SYBASE 技术服务园地 连载（73） SYBASE TECHNOLOGY SERVICE FIELD

解决方案

Sybase 15.3 横向扩展查询性能

使用 PlexQ 分布式查询平台、全共享的 MPP 架构（二）

（上接第8期）

2.3 Sybase IQ15 中对 Intra-Operator 的增强

Sybase IQ 15.0 版本显著的增强了 Intra-operator 并行化。许多查询操作现在可以使用多个线程并行执行。

多数的表 Join 操作

（1）Group By 操作。

（2）排序（Oder By 与 Merge Joins）。

（3）表中的谓词执行：例如："WHERE last_name like "%som%"，范围谓词，IN 条件，"Top N" 操作，以及更多。

在 Sybase IQ 15.3 之前，一个单独的查询的节点间（Inter-nodes）和节点内（Intra-Nodes）并行化仅能使用一个单一服务器上的 CPU 资源。当时，Multiplex 配置在扩展支持越来越多的并发用户或者查询上非常有效，但是对于利用跨 Multiplex 的所有计算带宽而减少查询执行时间上无所作为。新的 Sybase IQ 15.3 PlexQ 分布式查询处理功能消除了这种限制，允许一个查询使用 Sybase IQ Multiplex 中所有机器上的可能的 CPU 资源。

现在，让我们详细探索 Sybase IQ 15.3 PlexQ 全共享 MPP 架构中的 DQP 如何工作？

3 理解分布式查询处理

3.1 什么是 DQP？

分布式查询处理（DQP）将查询处理分散到 Sybase IQ Multiplex 中的多个服务器上。一个 Sybase Multiplex 是一组服务器，每个都运行 Sybase IQ。Multiplex 中的服务器连接到一个中心存储，比如一个共享的磁盘阵列，永久共享数据。Sybase IQ Multiplex 有一个混合的集群架构，包含永久的 IQ 数据的共享存储、存储目录元数据的独立的节点存储、私有的临时数据、事务日志。

当 Sybase IQ 查询优化器认为一个查询可能需要利用多个节点上更多的可用 CPU 资源的时候，它会将查询分为并行的"碎片"，这些"碎片"可以在 Multiplex 中的其他服务器上并行执行。DQP 是这样一个过程：将查询分解为多个独立的任务部分，将这些任务分布到 Multiplex 中的其他节点上，将结果及时的收集和组织并生成最终的查询结果集。

需要重点强调的是，如果一个查询不能完全利用一个单一机器上的 CPU 资源的话，它通常不会被分布。例如，如果优化器将把一个查询并行化为 7 条（保持 7 个线程），而 CPU 有 8 个核，则它不会分布这个查询。分布要求网络和存储的开销以分配任务，以及存储和传输即时的结果。在一个 DBMS 内的目标就是尽可能快速的执行查询。一个简单的查询在单一的机器上执行是快速的。然而，大的、复杂的查询，超出了一个机器的 CPU 能力，可能更好的方式就是利用分布式并承担因此带来的开销。如果性能提高了，那么分布式就是胜利。

3.2 DQP 如何工作？

DQP 对任何在 Multiplex 网格配置中部署了 Sybase IQ 15.3 的客户都是可用的。当你安装了 Sybase IQ，DQP 缺省的是打开的，所有 Multiplex 中的服务器均可被分布式处理所利用。

DQP 引入了"Leader"和"Worker"节点的概念。Leader 节点是查询发起的节点，Worker 节点可以是 Multiplex 中的任何有能力接受分布式查询处理工作的节点。所有的 Multiplex 节点类型(读节点、写节点、协调节点)都可以作为 Leader 或 Worker 节点。

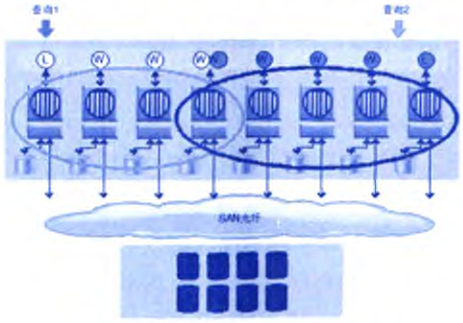

图3 一个运行中的分布式查询

在上图中，查询 1 和查询 2 的执行被分布到 Multiplex 中的节点中。这两个查询由不同的 Leader 节点和一组 Worker 节点来完成。这是一个可能的运行场景。你可以非常灵活的配置节点集（参见：下

面的“逻辑服务器”部分）参与一个分布式的查询。

Sybase IQ 15.3 也启用了一个新的共享DBSpace（数据库空间）以支持DQP，叫作共享临时存储（Shared Temporary Store）。这个DBSpace被命名为IQ_SHARED_TEMP，必须存放于可被Multiplex中所有节点访问和写入的共享磁盘存储上。这也是对IQ_SYSTEM_MAIN和用户自定义的DBSpace的共同要求。IQ_SHARED_TEMP的目的是允许即时数据在分布式查询所覆盖的服务器间双向传递。IQ_SHARED_TEMP与本地临时存储IQ_SYSTEM_TEMP都使用临时缓冲作为数据的内存缓冲区。

当一个客户端向Sybase IQ服务器提交一个查询，查询优化器使用成本分析选择是否并行化与分布查询的执行。一个可并行化的查询被分解为查询碎片—谓词和数据流动的子树。只有Sybase IQ引擎支持碎片中包含的所有查询操作的并行和分布式执行时，一个查询才被认为是适合分布的。

当一个查询被分布，Leader节点将查询碎片分派给Worker节点，并从Worker服务器收集即时的结果。Worker服务器不决定查询分布，他们只是简单的执行分派给自己的任务并返回结果。

如果一个查询优化器作出这样一个决定：一个分布式查询不适合分发，甚者可能会降低性能，那么这个查询将不会被分布，而是会在Multiplex中的单个节点上执行。查询可以被分为如下几类：

（1）不分布的：没有碎片在Multiplex的其他节点上执行。该查询仅在Leader节点上完成。

（2）部分分布的：一个或多个碎片在Multiplex的其他节点上执行，包含“Leader”节点。

（3）完全分布的：所有碎片在Multiplex的多个节点上执行。

3.3 逻辑服务器（Logical Server）

你可能不总是希望使用Multiplex中的所有服务器执行分布式查询，而且可能希望用户使用一个资源的子集。为了这个目的，Sybase IQ引入了逻辑服务器的概念。一个逻辑服务器允许一个Multiplex的一个或多个服务器组合在一起作为一个逻辑实体。用户基于登录政策授权访问逻辑服务器。

有一些内建的逻辑服务器。特别是，内建的OPEN逻辑服务器包含所有不属于任何用户自定义的逻辑服务器的成员的服务器。如果你不建立任何逻辑服务器，Multiplex中的所有节点可能会参与到DQP中，因为他们都是OPEN服务器的一部分。

用户登录政策可能允许访问一个或多个逻辑服务器。一个用户将会连接到一个物理服务器上运行一个查询。Sybase IQ查看用户登录政策，决定物理服务器属于哪个逻辑服务器。然后将查询执行分布到逻辑服务器的那些成员节点上。尽管一个物理服务器可能属于不止一个逻辑服务器，但是它不可能属于分配了相同登录政策的多个逻辑服务器。例如，一个用户连接到逻辑服务器A和B，物理服务器C不属于逻辑服务A,B的成员。这确保如果用户X连接到物理服务器C，在选择执行查询的逻辑服务器时不会发生歧义。你可以动态增加或减少逻辑服务器的成员服务器以适应不断变化的应用的资源需求。

3.4 Multiplex节点间通讯

为了支持参与查询分布的节点间的流线式通讯，IQ 15.3引入了Multiplex节点间通讯（MIPC）框架。MIPC网是一个点到点的节点间通讯基础架构，是对IQ15.0中增加的节点间通讯(INC)协议的补充。INC用于双向的心跳监测、版本数据同步、以及Multiplex中要求的其他类型的消息和数据的传播。INC允许节点间通过协调节点互相对话，对当个节点查询的相当有限的通讯需求来讲已经足够。MIPC允许Multiplex节点间彼此直接对话，支持更强健的DQP的通讯需求。

MIPC既有公共的也有私有的配置选项。私有选项允许你指定主机端口配对（此时仅限于TCP/IP协议），Multiplex服务器将仅对DQP相关的通讯使用该配对。如果没有提供私有的互联配置，MIPC使用为其他类型通讯所设定的主机端口配对：外部用户连接和INC连接。

在内部测试阶段，一个私有的MIPC网络比一个共享的MIPC网络提供了明显的性能优势—在一个特定的实例中，一个分布式查询，运行于私有MIPC网络中的2个节点上，和使用共享MIPC网络的3个节点配置的执行速度一样快。

3.5 DQP的先决条件

你无须为激活分布式查询处理而进行任何配置。除非你通过关闭dqp_enabled登录政策选项或dqp_enabled临时数据库选项禁止了DQP，DQP对合格的查询自动启用，当：

（1）该服务器是Multiplex的一部分。

（2）有一个逻辑服务器允许登录，而且至少一个节点可用。缺省的，有一个叫做OPEN逻辑服务器的内建的逻辑服务器，所以这个需求出厂时即已满足。

（3）共享临时DBSpace有可写的文件。最初，共享临时DBSpace中没有DBFiles，Multiplex管理员必须增加至少一个原始设备DBFile，以激活分布式查询处理。

3.6 哪种类型的查询可以在Multiplex中分布

一个查询操作能被分布，首先必须能够被并行执行。当一个操作并行执行的时候，多个线程可用于并行执行这个过程。在IQ15.3中，多数查询操作可以被并行化，但不是所有的查询都可以被分布。

下表显示了哪些查询操作可以被分布：

表1 查询操作分布

分类	操作
JOIN	Nested Loop / Nested Loop Pushdown
	Hash / Hash Pushdown
	Sort Merge / Sort Merge Pushdown
Group By	GROUP BY SINGLE
	GROUP BY (HASH)
	GROUP BY (SORT)
DISTINCT	DISTINCT(HASH)
	DISTINCT(SORT)
SORT	ORDER BY
	ORDER BY (IN)
	SORTED IN
SUBQUERY	Uncorrelated
PREDICATES	条件执行（使用FP / LF / HG索引）
OLAP	OLAP RANK与WINDOWS with PARTITION
SELECT component of	INSERT ... SELECT
INSERT operations	INSERT ... LOCATION

具有如下行为的查询碎片将不会被分布：

（1）写入数据库（包括DDL，INSERT，LOAD，UPDATE和DELETE）

（2）引用临时表。

（3）引用存于SYSTEM DBSpace中的表。

（4）引用代理表。

（5）使用不确定的函数，例如NEWID。

需要注意的是，LOAD操作仍然可以被“分布”，通过使用Multiplex中的多个节点并行加载单独的表。

3.7 你如何知道一个查询是否被分布？

Sybase IQ查询计划让你可以看到查询是否被分布。查询计划提供了详细的信息，指出哪些服务器参与了查询处理，评估任务如何被分布，以及显示时间信息。

当一个客户端连接到一个物理服务器并启动一个查询的时候，DQP开始运行。这个服务器查询的是Leader节点。该Leader节点调用查询优化器，建立查询执行计划。查询优化器建立一个查询树，将查询分解为碎片。一个碎片是下列其中之一：

（1）一个叶条件（一个谓词）。

（2）一个数据流动的子树，有特定的分区：范围分区或关键词分区碎片是查询树的一部分，可以被单独执行。当两个碎片可以无所谓前后顺序执行时，它们可能被并行执行。如果一个碎片依赖于另一个碎片的即时结果，那么两个碎片必须按照适当的顺序执行。如果碎片中所有的查询操作都是可并行化和可分布的，那么这个碎片就可以跨所有Worker节点分布。不能被分布的碎片将完全在Leader节点上执行。这个优化器将每个查询操作分解为碎片，作为一组“任务单元”。一个任务单元是一组数据集，一个处理线程基于该数据集进行工作。

这是一个查询计划分解查询碎片的例子。你实际上不会在现实的查询计划中看到下图中的虚线部分。这仅仅是给你一个认识，一个查询可能如何被优化器分解。在这个例子中，碎片1、2、3将会被并行执行：

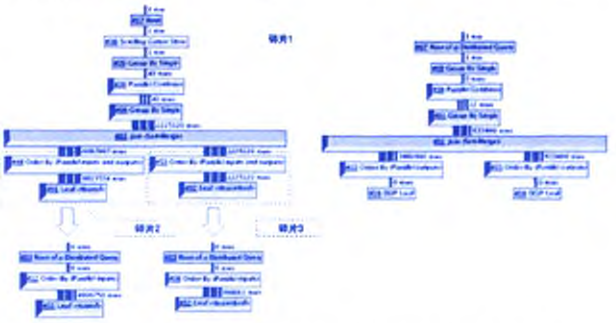

图4 描述分布式处理碎片的查询计划样板

当你打开创建查询计划文件的数据库选项，整个查询的查询计划将在Leader节点上被创建。

当一个查询碎片是一个数据流动子树，而且它被分布，参与执行该碎片的每个Worker节点将为这个碎片生成一个本地的查询计划。（注意：你仅需在Leader节点打开查询计划数据库选项，不需要为在Worker节点上创建查询碎片计划而在Worker节点上打开该选项。）一个碎片最顶端的查询操作器管理该碎片的任务单元并将任务单元分配到跨所有Worker的线程。

任务单元的线程分配是一个高度动态的过程，允许线程在查询执行时增加或删除。线程根据机器负载和资源可用性纵向或回溯扩展。临时缓冲和CPU时间的可用性是决定增加或删除线程的决定性因素。在Sybase IQ DQP中，物理服务器可以动态地添加到逻辑服务器上，并且在经过一些初始化之

后，一旦新的查询碎片被安排分布即可开始执行DQP工作。

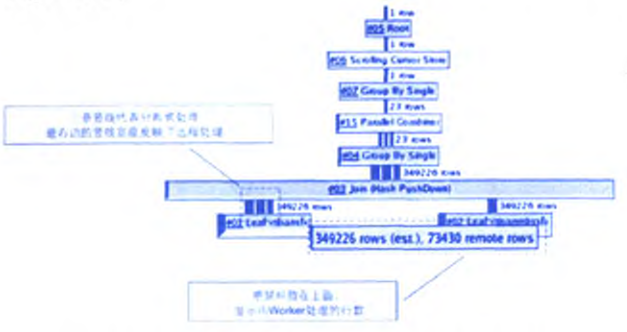

图5 描述分布式处理碎片的查询计划样板

如果一个查询仅有部分被分布，你将会看到在被分布的节点之间有一个三条的黑色竖线。当你将鼠标移动到紧邻这个并行线的行计数上时，将会显示远程的行数（有多少被分布）。最右端的竖线的宽度由远程行数来决定。

在查询树下，是时间图表。最上端，对每一个查询树中的节点，你将会看到执行的每个阶段的用时。现在这包括了Multiplex中所有服务器的用时。时间图表上关于CPU使用的部分将显示所有服务器的汇总用时。

在节点的阶段用时下面是线程显示。它显示了在特定的时间，哪个服务器的哪个线程在执行任务。线程分配以堆栈条图形显示：

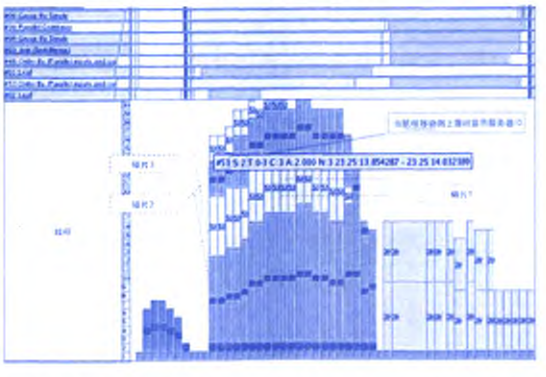

图6 查询计划部分显示Multiplex中的线程和分布

如果你将鼠标移动到线程条上，你会看到各种统计，比如：

（1）#53：在正在执行的查询碎片根部的节点号。

（2）S：2：服务器ID（2），拥有正在执行的线程。

（3）T：0-3：正在执行的线程的范围。

（4）A：2：在那个时间段内执行的线程的平均数。

（5）N：3：在那个时间段内用于计算线程统计的样本数。

（6）23：25：13...— 23：25：14：时间段的起始和结束时间。

如果一个查询碎片同时在多个服务器上执行，你会发现线程被碎片堆栈中彼此相邻且位于上方的碎片的相同的根节点而阻断。

下图的时间表是特定节点的详细信息：

图7 查询计划详细的特定节点的信息

对于一个特定的节点，你会发现任务是如何被分布到服务器和线程的。在上面的“碎片1”中，服务器“kwd_nc1615”的任务单元的值是“25(2,6,4,4,3,2,3,1)”。这意味着有25个任务单元分配到这个服务器，而且任务单元2,6,4,4,3,2,3和1分别被分配到8个不同的线程。你也会发现有多少私有和共享临时空间被用于执行这个碎片。

“碎片2”显示了首先分配到Worker节点的任务单元数量。大于1意味着Leader节点在Worker节点开始处理之前首先执行了某些任务。这可能是由于需要开始执行任务的Worker产生了一个延迟。

“碎片3”显示了“并行下池任务单元”，即整个碎片任务单元的总数。

4 错误是如何被处理的？

DQP可容忍Worker节点/网络的失败以及Worker节点的缓慢。如果一个Worker节点由于一个错误或者超时没能完成任务单元，这个任务单元会重新回到Leader节点。如果发生这种现象，该Worker节点在碎片执行期间将不会再分配给任务单元。

尽管一个Worker节点可能在执行某个查询碎片时发生失败，它仍然可以在稍后被分配给不同查询碎片的任务单元。

文/赛贝斯软件（中国）有限公司

（未完待续）