pkg原始数据 | |
| 星期三 | | 回所汇报测试进展 | | 刘敬  林蝉  吴大衍 | 3 | 两次相同查询的中间结果会有不同，但是最终结果相同 | 这跟每个节点执行的查询计划有关，每个节点不同查询读取的数据每次都不相同，导致聚合的结果都会有差别 | | | | |
| 中心调试csv文件哈希分发代码并按照哈希切分文件 | |
| 合并数据扩展代码和lucene索引生成代码 | |
| 星期四 | | 切分哈希分发后的大文件 | | 钟老师  吴大衍 | 4 |  |  | | | | |
| 上传comm\_pkg一天的数据 | |
| 由于代码原因，comm\_pkg表有两个节点的数据不准确，修改代码后重新生成这两个点的数据 | |
| 星期五 | | 撰写模拟数据生成方案 | | 钟老师  吴大衍 | 5 |  |  | | | | |
| 重新生成zb\_test表数据（fl字段值域不符合文档要求） | |
| 上传comm\_pkg表余下两节点的数据 | |
|  | |  | |  |  |  |  | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 11 月 2 日－2015 年11月6日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 中心加载数据 | | 代码修改完毕，目前正在服务器上生成数据（包括原始数据和模拟数据） | | 1 | 完成数据仓库的测试 | | 完成 | 周三 | 钟老师 | 吴大衍 |
| 2 | 测试用例查询时间 | | 未完成 | |  |  | |  |  |  |  |
| 3 |  | |  | |  |  | |  |  |  |  |
| 4 |  | |  | |  |  | |  |  |  |  |
| 5 |  | |  | |  |  | |  |  |  |  |
| 6 |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 测试parquet+snappy压缩比 | | 钟老师  王琳  吴大衍 | 1 | Parquet存储timestamp类型的数据解析会出错 | 用bigint类型代替timestamp | | | | |
| 根据压缩比和每天的数据量确定二级分区个数 | |
| 和钟老师分析测试用例，确定测试方案 | |
| 星期二 | | 和钟老师讨论数据生成方案 | | 钟老师  王琳  吴大衍 | 2 | 如何确定每个文件的大小实现更高的查询效率 | 实验对比每个parquet文件为600兆和100兆时的查询效率，发现100兆的查询时间是600兆的1/3，并且600兆的文件查询结束后会警告文件跨块，所以最终确定一个parquet文件在100兆左右，约120万条记录（snappy压缩后） | | | | |
| 开发模拟数据生成程序 | |
| 配合王琳开发原始数据加载程序 | |
|  | |
| 星期三 | | 根据测试用例简化数据生成方案 | | 钟老师  王琳  吴大衍 | 3 | 表的字段类型如何确定才能在保证出错率低的情况下达到高的压缩比 | 经过多次实验，尽量用整型代替字符型，但是要保证出错率要在尽量低的范围内 | | | | |
| 修改shell支持基于时间的一级分区 | |
| 配合王琳开发put程序 | |
| 生成一到十二号的原始数据 | |
| 生成一到六号每天1T的模拟数据 | |
| 星期四 | | 利用29个节点同时生成第一天的全部数据（8至9T） | | 钟老师  王琳  吴大衍 | 4 | 每天的源数据都不相同，甚至列数都会不同，导致程序解析出错 | 修改原程序以适应各种情况 | | | | |
| 利用生成的1T数据（1501）亿测试一条join+group by的测试用例 | |
| 利用29个节点，每个节点模拟生成一天数据 | |
| 星期五 | | 修改put程序，原有程序有问题，会导致脏数据上传 | | 钟老师  王琳  吴大衍 | 5 | 测试过程中会出现报错提示某个parquet文件元数据无效 | 删除脏数据并修改put程序使得脏数据不会上传 | | | | |
| 删除hdfs上的脏数据 | |
| 开发parquet的schema自动解析程序 | |
| 测试千亿join+group by用例 | |
| 利用29个节点同时生成第二天的全部数据用来测试 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 10 月 26 日－2015 年10月30日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | parquet格式（snappy压缩和不压缩）与rcfile格式加载、存储以及查询性能比较 | | 基本完成 | | 1 | 配合王琳加载数据 | |  |  |  |  |
| 2 | 完成第二轮原生impala测试结果的分析 | | 完成 | | 2 | 中心测试数据仓库查询性能 | |  |  |  |  |
| 3 | 完成写parquet格式文件代码 | | 完成 | |  |  | |  |  |  |  |
| 4 | 阅读fe端代码找出impala分区部分代码 | | 完成 | |  |  | |  |  |  |  |
| 5 | 修改impala-shell.py以支持hash分区 | | 完成 | |  |  | |  |  |  |  |
| 6 | 测试parquet格式文件的压缩比（snappy算法） | | 完成 | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 调研hive的分区规则 | | 吴大衍 | 1 | Hive不支持hash分区、范围分区等，仅支持精确分区 | 考虑借鉴hive中桶的概念 | | | | |
| 分析第二轮impala测试结果 | |
| 测试parquet和rcfile存储格式的查询性能 | |
| 星期二 | | 完成parquet、rcfile和orcfile格式文件的加载 | | 钟老师  吴大衍 | 2 | 不能简单替换pzid，因为取出分区后还需要此字段筛选 | 加入一个hash字段 | | | | |
| 测试parquet存储格式和rcfile存储格式的查询性能 | |
| 看fe端确定分区的代码 | |
| 和钟老师讨论hash分区实现 | |
| 星期三 | | 修改impala-shell.py代码在sql语句中加入hash查询字段以支持hash分区 | | 王琳  吴大衍 | 3 | 网上对parquet文件如何生成的讲解太少，主要集中在原理讲解 | 借鉴spark生成parquet文件的方法 | | | | |
| 中心和王琳商量数据加载的方案以及parquet格式文件如何写 | |
| 完成parquet格式文件的写开发 | |
|  | |
| 星期四 | | 完善写parquet格式文件代码 | | 吴大衍 | 4 | 无 | 无 | | | | |
| 阅读parquet格式相关论文，了解parquet格式中definition以及repitition的具体含义 | |
|  | |
| 星期五 | | 用中心测试用例测试shell是否运行正确 | | 王琳  吴大衍 | 5 | 王琳读csv格式文件的速度太慢 | 修改csv格式文件的解析代码 | | | | |
| 跟王琳讲解parquet中schema如何生成，并尝试加载数据 | |
| 测试snappy+parquet的压缩比 | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 10 月 19 日－2015 年10月23日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 第二轮原生impala数据加载（包括大表240亿数据以及parquet格式以及orcfile格式的实验数据加载） | | 完成 | | 1 | 完成第二轮测试结果分析 | | 完成 | 周一 | 吴大衍 | 吴大衍 |
| 2 | 第二轮原生impala测试（小表10万，100万） | | 完成 | | 2 | 完成parquet格式和orcfile格式的测试，并与rcfile对比 | | 完成 | 周二 | 吴大衍 | 吴大衍 |
| 3 | 阅读原生impala内存管理部分代码 | | 基本完成 | | 3 | 调研hive的分区规则（是否支持模糊分区或者实验测试百万级分区的查询性能） | | 完成 | 周三 | 吴大衍 | 吴大衍 |
| 4 | 中心听取测试要求以及测试样例 | | 完成 | | 4 |  | |  |  |  |  |
| 5 | 分析第二轮测试结果 | | 未完成 | | 5 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 编写第二轮原生impala测试数据加载代码 | | 吴大衍 | 1 | 无 | 无 | | | | |
|  | |
|  | |
| 星期二 | | 阅读原生impala内存管理部分代码 | | 吴大衍 | 2 | 内存使用峰值远没有达到内存限制，却存在溢出磁盘操作 | 可能是因为分配给每个node的buffer是有限制的 | | | | |
|  | |
|  | |
|  | |
| 星期三 | | 加载parquet格式以及orcfile格式的测试数据 | | 吴大衍 | 3 | 没有找到写Parquet以及orcfile格式文件的方法 | 采用insert+select的方法加载parquet和orcfile格式的数据，各1200万 | | | | |
| 编写第二轮impala测试的测试样例 | |
|  | |
|  | |
| 星期四 | | 完成第二轮原生impala测试数据的加载 | | 吴大衍 | 4 | 无 | 无 | | | | |
| 编写第二轮原生impala测试程序 | |
|  | |
| 星期五 | | 完成10万小表的测试并分析结果 | | 吴大衍 | 5 | 中心测试要求100万不同的pzid值，如果不对pzid分区就需要扫描当天的所有数据，否则就要进行百万级的分区 | 调研hive是否支持模糊分区（多个列值映射到一个分区中，如果不支持就实验测试百万级分区的查询性能，因为如果分区太多会导致每个分区的文件太碎，效果可能不如扫描全表） | | | | |
| 中心听取测试方案以及测试样例 | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 10 月 12 日－2015 年10月16日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 阅读impala2.2be端代码 | | 基本完成了对hdfs-scan-node,hash-join-node(partitioned-hash-join-node),Aggregation-node(partitioned-aggregation-node)算子代码的阅读 | | 1 | 完成impala第二轮测试数据的生成并完成第二轮测试并分析测试结果 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 完成小数据集上原生impala性能测试 | | 完成 | | 2 | 进一步阅读代码了解impala的内存管理机制 | | 完成 | 周五 | 吴大衍 | 吴大衍 |
| 3 | 生成第二轮测试数据 | | 未完成 | | 3 |  | |  |  |  |  |
| 4 |  | |  | |  |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 阅读be端代码的hdfs-scan-node和hash-join-node的部分 | | 吴大衍 | 1 | 无 | 无 | | | | |
| 了解be端整个planfragment的执行流程 | |
|  | |
| 星期二 | | 阅读be端aggregation-node算子 | | 吴大衍 | 2 | Group by+join查询的时间较长，查询后发现瓶颈出现在datastreamsender上 | 由于是在shell上进行小数据集上测试，测试结果全部输出到shell中，导致coordinator节点拉去数据的速度较慢，可以考虑将结果重定位至文件或者jdbc测试 | | | | |
| 益园测试小数据上的原生impala性能（大表2000万、小表10万） | |
|  | |
|  | |
| 星期三 | | 阅读impala官方文档关于impala回写磁盘的介绍 | | 李老师  吴大衍 | 3 | 何时使用partitioned算子？何时进行回写磁盘操作？ | Impala-2.2默认的hash-join以及aggregation操作都是走partitioned-hash-join-node以及partitioned-aggregation-node,代码中有相应的bool值设定；  Impala有一套block管理方案，当block对应的buffer超过限制时就会选择一个partiition回写磁盘 | | | | |
| 代码中定位磁盘回写部分，并阅读be端的partitioned-hash-join-node以及partitioned-aggregation-node部分的代码 | |
|  | |
|  | |
| 星期四 | | 阅读partitioned-hash-join-node以及partitioned-aggregation-node部分的代码以及相应的内存管理代码 | | 吴大衍 | 4 | 为什么profile中显示block的内存使用峰值远远不到limit值，但还是会有回写磁盘操作？ | 还未弄清楚，准备通过profile中的具体变量精确定位到代码中的相应位置 | | | | |
| 去华严和加载组讨论第二轮测试数据加载的方案并阅读加载部分代码方便以后自己加载 | |
|  | |
| 星期五 | | 阅读be端有关哈希表操作的部分代码 | | 吴大衍 | 5 | 为什么小表join 列只有25个不同值，但是hash buckets值却远远超过25 | 由于aggregation操作和hash join操作都需要用到hash表，所以hash表的初始值选作刚好大于行数的2的幂 | | | | |
| 修改数据加载代码，加快数据加载速度 | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 10 月 8 日－2015 年 10 月10日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 第一轮原生impala性能测试 | | 完成 | | 1 | 改变impala数据分区方式，重新测试 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 阅读impala 2.2be端代码 | | 未完成 | | 2 | 阅读be端代码定位查询慢的原因 | | 完成 | 周五 | 吴大衍 | 吴大衍 |
| 3 |  | |  | | 3 |  | |  |  |  |  |
| 4 |  | |  | | 4 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期四 | | 完成第一轮原生impala性能测试 | | 钟老师  吴大衍 | 3 | 小表c\_pzid的值太少（1-25）导致和大表join后的结果过多,大表每一行平均要跟1万行的小表行进行join，这也跟实际情况不符 | 调整小表中c\_pzid的范围为1-10万 | | | | |
| 和钟老师讨论优化方案 | |
|  | |
|  | |
| 星期五 | | 了解impala2.2的代码整体结构，包括fe端代码和be端代码的执行流程 | | 吴大衍 | 4 | 无 | 无 | | | | |
|  | |
|  | |
| 星期六 | | 阅读be端hash-join-node和hdfs-scan-node相关代码 | | 吴大衍 | 5 | 代码中只有主线程在执行hash-join，但是代码中显示开启了多线程进行hdfs-scan，可能这也是为什么监控显示只有一个线程满负荷执行但是其他线程几乎没有执行的原因 | 结合第二轮测试的监控情况以及代码找出原因 | | | | |
|  | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 9 月 21 日－2015 年 9 月 25日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 调整原生impala测试用例 | | 完成 | | 1 | 定位count(\*)+join查询慢的原因 | | 完成 | 周三 | 吴大衍 | 吴大衍 |
| 2 | 测试原生impala | | 未完成（count(\*)+join查询太慢） | | 2 |  | |  |  |  |  |
| 3 | 了解impala查询计划，找到count(\*)+join查询慢的具体原因 | | 查询计划基本了解，原因还在定位 | | 3 |  | |  |  |  |  |
| 4 | 阅读space saving算法对随机生成数据和zipf数据上进行频繁项以及top-k查找实验的相关论文 | | 未完成 | | 4 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 阅读space saving算法对随机生成数据和zipf数据上进行频繁项以及top-k查找实验的相关论文 | | 吴大衍 | 1 | 无 | 无 | | | | |
| 开组会 | |
|  | |
| 星期二 | | 和钟老师讨论并修改测试原生impala用例 | | 钟老师  吴大衍 | 2 | 原生的impala测试用例中的查询都带有精确条件，缺乏普遍性。 | 添加分跨度的只包含count(\*)+join的查询语句 | | | | |
|  | |
|  | |
|  | |
| 星期三 | | 编写原生impala的jdbc测试代码 | | 吴大衍 | 3 | 通过shell测试impala的查询性能需要保持shell打开，但是关闭shell就会结束查询，但是往往一条查询语句的执行时间会很长，不可能等着它执行完毕 | 编写impala的jdbc测试代码，并将其在后台运行 | | | | |
|  | |
|  | |
|  | |
| 星期四 | | 测试原生impala的count(\*)+join查询性能 | | 吴大衍 | 4 | Select count(\*) from t\_ybrz join t\_pzid on (t\_ybrz.c\_pzid = t\_pzid.c\_pzid) where c\_capturetime=1441814400;这条查询语句的查询响应时间为15h27min，查询时间太慢 | 暂时未找到具体原因 | | | | |
|  | |
|  | |
| 星期五 | | 了解impala的查询计划 | | 吴大衍 | 5 | 查询计划表明hash join函数中的probe部分占用了百分之九十的时间，但是hash join函数的probe部分可以分为三个部分：读，定位，写中间结果，究竟是哪个环节慢了呢？ | 首先定位的时间复杂度是在O（n）并且都是在内存中完成，应该不是瓶颈所在；读和写哪个是瓶颈还需要进一步观察硬盘的IO情况，并阅读be端代码了解具体流程 | | | | |
| 去益园查看impala的具体查询计划和日志定位查询瓶颈 | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 9 月 14 日－2015 年 9 月 18日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 设计Dbroker和原生impala检索测试用例 | | 完成 | | 1 | 完成原生impala的检索测试 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 2 | 测试原生impala检索性能 | | 基本完成 | | 2 | 完成Dbroker和原生impala的并发查询性能测试 | | 完成 | 周五 | 钟老师 | 吴大衍 |
| 3 | 准备讨论班材料 | | 完成 | | 3 | 设计基于count-min的低频项挖掘算法实验方案 | | 基本完成 |  | 吴大衍 | 吴大衍 |
| 4 |  | |  | | 4 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 开组会 | | 韩笑  吴大衍 | 1 | Dbroker数据没有分区；  Dborker中t\_ybrz表中的c\_pzid列没有建索引。 | 和加载协商解决 | | | | |
| 去益园了解系统部署以及数据加载情况 | |
| 结合tpc-h设计Dbroker检索用例 | |
| 星期二 | | 准备讨论班材料 | | 吴大衍 | 2 | 原生impala数据存在问题；  Dbroker进行联合查询时会报内存溢出的错误。 | 协调加载解决原生impala的数据问题；  改变sql中表的join顺序，内存溢出问题即可解决 | | | | |
| 设计Dbroker检索测试用例 | |
|  | |
|  | |
| 星期三 | | 设计Dbroker测试用例 | | 吴大衍 | 3 | 测试方向有偏差 | 根据已有测试方案设计原生impala的检索测试用例 | | | | |
| 准备讨论班材料 | |
| 设计原生impala测试用例 | |
|  | |
| 星期四 | | 测试原生impala | | 吴大衍  韩笑 | 4 | 无 | 无 | | | | |
| 讨论班报告 | |
|  | |
| 星期五 | | 体检 | | 韩笑  吴大衍 | 5 | Dbroker在进行联合查询+关键词查询时select count(\*)结果永远是0，select \*可以返回结果，查看dbk\_query日志发现执行select count(\*)命令时search条件没有被解析 | 记录问题上报古井子老师 | | | | |
| 测试原生impala和Dbroker检索性能 | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**： 吴大衍 **时间**：2015 年 9 月 7 日－2015 年 9 月 11日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 优化jdbc向se端的传值格式 | | 完成 | | 1 | 设计\*\*号项目的测试用例 | |  | 周二 | 吴大衍 | 吴大衍 |
| 2 | 学习安装mpp-engine | | 完成 | | 2 | \*\*号项目的测试 | |  | 未知 | 钟老师 | 吴大衍 |
| 3 | 学习counting sample和concise sample算法 | | 完成 | | 3 | 讨论班汇报 | |  | 周四 | 古老师 | 吴大衍 |
| 4 | 调研tpch的原理和使用方法 | | 未完成 | | 4 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 按照钟老师要求，优化jdbc向se端的传值格式并调整代码使得图像检索部分相对独立 | |  | 1 | 需要加入图像检索部分索引每个域的元数据信息 | 在se端找到元数据信息定义部分的代码并按照格式加入图像检索部分索引域的元数据信息 | | | | |
|  | |
|  | |
| 星期二 | | 学习安装mpp-engine | |  | 2 | Mpp-engine部署完成后ice服务始终无法正常启动 | 经查找缺少两个lib文件，加入后即可正常启动 | | | | |
|  | |
|  | |
|  | |
| 星期三 | | 学习counting sample和concise sample算法 | |  | 3 | 无 | 无 | | | | |
|  | |
|  | |
|  | |
| 星期四 | | 证明counting sample算法和concise sample算法的有效性 | |  | 4 | 数据流上的频繁项挖掘算法有一类很重要的解决方法就是抽样，抽样中最具代表性的算法是sticky sampling算法那，但是sticky sampling算法的原始论文并没有严谨证明其有效性，counting sample算法作为sticky sampling算法的前身其证明对理解sticky sampling算法很有帮助，但是如何counting sample算法的有效性就是个问题 | 将数据流上的抽样问题抽象成与其等效的一个概率模型 | | | | |
|  | |
|  | |
| 星期五 | | 调研tpch原理以及使用方法 | |  | 5 | 无 | 无 | | | | |
| 调研sql的执行过程(中间生哪些表等) | |
|  | |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **工作周报** | | | | | | | | | | | |
| **部门**：二室  **报告填写人**：吴大衍 **时间**：2015 年 8 月 31 日－2015 年 9 月 2日 | | | | | | | | | | | |
|
| **本周总结** | | | | | **下 周 工 作 计 划** | | | | | | |
| **编号** | **本周完成主要工作** | | **进展情况** | | **编号** | **下周主要事项** | | **预期进展** | **计划完成时间** | **负责人** | **参与人** |
| 1 | 和钟老师讨论传值格式并修改se端代码 | | 完成 | | 1 | \*\*号项目测试 | |  |  | 钟老师 | 吴大衍 |
| 2 | 调研elastic search和lucene的打分机制 | | 初步完成 | | 2 |  | |  |  |  |  |
| 3 | 调研tpc-h | | 完成 | | 3 |  | |  |  |  |  |
| 4 | 学习resevior sampling算法 | | 完成 | | 4 |  | |  |  |  |  |
| **本 周 工 作 记 录** | | | | | **本周工作中存在问题及建议解决办法** | | | | | | |
| **具 体 时 间** | | **工作内容记录** | | **参与人** | **编号** | **存在问题** | **提议解决办法** | | | | |
| 星期一 | | 调研lire网上demo并思考参数传递方式 | | 钟老师  　吴大衍 | 1 | 究竟需要传给se端哪些参数和图像检索模块的展现形式息息相关；  参数的传递形式越简单越好，减少jdbc和se端的通信量； | 调研LIRe的网上demo，借鉴其UI以及传递的参数类型；  采用十六进制表示哈希值，减少jdbc和se端的通信量； | | | | |
| 和钟老师讨论并修改se端代码 | |
|  | |
| 星期二 | | 结合钟老师的反馈意见进一步修改se端代码 | | 钟老师  吴大衍 | 2 | Elastic search的重打分机制工作原理是什么？为什么要引入重打分机制？  为什么elastic search要引入多种打分机制？好处是什么？ | 进一步阅读elastic search的源码和相关资料 | | | | |
| 调研elastic search和lucene的打分机制 | |
|  | |
|  | |
| 星期三 | | 学习resevior sampling算法 | | 钟老师  吴大衍 | 3 | 无 | 无 | | | | |
| 调研tpc-h工作原理 | |
|  | |
|  | |
|  | |
|  | |