Atitit 数据存储的分组聚合 groupby的实现attilax总结

[1.](#_Toc15785) **[聚合操作](#_Toc15785)** [1](#_Toc15785)

[1.1. a、标量聚合 流聚合 1](#_Toc7948)

[1.2. b、哈希聚合 2](#_Toc24905)

[1.3. 所有的最优计划的选择都是基于现有统计信息来评估 3](#_Toc13203)

[1.4. 参考资料 3](#_Toc19748)

# ****聚合操作****

聚合也是我们在写T-SQL语句的时候经常遇到的，我们来分析一下一些常用的聚合操作运算符的特性和可优化项。

## a、标量聚合 流聚合

标量聚合是一种常用的数据聚合方式，比如我们写的语句中利用的以下聚合函数：MAX()、MIN()、AVG()、COUNT()、SUM()

以上的这些数据结果项的输出基本都是通过流聚合的方式产生，并且这个运算符也被称为：标量聚合

其实，流聚合这种算法最常用的方式是分组（GROUP BY）计算，上面的标量计算也是利用这个特性，只不过把整体形成了一个大组进行聚合。

我么通过如下代码理解

[IMG_256](http://www.cnblogs.com/zhijianliutang/p/javascript:void(0);)

clear the current aggredate results

clear the current group by columnsfor each input row

begin

if the input row does not match the current group by columns

begin

output the current aggreagate results(if any)

clear the current aggreagate results

set the current group by columns to the input row

end

update the aggregate results with the input row

end

[IMG_257](http://www.cnblogs.com/zhijianliutang/p/javascript:void(0);)

流聚合运算符其实过程很简单，维护一个聚合组和聚合值，依次扫描表中的数据，如果能不匹配聚合组则忽略，如果匹配，则加入到聚合组中并且更新聚合值结果项。

举个例子

SELECT ShipAddress,ShipCity,COUNT(\*)

FROM Orders

GROUP BY ShipAddress,ShipCity

所以这里我们已经总结出对于流聚合的一种优化方式：尽量避免排序产生，而要避免排序就需要将分组（Group by）字段在索引覆盖范围内。

## b、哈希聚合

上述的流聚合的方式需要提前排序，我们知道排序是一个非常大的消耗过程，所以不适合大表的分组聚合操作，为了解决这个问题，又引入了另外一种聚合运算：哈希聚合

所谓的哈希聚合内部的方法和本篇前面提到的哈希连接机制一样。

哈希聚合不需要排序和过大的内存消耗，并且很容易并行执行计划，利用多CPU同步进行，但是有一个缺点就是：这一过程是阻塞的，也就说哈希聚合不会产生任何结果直到完整的输入。

所以在大数据表中采用哈希聚合是一个很好的应用场景。

我们知道排序是很耗资源的一件事情，但是利用哈希匹配只需要将不同的列值进行提取就可以，所以相比性能而言，无疑哈希匹配算法在这里是略胜一筹的算法。

## 所有的最优计划的选择都是基于现有统计信息来评估

而上面关于这两列内容分布类型SQL Server是怎样知道的？这就是SQL Server的强大的统计信息在支撑了。

在SQL Server中并不是固定的语句就会形成特定的计划，并且生成的特定计划也不是总是最优的，这和数据库现有数据表中的内容分布、数据量、数据类型等诸多因素有关，而记录这些详细信息的就是统计信息。

所有的最优计划的选择都是基于现有统计信息来评估，如果我们的统计信息未及时更新，那么所评估出来最优的执行计划将不是最好的，有时候反而是最烂的。

## 参考资料

SQL Server调优系列基础篇（常用运算符总结——三种物理连接方式剖析） - 指尖流淌 - 博客园.html

作者:: 绰号:老哇的爪子claw of Eagle 偶像破坏者Iconoclast image-smasher

捕鸟王"Bird Catcher 王中之王King of Kings 虔诚者Pious 宗教信仰捍卫者 Defender of the Faith. 卡拉卡拉红斗篷 Caracalla red cloak

简称：： Emir Attilax Akbar 埃米尔 阿提拉克斯 阿克巴

全名：：Emir Attilax Akbar bin Mahmud bin attila bin Solomon Al Rapanui

埃米尔 阿提拉克斯 阿克巴 本 马哈茂德 本 阿提拉 本 所罗门  阿尔 拉帕努伊

常用名：艾提拉（艾龙），   EMAIL:1466519819@qq.com

转载请注明来源：attilax的专栏  http://blog.csdn.net/attilax

--Atiend