



飞总的IT世界
面面观_wtg

29 文章 | 1.6万 总阅读

[查看TA的文章>](#)

0

[分享到](#)

SparkSQL-有必要坐下来聊聊Join

2017-06-02 09:24

[大数据 / 技术](#)

这个系列是我偶然从数据管理公众号读到得，讲查询优化是怎么做得。写得很好。虽然查询优化是本行也是我现在每天要做的事情。我自问自己来写，没办法写得做这样这样的通俗易懂。我问作者要了授权。感谢数据管理公众号的授权。

Join背景介绍

Join是数据库查询永远绕不开的话题，传统查询SQL技术总体可以分为简单操作（过滤操作-where、排序操作-limit等），聚合操作-groupBy等以及Join操作等。其中Join操作是最复杂、代价最大的操作类型，也是OLAP场景中使用相对较多的操作。因此很有必要聊聊这个话题。

另外，从业务层面来讲，用户在数仓建设的时候也会涉及Join使用的问题。通常情况下，数据仓库中的表一般会分为“低层次表”和“高层次表”。

所谓“低层次表”，就是数据源导入数仓之后直接生成的表，单表列值较少，一般可以明显归为维度表或者事实表，表和表之间大多存在外键依赖，所以查询起来会遇到大量Join运算，查询效率相对比较差。而“高层次表”是在“低层次表”的基础上加工转换而来，通常做法是使用SQL语句将需要Join的表预先进行合并形成“宽表”，在宽表上的查询因为不需要执行大量Join因而效率相对较高，很明显，宽表缺点是数据会有大量冗余，而且生成相对比较滞后，查询结果可能并不及时。

因此，为了获得实效性更高的查询结果，大多数场景还是需要进行复杂的Join操作。Join操作之所以复杂，不仅仅因为通常情况下其时间空间复杂度高，更重要的是它有很多算法，在不同场景下需要选择特定算法才能获得最好的优化效果。关系型数据库也有关于Join的各种用法，关注公众号InsideMySQL并回复join可以查看相关文章，这里面详细介绍了MySQL Join的各种算法以及调优方案。本文接下来会介绍SparkSQL所支持的几种常见的Join算法以及其适用场景。

Join常见分类以及基本实现机制

当前SparkSQL支持三种Join算法 - shuffle hash join、broadcast hash join以及sort merge join。其中前两者归根到底都属于hash join，只不过在hash join之前需要先shuffle还是先broadcast。其实，这些算法并不是什么新鲜玩意，都是数据库几十年前的老古董了（参考），只不过换上了分布式的皮而已。不过话说回来，SparkSQL/Hive...等等，所有这些大数据技术哪一样不是来自于传统数据库技术，什么语法解析AST、基于规则优化（CRO）、基于代价优化（CBO）、列存，都来自于传统数据库。就拿shuffle hash join和broadcast hash join来说，hash join算法就来自于传统数据库，而shuffle和broadcast是大数据的皮，两者一结合就成了大数据的算法了。因此可以这样说，大数据的根就是传统数据库，传统数据库人才可以很快的转型到大数据。好吧，这些都是闲篇。

继续来看技术，既然hash join是“内核”，那就刨出来看看，看完把“皮”再分析一下。

Hash Join

大家都在搜：十大最蠢密码公布



热门图集



官方撒糖！哈里王子与准王妃甜蜜拥抱



她是姜文喜欢和姜文



黑龙江漠河：选手在零下33℃比赛钢管舞



20年前经典画面，超乎



24小时热文

- 苹果强行降低旧手机诉讼 这次会如何公
- 世界最美女性的国绝对想不到
- 阿里腾讯公共互怼全靠小学生，腾讯马云没有马仔

个Join节点，参与join的两张表是item和order，join key分别是item.id以及order.i_id。现在假设这个Join采用的是hash join算法，整个过程会经历三步：

- 1. 确定Build Table以及Probe Table：这个概念比较重要，Build Table使用join key构建Hash Table，而Probe Table使用join key进行探测，探测成功就可以join在一起。通常情况下，小表会作为Build Table，大表作为Probe Table。此事例中item为Build Table，order为Probe Table；
- 2. 构建Hash Table：依次读取Build Table（item）的数据，对于每一行数据根据join key（item.id）进行hash，hash到对应的Bucket，生成hash table中的一条记录。数据缓存在内存中，如果内存放不下需要dump到外存；
- 3. 探测：再依次扫描Probe Table（order）的数据，使用相同的hash函数映射Hash Table中的记录，映射成功之后再检查join条件（item.id = order.i_id），如果匹配成功就可以将两者join在一起。

基本流程可以参考上图，这里有两个小问题需要关注：

- 1. hash join性能如何？很显然，hash join基本都只扫描两表一次，可以认为 $O(a+b)$ ，较之最极端的笛卡尔集运算 $a*b$ ，不知甩了多少条街；
- 2. 为什么Build Table选择小表？道理很简单，因为构建的Hash Table最好能全部加载在内存，效率最高；这也决定了hash join算法只适合至少一个小表的join场景，对于两个大表的join场景并不适用。

上文说过，hash join是传统数据库中的单机join算法，在分布式环境下需要经过一定的分布式改造，说到底就是尽可能利用分布式计算资源进行并行化计算，提高总体效率。hash join分布式改造一般有两种经典方案：

- 1. broadcast hash join：将其中一张小表广播分发到另一张大表所在的分区节点上，分别并发地与其上的分区记录进行hash join。broadcast适用于小表很小，可以直接广播的场景；
- 2. shuffle hash join：一旦小表数据量较大，此时就不再适合进行广播分发。这种情况下，可以根据join key相同必然分区相同的原理，将两张表分别按照join key进行重新组织分区，这样就可以将join分而治之，划分为很多小join，充分利用集群资源并行化。

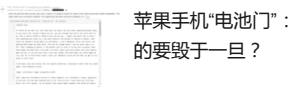
Broadcast Hash Join

如下图所示，broadcast hash join可以分为两步：

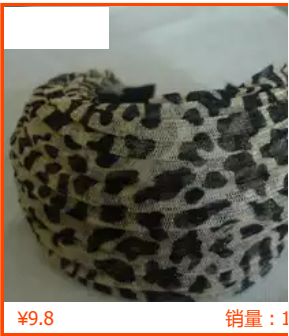
- 1. broadcast阶段：将小表广播分发到大表所在的所有主机。广播算法可以有很多，最简单的是先发给driver，driver再统一分发给所有executor；要不就是基于bittorrent的p2p思路；
- 2. hash join阶段：在每个executor上执行单机版hash join，小表映射，大表试探；

SparkSQL规定broadcast hash join执行的基本条件为被广播小表必须小于参数spark.sql.autoBroadcastJoinThreshold，默认为10M。

Shuffle Hash Join



苹果手机“电池门”：的要毁于一旦？



搜狐号推荐

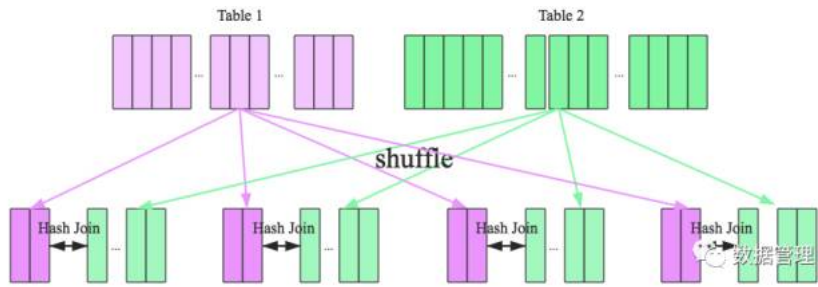
- 视觉 搜狐科技视界 搜狐科技官方原创账号。聚焦件、大趋势和新变化，用我们
- 卫星情报局 地球的事儿，我们都知道！
- 猎云网 猎云网是一家科技新媒体，聚势、创业创新报道，关注新产
- 驱动之家 驱动之家网站是在IT行业内居资讯、产品评测、驱动程序
- 蓝鲸TMT 蓝鲸TMT网,关注移动互联网创互联网热点新闻事件。



联系我们

率最高。但是一旦小表数据量增大，广播所需内存、带宽等资源必然就会太大，broadcast hash join就不再是最优方案。此时可以按照join key进行分区，根据key相同必然分区相同的原理，就可以将大表join而治之，划分为很多小表的join，充分利用集群资源并行化。如下图所示，shuffle hash join也可以分为两步：

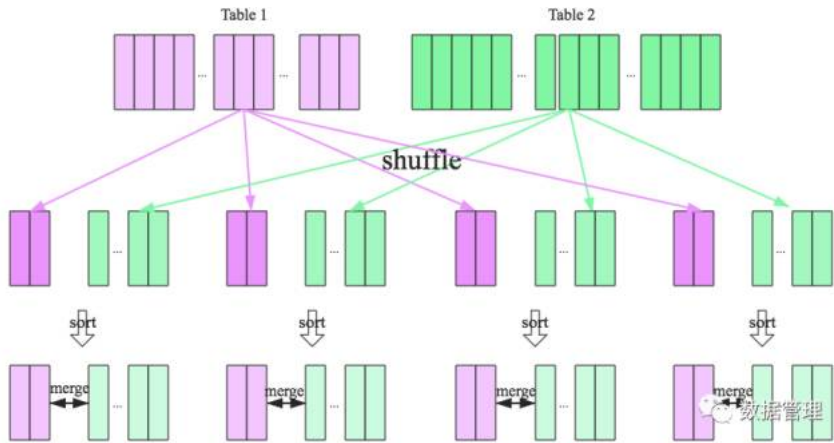
- 1. shuffle阶段：分别将两个表按照join key进行分区，将相同join key的记录重分布到同一节点，两张表的数据会被重分布到集群中所有节点。这个过程称为shuffle
- 2. hash join阶段：每个分区节点上的数据单独执行单机hash join算法。



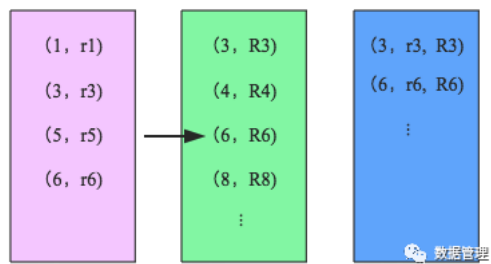
看到这里，可以初步总结出来如果两张小表join可以直接使用单机版hash join；如果一张大表join一张极小表，可以选择broadcast hash join算法；而如果是一张大表join一张小表，则可以选择shuffle hash join算法；那如果是两张大表进行join呢？

Sort-Merge Join

SparkSQL对两张大表join采用了全新的算法 - sort-merge join，如下图所示，整个过程分为三个步骤：



- 1. shuffle阶段：将两张大表根据join key进行重新分区，两张表数据会分布到整个集群，以便分布式并行处理；
- 2. sort阶段：对单个分区节点的两表数据，分别进行排序；
- 3. merge阶段：对排好序的两张分区表数据执行join操作。join操作很简单，分别遍历两个有序序列，碰到相同join key就merge输出，否则取更小一边，见下图示意：



经过上文的分析，可以明确每种Join算法都有自己的适用场景，数据仓库设计时最好避免大表与大表的join查询，SparkSQL也可以根据内存资源、带宽资源适量将参数spark.sql.autoBroadcastJoinThreshold调大，让更多join实际执行为broadcast hash join。

总结

Join操作是传统数据库中的一个高级特性，尤其对于当前MySQL数据库更是如此，原因很简单，MySQL对Join的支持目前还比较有限，只支持Nested-Loop Join算法，因此在OLAP场景下MySQL是很难吃的消的，不要去用MySQL去跑任何OLAP业务，结果真的很难看。不过好消息是MySQL在新版本要开始支持Hash Join了，这样也许在将来也可以用MySQL来处理一些小规模的OLAP业务。

和MySQL相比，PostgreSQL、SQLServer、Oracle等这些数据库对Join支持更加全面一些，都支持Hash Join算法。由PostgreSQL作为内核构建的分布式系统Greenplum更是在数据仓库中占有一席之地，这和PostgreSQL对Join算法的支持其实有很大关系。

总体而言，传统数据库单机模式做Join的场景毕竟有限，也建议尽量减少使用Join。然而大数据领域就完全不同，Join是标配，OLAP业务根本无法离开表与表之间的关联，对Join的支持成熟度一定程度上决定了系统的性能，夸张点说，“得Join者得天下”。本文只是试图带大家真正走进Join的世界，了解常用的几种Join算法以及各自的适用场景。如果大家有兴趣的，可以查看Spark相关代码进行更深入的学习。

欢迎打赏支持飞总的写作

往期精选

加飞总小密圈和大咖嘉宾聊天



[返回搜狐，查看更多](#)

阅读 (187)

不感兴趣

投诉

本文相关推荐

mysql+left+join

sparksql+和hive的区别

update+join

join+实现原理

join在数据库中的用法

join的过去式

hive怎么进行join操作

join.me怎么使用

sql语句中的join

大表和小表join+哪个在先

join创意实战笔记

join+走索引


¥49.00 ¥159


¥900


¥49.00 ¥159

客厅灯

床垫

冲锋衣

沙发垫

电脑椅

广告

我来说两句

0人参与，0条评论

来说两句吧.....

登录并发表





搜狐“我来说两句” 用户公约

还没有评论，快来抢沙发吧！

- 推荐
- OPPO
- iPhone
- QQ
- 百度
- 荣耀
- 唯品会
- 小红书
- 乐视
- TCL
- 5G
- 美团
- 华为

推荐阅读


苹果强行降低旧手机性能 引集体诉讼 这次会如何公关呢？



零镜网 · 昨天 21:56

2


比特币24小时重挫逾30%，跌破11000美元



IT之家 · 昨天 23:07



5

西媒称“网红”热席卷中国:业内竞争激烈 多数人收入微薄



参考消息 · 今天 00:24



暴走CEO任剑再回应王尼玛出走风波：源起前员工被高利贷追债



雷帝触网 · 昨天 21:49

1

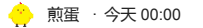

马云没有马仔



接招 · 昨天 21:34

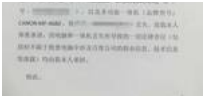
9

如果蚊子灭绝了会怎样





煎蛋 · 今天 00:00

1





驰？


 雷帝触网 · 昨天 21:53


 4

B站首推年度弹幕，竟是这个字！王安石还成了幕后推手



 刺猬公社 · 今天 07:47








A股最坑新股，开板之后重组，复牌连续跌停，散户骂声一片！散户：该拿什么拯救你，我的A股！

广告 · 今天 11:14

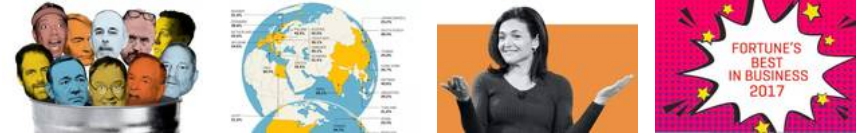
Elon Musk的完美圣诞：特斯拉利好不断、Space X里程碑式突破在即





 36氪 · 今天 00:44




有人闷声发大财，有人遭遇滑铁卢——2017科技领域大盘点





 新智造 · 昨天 21:00







埃里克·施密特卸任 Alphabet 执行董事长，回顾「Google 三巨头」的这 17 年


 极客公园 · 昨天 21:39






36氪领读 | 从“饮食简史”到“好好吃饭”，我们如何重新做好味蕾的“经营者”？


 36氪 · 今天 09:01







90后美女，辞职半年，在家炒股。每日推荐1股！过几天看效果！

广告 · 今天 11:14





网易推白金卡 圈地运动下的用户需求


 Gamewower丁鹏 · 今天 10:43




差距不过100亿！董明珠雷军10亿赌约大限将至，输赢不重要，广告费已经...


 投资界 · 12-21 14:18

 7



解救帕金森患者被冻结的脚步！研究人员发明出一双激光鞋和一条“安全带”

 36氪 · 今天 07:21



音频产品市场来了新巨头，传今日头条将上线音频内容



90后女股民自爆一年打中18只新股，原因竟如此简单！

广告 · 今天 11:14

橙意家人获飞利浦战略投资，联手加强睡眠呼吸慢病管理

投资界 · 12-21 19:46

股票	涨幅	成交量	换手率
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%
东方财富	1.15%	1.15%	1.15%

阿里云计算将在明年1月开放印度数据中心 | 印度创投周报

钛媒体 APP · 今天 10:54

途牛再陷“裁员”风波，回应：研发架构调整，将确保员工权益

36氪 · 昨天 17:04

请把圣诞节变成愚人节 @微信官方



硅谷密探 · 今天 08:04

广告 · 今天 11:14



美国警察的执法记录仪有用吗？

TechCrunch 中文版 · 昨天 14:37

差距不过100亿！董明珠雷军10亿赌约大限将至，输赢不重要，广告费已经...

投资界 · 12-21 14:18

有人闷声发大财，有人遭遇滑铁卢——2017科技领域大盘点



新智造 · 昨天 21:00

时钟基因解开不吃早餐反倒增胖之谜

TechNews 科技新报 · 今天 10:15

A股最坑新股，开板之后重组，复牌连续跌停，散户骂声一片！散户该怎么办！



广告 · 今天 12:55

职场中层，更要懂得如何取舍

滚石特写: 沉默7年后, Magic Leap用魔幻现实主义式科技重新定义了自己







机器之能 · 昨天 13:00

阿里讽腾讯整条命是小学生给的，被怼不识数；“吃鸡”游戏官方：99%外挂来自中国，但绝...

创业邦 · 今天 09:21

那个给南仁东制造“大麻烦”的人，今天站在台上说.....




动静贵州 · 昨天 22:25

圣诞节要来了，精选五大千元以下 3C 圣诞礼物清单



电獭少女
台湾知名女生科技媒体
ZEALER 科技生活第一站 视频

ZEALER · 昨天 19:46







why?老师在课堂上一言不合就尬舞而走红

男人街 · 12-12 16:11


悍卫领空？美军为何年年要追踪圣诞老人，来看看圣诞老人飞到哪里吧

TechNews科技新报 · 今天 10:20

科氮做电商，被「忽悠」了3次




36氪 · 昨天 16:39



上班不喝茶，喝茶就喝陈皮普洱。卖爆了，陈皮普洱仅198元/桶

广告 · 今天 12:55



软文营销：看了50个教程，终于知道故事要这样写！

砍柴网 · 今天 09:56

爱点击在美国悄无声息的上市了 蓝色光标曾投资6000万美元



上谷歌搜施密特，你会回来转发的



虎嗅 虎嗅APP · 昨天 21:47



Taylor Swift 登基社交女王！粉丝限定的社交 App 《The Swift Life》

Tech News TechNews科技新报 · 今天 10:35



纳什空间获得来自华融融德的战略融资，鸟巢旗舰项目正式开业

钱眼网 投资界 · 12-21 17:53



世界首例悲剧：俄罗斯男子玩VR游戏不幸身亡

好特游戏 好特游戏 · 今天 10:10



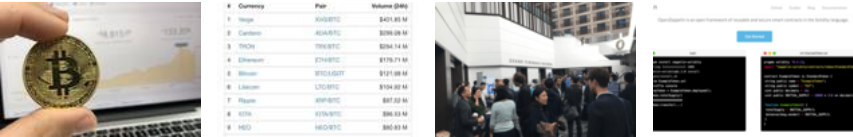
《芳华》：人与人为什么会疏远，这是我见过的最好答案



黑马 i黑马 · 昨天 20:34

1

从硅谷看区块链市场：中国热潮依旧，重要机构入场



钛媒体APP 钛媒体APP · 今天 09:00



90后女股民自爆一年打中18只新股，原因竟如此简单！



广告 · 今天 12:55

“双旦”将至 最适合送男友的手机推荐



蓝鲸TMT 蓝鲸TMT · 今天 11:25



博恩集团董事长熊新翔：互联网已不再是热点，投资要看这“六刀”

猎云网 猎云网 · 今天 12:09





诺基亚2手机暴力测试：质量延续传统


IT之家 IT之家 · 今天 09:07





Eilon Musk的元美全诞：特斯拉利好个断、Space X里程碑式突破在即



Rank	Space	Current Status	Percentage of all launches
1	SpaceX	Active	38%
2	ULA	Active	25%
3	Orion	Active	15%
4	Orion	Active	15%
5	Orion	Active	15%
6	Orion	Active	15%
7	Orion	Active	15%
8	Orion	Active	15%
9	Orion	Active	15%
10	Orion	Active	15%



 36氪 · 今天 00:44





 3


想做我男朋友？先有2000万吧


 猫聘同道 · 今天 10:00



除了《真爱至上》，今年圣诞你还需要看这些暖心广告



 爱范儿 · 今天 08:06





雷军在印度通过打车软件呼叫摩托

 IT之家 · 昨天 20:29

 1



【钛晨报】故意让老iPhone速度变慢，美国用户已起诉苹果

 钛媒体 APP · 今天 09:01





36氪领读 | 从“饮食简史”到“好好吃饭”，我们如何重新做好味蕾的“经营者”？

 36氪 · 今天 09:01







玉函路隧道今晚24点试通车！怎么走看这里


 济南发布 · 昨天 16:15




知识图谱大牛组团来阿里，他们都聊了什么？





 阿里技术 · 今天 09:59



这个人帮Google打下了半壁江山，还间接影响了互联网的格局



 爱范儿 · 昨天 17:07



我待小米 AI 音箱如初恋，它却虐我千百遍

 ZEALER · 今天 08:00



贾跃亭还有机会翻身吗？

智谷趋势 · 昨天 19:29

3



逼人买新机？苹果承认限制旧款iPhone性能

KOM 123 陆家嘴杂志 · 今天 06:00

...

吴恩达究竟是人工智能的布道者还是卖水人









脑极体 · 昨天 21:50

...



我炒股，只信自己！每天送你一只股，牛不牛，你先看看再说！

广告 · 今天 14:4

小米米家重磅新品12月25日揭晓

IT之家 · 今天 10:42

...

中国联通腾讯大王/天王卡添福利

IT之家 · 今天 09:30

...

三星存储器之神坐上CEO大位，SK 海力士发信吁员工备战

Tech News TechNews科技新报 · 今天 11:00

...



阿里王帅：腾讯光靠小学生 刘强东擅长碰瓷

威锋网 · 昨天 18:22

19

加载更多