友乾营

专注数据技术的社群

如何实现

海量数据清单和分组报表

2 期

www.raqsoft.com.cn/yqy

什么是海量数据报表?

友乾营

报表要展现的数据可达百万、甚至千万级别，通常以清单列表方式呈现，少数情况以分组报表形式呈现

订单ID	客户ID	订购日期	运货费	货主名称	货主
10248	VINET	2012-07-04	¥32.38	余小姐	北
10249	TOMSP	2012-07-05	¥11.61	谢小姐	济
10250	HANAR	2012-07-08	¥65.83	谢小姐	秦
10251	VICTE	2012-07-08	¥41.34	陈先生	南
10252	SUPRD	2012-07-09	¥51.30	刘先生	长
10253	HANAR	2012-07-10	¥58.17	谢小姐	长
10254	CHOPS	2012-07-11	¥22.98	林小姐	武
10255	RICSU	2012-07-12	¥148.33	方先生	北
10256	WELLI	2012-07-15	¥13.97	何先生	济
10257	HILAA	2012-07-16	¥81.91	王先生	上
10258	ERNSH	2012-07-17	¥140.51	王先生	济
10259	CENTC	2012-07-18	¥3.25	林小姐	上
10260	OTTIK			徐文彬	北
10261	QUEDE			刘先生	济
10262	RATTC	2012-07-22	¥48.29	王先生	上

列表

货主地区	订单ID	客户ID	订购日期	运货费	订单金额
东北	10783	HANAR	2013-12-18	¥124.98	¥1442.50
	10932	BONAP	2014-03-06	¥134.64	¥1788.63
	10723	WHITC	2013-10-30	¥21.72	¥468.45
	10641	HILAA	2013-08-22	¥179.61	¥2054.00
	10700	SAVEA	2013-10-10	¥65.10	¥1638.40
	10701	SAVEA	2013-10-13	¥220.31	¥2864.50
	10730	BONAP	2013-11-05	¥20.12	¥484.26
	10760	MAISD	2013-12-01	¥155.64	¥2917.00
	10670	FRANK	2013-09-16	¥203.48	¥2301.75
	11068	QUEEN	2014-05-04	¥81.75	¥2027.08
				订单金额小计:	**¥102410.19**
华东	10737	VINET	2013-11-11	¥7.79	¥139.80
	10894	SAVEA	2014-02-18	¥116.13	¥2753.10
	10502	PERIC	2013-04-10	¥69.32	¥816.30
	10681			.13	¥1287.40
	10503			.74	¥2048.50
	10689			.42	¥472.50
	10885	SUPRD	2014-02-12	¥5.64	¥1209.00

分组

数据库分页存在的问题`

友乾营

翻页效率差

页码小时，感觉不明显；
页码较大时，翻页会有明显的等待感

可能出现汇总错误

每页的SQL都是单独发送的，如果期间数据库发生增删操作，可能导致数据汇总错误

无法实现分组效果

由于每页取数无法保证取出整组数据，因而也无法实现分组报表

无法使用非RDB数据源

数据库分页基于RDB，其他数据源（如NoSQL或文件则无法使用）

还能想到哪些办法？

友乾营

游标

向数据库发出取数SQL生成游标，从中取出一页后呈现，但并不终止这个游标，要取下一页的时候再继续取数

只能向后翻页

混用

向后翻页时用游标，一旦发生向前翻页时，则重新执行取数SQL

解决问题不彻底

一个很大的问题

友乾营

数据源耦合性太强

每种数据库的实现方式不同，更换数据源都要重新实现

强耦合

如何从根本上解决问题？

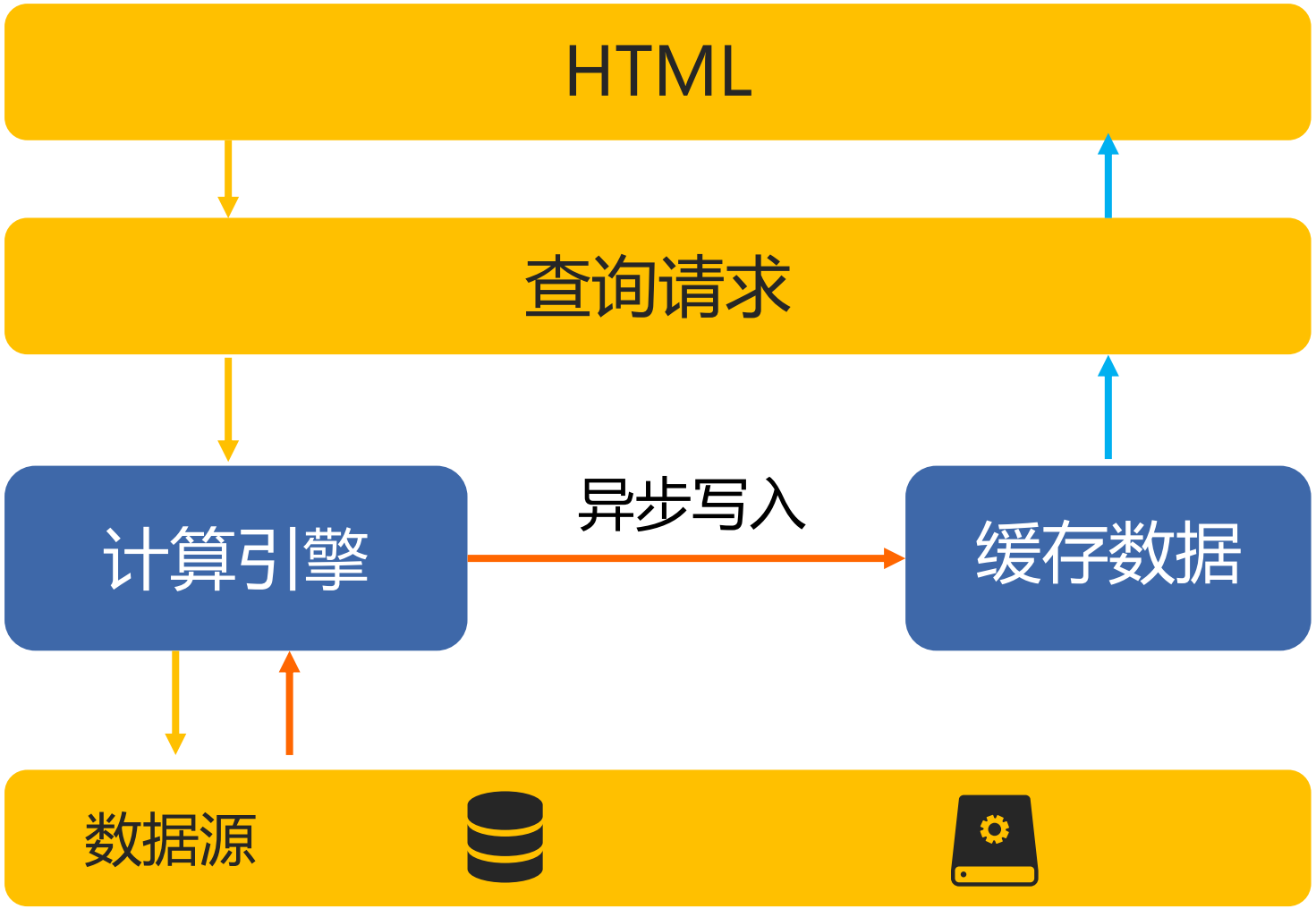
友乾营

取数和呈现做成两个异步线程

- 取数线程发出SQL后就不断取出数据并缓存到本地存储中，呈现线程根据页数计算出行数到本地缓存中去获取数据显示
- 只要已经取过的数据就能快速呈现，不会有等待感，还没取到的数据需要等待一下也是正常可理解的
- 取数线程只涉及一句SQL，在数据库中是同一个事务，也不会有不一致的问题

应用层面

异步双线程目标架构



要具备哪些能力

友乾营

多源支持

可以基于RDB和其他类型数据源

游标机制

可以通过游标分批返回数据

5个

必备能力

高性能存储

具备高性能缓存功能，可以按行号随机访问记录

较完备的计算能力

可以基于不具备计算能力的数据源（如文件）实施计算

按分组返回数据

可以返回完成分组数据，以实现分组效果

应该注意什么问题

友乾营

不要全表排序

大报表的数据集都比较大，如果在意响应时间（谁会不在意呢），那么应该尽量不对数据集进行全表排序（注意我说的是全表排序），毕竟，等排完序再呈现，时间已经过去很久了

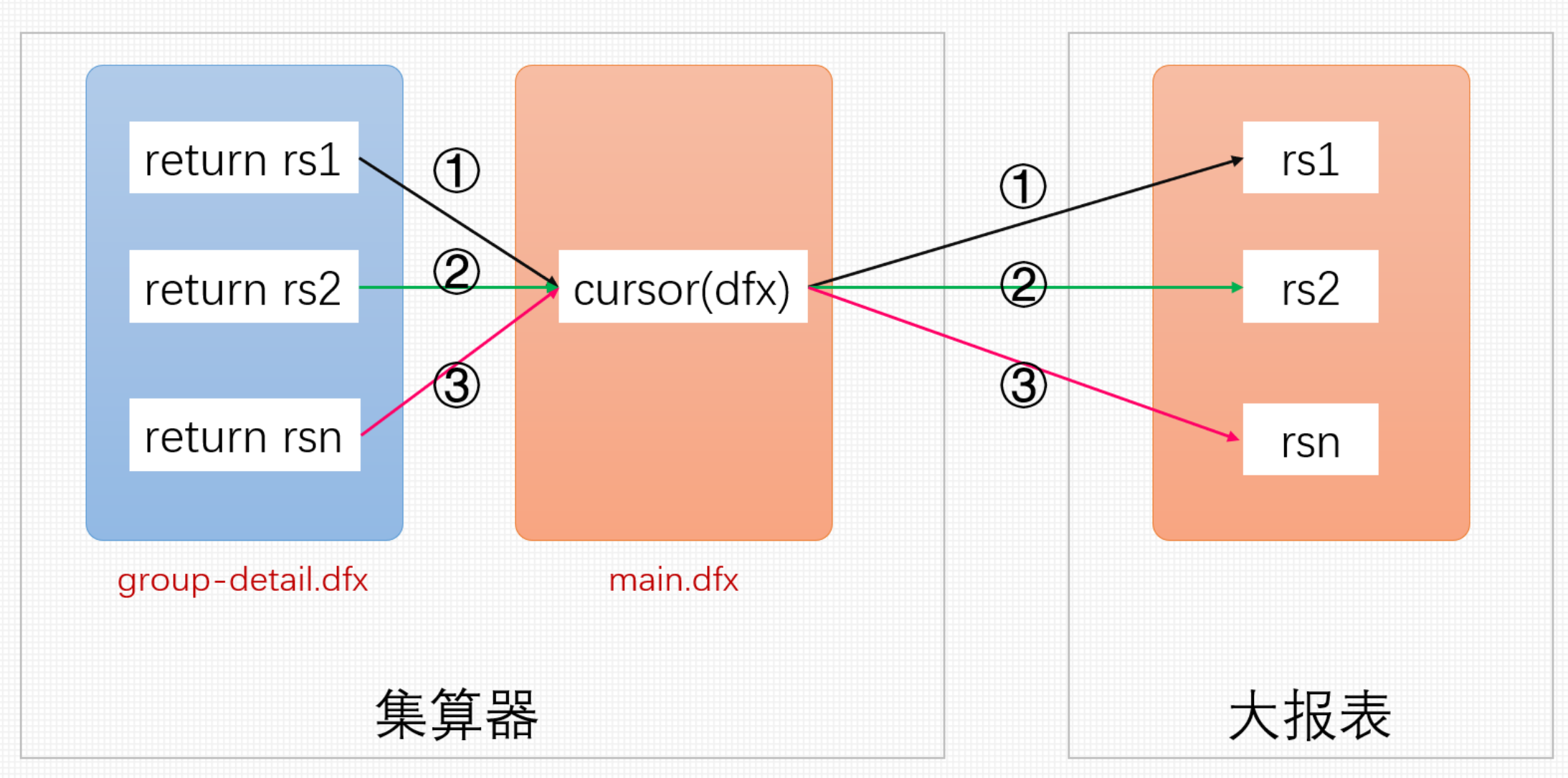
不适合高并发场景

大报表采用异步机制，将数据分批加载到内存再交给前端呈现，减少了内存占用，但同时增加了 CPU 和磁盘 I/O 负载，并发高时 CPU 和硬盘可能成为瓶颈从而影响呈现效果，因此大报表不适合高并发的场景

分组报表单个分组不宜过大

由于计算分组明细和汇总值时需要将某一个分组数据全部加载到内存中进行计算，因此分组相对内存容量不宜过大，从而确保单个分组数据能进行全内存计算。

大分组举例


return rs1
return rs2
return rsn
①
②
③
cursor(dfx)
①
②
③
rs1
rs2
rsn
group-detail.dfx
main.dfx
集算器
大报表

下期预告

友乾营

《在word报告中插入报表》

报告格式很复杂，人肉生成不现实！

依靠技术硬编码，内容太广不现实！

使用工具来实现，没有适合很痛苦！

下周三19:30 逐一探讨上述问题，并分享实用解决思路！

掰开揉碎，讲给你听！

润乾线上直销系统

好多乾 – 润乾互联网营销

http://www.raqsoft.com.cn/wx/hdq-strategy.html