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

编号	本周完成主要工作	进展情况
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

下 周 工 作 计 划

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	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

具 体 时 间	工作内容记录	参与人
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
星期二	BFSS 存在问题及其 SBFSS 算法描述部分撰写	吴大衍
星期三	SBFSS 算法描述及其 false positive 理论分析部分撰写	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
1		
2		
3		

星期四	阅读 SBF 算法论文，总结 SBF 参数选择策略 SBFSS 算法 false negative 部分撰写，包括符合选择参数等	吴大衍	4		
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 修改论文已完成部分，主要包括用词和语法	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 23 日－2015 年 5 月 27 日

本周总结

编号	本周完成主要工作	进展情况
1	线上系统 fastdfs 的备份以及过期数据的删除	完成
2	撰写论文理论分析部分	基本完成

下 周 工 作 计 划

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	实现线上系统的实时备份	完成	周五	钟 老师	吴大衍
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） 编写存储过程定时调用任务给华宇生成数据	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
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

编号	本周完成主要工作	进展情况
1	完成缩略图程序的开发和部署	完成
2	调研 fastdfs 的工作原理（包括安装、配置等）	完成
3	论文完成研究现状以及部分算法介绍的撰写	完成

下 周 工 作 计 划

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	完成	周五	钟 老师	吴大衍
2	给华宇生成展示所需数据	完成	周二	钟老师	吴 大衍
3	论文完成算法部分的撰写	完成	周五	王老师	吴 大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	党课学习 学习 fastdfs 工作原理 完成论文概述部分撰写	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法

星期二	部署分布式 fastdfs 撰写论文研究现状部分	吴大衍			
星期三	按照华宇展示需求修改存储过程 完成论文研究现状部分的撰写	王蒙 吴大衍			
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） 撰写论文算法部分	魏美茹 吴大衍			
星期五	撰写论文算法部分 参加定向越野比赛	贾德宾 王卓 吴大衍			

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

编号	本周完成主要工作	进展情况
1	图像检索部分 V3 的撰写	完成
2	声纹识别以及移动终端安全防护 v2 和 v3 的撰写	完成

下 周 工 作 计 划

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	调研 fastdfs 的基本原理，测试其鲁棒性	完成	周五	钟 老师	吴大衍
2	论文概述和研究现状部分的撰写	完成	周五	王老	吴 大

							师	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 协调落实 fileicon 字段的填充过程	钟老师 吴大衍			
星期二	完成 2-pass 算法和 mysql 的对比实验 和王老师讨论并确认下一步工作	王老师 吴大衍			
星期三	硕士答辩记录 开发视频缩略图生成部分代码	组内老师 吴大衍			
星期四	开会安排文档中支撑组件分布的撰写分工 和魏美茹协同开发 组内讨论班	小组成员 吴大衍			
星期五	撰写声纹识别和移动终端安全防护 v2v3 撰写论文摘要和概述部分	吴大衍			

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

编号	本周完成主要工作	进展情况
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
2	查找图像检索系统的相关材料，撰写 V1	完成
3		

下 周 工 作 计 划

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 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	完成 V2 撰写	完成	周二	李 老师	吴 大衍
3					

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
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
星期三	基于 BFSS 算法设计 two pass 算法查找精确低频项	吴大衍
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

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
1		
2		
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

编号	本周完成主要工作	进展情况
1	完成 BFSS 算法在合成数据集上的实验	完成
2	撰写材料	完成
3		

下 周 工 作 计 划

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	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	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					

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	改进实验代码 学习 java 进程资源监控工具 jprofile 的用法	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
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 按照分工撰写材料	吴大衍	3		
星期四	阅读 bloom filter 改进算法 stable bloom filter 算法论文 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	王老师 吴大衍	4		
星期五	和钟老师讨论工作进展 按照分工撰写材料 按照李老师的修改意见修改材料	李老师 钟老师 吴大衍			

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

编号	本周完成主要工作	进展情况
1	详细阅读 ES search 阶段代码	完成

下 周 工 作 计 划

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 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 和王老师讨论低频项挖掘工作进展	王老师 吴大衍	3		
星期四	编写 BSS 算法实验代码 和钟老师讨论低频项研究进展 讨论班学习	钟老师 吴大衍	4		
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

编号	本周完成主要工作	进展情况
1	讨论班报告	完成
2	完善 3.2 全文检索模块	完成
3		

下 周 工 作 计 划

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	合并 3.2 代码	完成	周一	钟 老师	吴大衍
2	调研多磁盘并发写技术	完成	周五	钟老师	吴 大衍
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

具 体 时 间	工作内容记录	参与人
星期一	完善 3.2 全文检索模块，使得 3.2 支持 select _score	吴大衍
	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
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

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
2		
3		

星期四	阅读哈尔小波变换算法相关论文 制作汇报 ppt 讨论班汇报	组内成员 吴大衍	4		
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

编号	本周完成主要工作	进展情况
1	调试 3.2 代码	完成
2		
3		

下 周 工 作 计 划

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	准备讨论班报告	完成	周三	吴大衍	吴大衍
2	整理 3.2 代码	完成	周五	钟老师	吴大衍
3	调研中间件相关技术			钟老师	吴大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人

本周工作中存在问题及建议解决办法

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

编号	本周完成主要工作	进展情况
1	3.1 全文检索代码移植到 3.2 上	完成
2	编译 3.2 代码	完成
3		

下 周 工 作 计 划

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	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

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 搭建 DBroker-3.2fe 端调试环境 组会讨论中间件开发问题	钟老师 林蝉 吴大衍			
星期五	合并 fe 端代码 编译 DBrokerSE-3.2	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 25 日－2015 年 1 月 29 日

本周总结

编号	本周完成主要工作	进展情况
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
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

下 周 工 作 计 划

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					

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	组会 阅读 elasticsearch 源码并做汇报	李老师 钟老师 吴大衍
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 室年会	吴大衍
星期四	理论推导基于 count-min 算法的低频项挖掘空间复杂度 陪王卓去医院	吴大衍
星期五	阅读 PLA 算法相关论文	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
1	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

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

编号	本周完成主要工作	进展情况
1	全文检索单 field 单 query 开发	完成
2	整理 Elasticsearch 相关资料并制作汇报 ppt	完成

下 周 工 作 计 划

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	全文检索加入 boost	完成	周五	钟老师	吴大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
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
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
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	组内成员

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
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

	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 讨论班	吴大衍			
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

编号	本周完成主要工作	进展情况
1	阅读 crate 以及 elasticsearch 全文搜索部分源码	完成
2	基于 DBroker3.1 进行全文检索功能的开发	未完成

下 周 工 作 计 划

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	全文检索功能开发	完成	周五	钟老师	吴大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	调研 elasticsearch 基本概念 和钟老师讨论全文检索设计方案	钟老师 吴大衍
星期二	部署 elasticsearch 调试环境	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
2	Sql 中同时存在 match 和 search 时该如何	先和各位老师讨论是否有这种需求，如

	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			处理	果有再讨论具体方案
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 组会	小组成员 吴大衍			
星期五	问林蝉语法文件修改方法 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	林蝉 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 4 日－2015 年 1 月 8 日

本周总结

编号	本周完成主要工作	进展情况
1	非时间分区列排序的开发与调试	完成
2	搭建 DBroker3.1 测试环境	完成

下 周 工 作 计 划

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	全文检索调研与设计	完成	周五	钟老师	吴大衍

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 调试代码使得查询有结果出现	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
星期三	组会部署下一步工作 调试代码使得查询支持 limit 数目较大的情况	钟老师 吴大衍	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
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 优化代码 fetch 部分的调度策略	吴大衍			
星期五	搭建 DBroker3.1 测试环境 调研 ElasticSearch 打分机制	吴大衍			

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

编号	本周完成主要工作	进展情况
1	非分区列排序代码开发	基本完成

下 周 工 作 计 划

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

具 体 时 间	工作内容记录	参与人
星期一	益园志愿者活动	志愿者小组 吴大衍
星期二	完成非分区列排序中 search 部分的开发	吴大衍
星期三	完成非分区列排序中 broker 部分的开发	吴大衍
星期四	完成非分区列排序中 fetch 部分的开发	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
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
3		
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

编号	本周完成主要工作	进展情况
1	完成以图搜图代码调试工作并整理上传 gitlab	完成
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

下 周 工 作 计 划

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	周四	钟老师	吴大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	整理以图搜图代码并上传 gitlab	吴大衍
星期二	在结果表中加入 score 属性并测试按照 score 排序	钟老师

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
1		
2		

	的性能 钟老师讲解非分区列排序方案	吴大衍			
星期三	阅读时间分区列排序代码 益园开会听取志愿者活动安排 修改 prezi	志愿者小组 吴大衍	3		
星期四	阅读时间分区列排序代码 理解非分区列排序方案并和钟老师讨论 搭建代码框架	钟老师 吴大衍	4		
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

编号	本周完成主要工作	进展情况

下 周 工 作 计 划

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

具 体 时 间	工作内容记录	参与人
星期一	整体调试以图搜图代码	吴大衍
	整理代码	
	代码冲突合并	
星期二	代码冲突合并	吴大衍
	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
	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
	确定 prezi 模板	
	寻找 prezi 素材	
	制作 prezi 整体框架+汇报收尾部分	

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
2		
3		
4		

星期五	合并 prezi 各部分制作 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	王老师 岳老师 视频制作小组	5		

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

编号	本周完成主要工作	进展情况
1	阅读 fe 端代码	完成
2	修改 fe 端和 be 端代码	完成
3	调试 se 端代码	完成
4	整体调试	基本完成

下 周 工 作 计 划

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	整体调试	完成	周二	吴大衍	吴大衍
2	准备和一室联调			钟老师	吴大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	阅读 fe 端代码	

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
1	Catalog 在 load 时将表的某一属性写到了	Be 端代码版本不对，替换 be 端代码，

	了解配置文件的具体格式 	吴大衍		thrift 结构中，但是在 catalog 中却看不到这一项	重新编译后运行即可
星期二	调试 se 端代码 构思 fe 端代码修改档案 搭建 fe 端代码调试环境 	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
星期三	搭建 fe 端代码调试环境 修改 fe 端代码 了解 cmake 以及 thrift 工作原理	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
星期四	修改 thrift 并重新编译项目 修改 fe 端和 be 端代码 	吴大衍	4		
星期五	听取 2015 年大数据技术大会报告 整体调试	李老师 钟老师 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 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

具 体 时 间	工作内容记录	参与人
星期一	开组会 阅读 se 端代码查询解析部分代码	小组人员 吴大衍
星期二	设计第一版修改方案 按照修改方案修改 se 端代码	吴大衍
星期三	调试 se 端修改代码	吴大衍
星期四	和钟老师讨论修改方案	钟老师

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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。
2	se 端有些参数的具体含义不明	搭建测试环境跟踪测试
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
4		

	设计第二版修改方案	吴大衍			
	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

编号	本周完成主要工作	进展情况
1	和一室讨论图像接口事宜并准备相关材料	完成
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

下 周 工 作 计 划

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 端代码并设计以图搜图接口调用方案	完成	周五	钟老师	吴大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	开组会	钟老师 吴大衍
	与一室讨论图像场景和以图搜图的接口对接问题	

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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	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

	写测试问题报告			可能平均	量尽可能平均
星期二	阅读 rocketmq 官方文档，了解 rocketmq 工作原理 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	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
星期三	搭建分布式的 rocketmq 环境 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	吴大衍	3		
星期四	阅读 SE 端代码 听取金超师兄的讨论会报告 学习使用 gitlab	张金超 吴大衍	4		
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

编号	本周完成主要工作	进展情况
1	数据仓库测试	完成

下 周 工 作 计 划

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	钟老师安排工作			钟老	吴大

							师	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

具 体 时 间	工作内容记录	参与人
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 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	钟老师 吴大衍
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） 完善测试文档	钟老师 吴大衍
星期三	容错测试（杀进程，杀 namenode，umount 磁盘） 整理测试文档并上传 测试文档脱密	钟老师 吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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)
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
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。	
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

编号	本周完成主要工作	进展情况
1	测试数据加载完成	完成
2	测试用例查询时间	基本完成

下 周 工 作 计 划

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	测试并发查询以及系统容错性能	完成	周二	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	完成一二号全量数据生成并上传 hdfs 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	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
星期二	完成初步测试 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	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
星期三	回所汇报测试进展 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	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
星期四	切分哈希分发后的大文件 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	钟老师 吴大衍	4		
星期五	撰写模拟数据生成方案 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 	钟老师 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 2 日－2015 年 11 月 6 日

本周总结

编号	本周完成主要工作	进展情况
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）
2	测试用例查询时间	未完成
3		
4		
5		
6		

下 周 工 作 计 划

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	完成数据仓库的测试	完成	周三	钟老师	吴大衍

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	测试 parquet+snappy 压缩比	钟老师

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
1	Parquet 存储 timestamp 类型的数据解析	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	王琳 吴大衍		会出错	
星期二	和钟老师讨论数据生成方案 开发模拟数据生成程序 配合王琳开发原始数据加载程序	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）
星期三	根据测试用例简化数据生成方案 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	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
星期四	利用 29 个节点同时生成第一天的全部数据（8 至 9T） 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	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
星期五	修改 put 程序，原有程序有问题，会导致脏数据上传 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	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

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

编号	本周完成主要工作	进展情况
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
2	完成第二轮原生 impala 测试结果的分析	完成
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

下 周 工 作 计 划

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	配合王琳加载数据				
2	中心测试数据仓库查询性能				

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期一	调研 hive 的分区规则 分析第二轮 impala 测试结果	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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

	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 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	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
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 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 	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
星期四	完善写 parquet 格式文件代码 阅读 parquet 格式相关论文，了解 parquet 格式中 definition 以及 repitition 的具体含义 	吴大衍	4	无	无
星期五	用中心测试用例测试 shell 是否运行正确 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	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

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

编号	本周完成主要工作	进展情况
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
2	第二轮原生 impala 测试（小表 10 万，100 万）	完成
3	阅读原生 impala 内存管理部分代码	基本完成
4	中心听取测试要求以及测试样例	完成
5	分析第二轮测试结果	未完成

下 周 工 作 计 划

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	完成第二轮测试结果分析	完成	周一	吴大衍	吴大衍
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	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					
5					

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人

本周工作中存在问题及建议解决办法

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 编写第二轮 impala 测试的测试样例	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
星期四	完成第二轮原生 impala 测试数据的加载 编写第二轮原生 impala 测试程序	吴大衍	4	无	无
星期五	完成 10 万小表的测试并分析结果 中心听取测试方案以及测试样例	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）

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

编号	本周完成主要工作	进展情况
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
2	完成小数据集上原生 impala 性能测试	完成
3	生成第二轮测试数据	未完成
4		

下 周 工 作 计 划

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 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	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					

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 了解 be 端整个 planfragment 的执行流程	吴大衍	1	无	无
星期二	阅读 be 端 aggregation-node 算子 益园测试小数据上的原生 impala 性能（大表 2000 万、小表 10 万）	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
星期三	阅读 impala 官方文档关于 impala 回写磁盘的介绍 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	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
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 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	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

星期五	阅读 be 端有关哈希表操作的部分代码 修改数据加载代码，加快数据加载速度	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

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

编号	本周完成主要工作	进展情况
1	第一轮原生 impala 性能测试	完成
2	阅读 impala 2.2be 端代码	未完成
3		
4		

下 周 工 作 计 划

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	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 be 端代码定位查询慢的原因	完成	周五	吴大衍	吴大衍
3					
4					

本 周 工 作 记 录

具 体 时 间	工作内容记录	参与人
星期四	完成第一轮原生 impala 性能测试	钟老师

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
3	小表 c_pzid 的值太少（1-25）导致和大表	调整小表中 c_pzid 的范围为 1-10 万

	和钟老师讨论优化方案	吴大衍		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	
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

具 体 时 间	工作内容记录	参与人
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 开组会	吴大衍
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
星期三	编写原生 impala 的 jdbc 测试代码	吴大衍

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
1	无	无
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
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	编写 impala 的 jdbc 测试代码，并将其在后台运行

				会很长，不可能等着它执行完毕	
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 去益园查看 impala 的具体查询计划和日志定位查询瓶颈	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 0（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

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

编号	本周完成主要工作	进展情况
1	设计 Dbroker 和原生 impala 检索测试用例	完成

下 周 工 作 计 划

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

具 体 时 间	工作内容记录	参与人
星期一	开组会 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	韩笑 吴大衍
星期二	准备讨论班材料 设计 Dbroker 检索测试用例	吴大衍
星期三	设计 Dbroker 测试用例 准备讨论班材料 设计原生 impala 测试用例	吴大衍
星期四	测试原生 impala 讨论班报告	吴大衍 韩笑
星期五	体检	韩笑

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
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
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
4	无	无
5	Dbroker 在进行联合查询+关键词查询时	记录问题上报古井子老师

	测试原生 impala 和 Dbroker 检索性能	吴大衍		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	

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

编号	本周完成主要工作	进展情况
1	优化 jdbc 向 se 端的传值格式	完成

下 周 工 作 计 划

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**号项目的测试用例		周二	吴大	吴大

							衍	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

具 体 时 间	工作内容记录	参与人
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	
星期二	学习安装 mpp-engine	
星期三	学习 counting sample 和 concise sample 算法	
星期四	证明 counting sample 算法和 concise sample 算法的有效性	

本周工作中存在问题及建议解决办法

编号	存在问题	提议解决办法
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
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
3	无	无
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	将数据流上的抽样问题抽象成与其等效的一个概率模型

				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	
星期五	调研 tpch 原理以及使用方法 调研 sql 的执行过程(中间生哪些表等)		5	无	无

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

编号	本周完成主要工作	进展情况
1	和钟老师讨论传值格式并修改 se 端代码	完成

下 周 工 作 计 划

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	**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 和钟老师讨论并修改 se 端代码	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；
星期二	结合钟老师的反馈意见进一步修改 se 端代码 调研 elastic search 和 lucene 的打分机制	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
星期三	学习 resevior sampling 算法 调研 tpc-h 工作原理	钟老师 吴大衍	3	无	无