文章
姚 鑫 · 四月 19, 2021 阅读大约需 11 分钟

第三章 优化表（二）

第三章 优化表（二）

调整表计算值

调优表操作根据表中的代表性数据计算和设置表统计信息:
- ExtentSize，它可能是表中的实际行数(行数)，也可能不是。
- 表中每个属性(字段)的选择性。
可以选择性地阻止单个属性的选择性计算。
- 属性的离群选择性，其中一个值比其他值出现得更普遍。
有效的查询可以利用离群值优化。
- 标识某些属性特征的每个属性的注释。
- 每个属性的平均字段大小。
- 表的SQL Map Name、BlockCount和Source of BlockCount。

区段大小和行计数

从管理门户运行Tune Table工具时，ExtentSize是表中当前行的实际计数。默认情况下，GatherTableStats()方法还将实际行数用作ExtentSize。当表包含大量行时，最好对较少的行执行分析。可以使用SQL tune table命令并指定%SAMPLEPERCENT来仅对总行的一定百分比执行分析。在针对包含大量行的表运行时，可以使用此选项来提高性能。此%SAMPLEPERCENT值应该足够大，以便对代表性数据进行采样。如果ExtentSize<1000，则无论%SAMPLEPERCENT值如何，TUNE TABLE都会分析所有行。

指定的ExtentSize可以小于或大于实际行数。但是，ExtentSize不应显著超过当前表数据中的实际行数。指定ExtentSize时，Tuning Table会为该数量的行外推行ID，然后执行采样。如果ExtentSize大大超过实际行数，则大多数采样的行ID将与实际行数据不对应。如果是这种情况，则无法计算字段选择性；相反，Tune Table将列出指定的ExtentSize作为计算的ExtentSize，并列出一个较小的数字作为SAMPLESIZE；Tune Table为这些不存在的计算值返回<Not Specified>。

可以将ExtentSize设置为0。如果表从来不打算填充数据，但用于其他目的(如查询联接)，则可能需要这样做。当将ExtentSize设置为0时，InterSystems IRIS会将每个字段的选择性设置为100%，并将每个字段的平均字段大小设置为0。

选择性和异常值选择性

Tune Table以百分比计算每个属性(字段)值的选择性。
它通过对数据进行抽样来实现这一点，因此选择性总是一种估计，而不是一个精确的值。
选择性是基于所有属性值都是(或可能是)等可能的假设。
对于大多数数据来说，这是一个合理的假设。
例如，在一个普通人口表中，大多数数据值都是典型的:任何特定的出生日期都会出现在大约0.27%的数据中(365分之一);
大约一半是女性，一半是男性(50%)。
被定义为Unique的字段的选择性为1(不应与选择性为1.0000(1%)混淆)。
对于大多数性质，选择性百分比就足够了。

对于一些属性，Tune Table还计算离群值选择性。
这是单个属性值的百分比，与其他数据值相比，该属性值在示例中出现的频率更高。

只有当一个数据值的频率与其他数据值的频率存在显著差异时，调优表才会返回离群值选择性。
无论数据值的分布情况如何，Tune Table最多为表返回一个离群值。
如果选择了一个离群值，那么调优表将此值显示为离群值。
NULL表示为< NULL >。

如果TuneTable返回异常值选择性，则正常选择性仍然是整个行集内每个非异常值数据值的百分比。例如，如果在1000个随机选择的值中检测到11个不同的值，其中一个是异常值，则选择性为1/11(9.09%)：平均每个条目出现的几率为十一分之一。如果异常值选择性是80%，常规选择性是1%，那么除了异常值之外，还可以找到大约20((1-0.80)/0.01)个额外的非异常值。

如果优化表初始采样仅返回单个值，但附加采样返回多个不同的值，则这些采样结果会修改正常选择性。例如，990个值的初始随机采样仅检测一个值，但后续采样检测其他不同值的10个单个实例。在这种情况下，初始离群值会影响选择性值，该值现在被设置为1/1000(0.1%)，因为10个非离群值中的每一个在1000个记录中只出现一次。

异常值选择性的最常见示例是允许NULL的属性。如果某个特性具有NULL的记录数大大超过该特性具有任何特定数据值的记录数，则NULL为异常值。以下是FavoriteColors字段的选择性和异常值选择性：

```
SELECTIVITY of FIELD FavoriteColors
   CURRENT =     1.8966%
   CALCULATED =  1.4405%
   CURRENT OUTLIER = 45.0000%, VALUE = <Null>
   CALCULATED OUTLIER = 39.5000%, VALUE = <Null>
```

如果一个字段只包含一个不同的值(所有行都具有相同的值)，则该字段的选择性为100%。选择性为100%的值不被视为异常值。调谐表通过采样数据来建立选择性和异常值选择值。为了确定这一点，优选表首先测试少量或几条记录，如果这些记录都具有相同的字段值，它将测试多达100,000条随机选择的记录，以支持非索引字段的所有值都相同的假设。只有在字段已编制索引，字段是索引的第一个字段，并且字段和索引具有相同的排序规则类型的情况下，优化表才能完全确定该字段的所有值是否相同。

- 如果已知未编制索引的字段具有在测试100,000条随机选择的记录中可能检测不到的其他值，则应手动设置选择性和离群值选择性。
- 如果已知非索引字段没有其他值，则可以手动指定100%的选择性，删除任何异常值选择性，并设置CALCSELECTIVITY=0以防止优选表尝试计算选择性或将此值指定为异常值。

要修改这些选择性、异常值选择性和异常值计算值，请从调谐表显示中选择单个字段。这会在显示屏右侧的详细信息区域中显示该字段的这些值。可以将选择性、异常值选择性和/或异常值修改为更适合预期完整数据集的值。

- 可以将选择性指定为带有百分号(%)的行的百分比，也可以指定为整数行(没有百分号)。如果指定为整数行数，InterSystems IRIS将使用区大小来计算选择性百分比。
- 可以为以前没有异常值的字段指定异常值选择性和异常值。将异常值选择性指定为带百分号(%)的百分比。如果仅指定异常值选择性，则Tune Table假定异常值为<Null>。如果仅指定异常值，则除非还指定异常值选择性，否则调谐表不会保存此值。

CALCSELECTIVITY参数与不计算选择性

在某些情况下，可能不希望优化表工具计算属性的选择性。要防止计算选择性，请将属性的CALCSELECTIVITY参数的值指定为0(默认值为1)。在Studio中，可以在“ 新建属性向导 ”的“ 属性参数 ”页上设置CALCSELECTIVITY，也可以在检查器中的属性参数列表中设置CALCSELECTIVITY(可能需要收缩并重新展开属性参数列表才能显示它)。

应该指定CALCSELECTIVITY=0的一种情况是，如果该字段未编制索引，则已知该字段在所有行中只包含一个值(选择性=100%)。

离群值的优化

默认情况下，查询优化器假定查询不会选择离群值。
例如，查询通常选择特定的字段值并从数据库返回少量记录，而不是返回大量记录，其中该字段值是离群值。
查询优化器总是使用选择性来构造查询计划，除非执行一些要求考虑离群选择性的操作。

根据选择离群值，可以执行以下几个操作来调整查询优化:

- 如果异常值是<null>，则在查询WHERE子句中为该字段指定一个is null或is NOT null条件。
  这将导致查询优化器在构造查询时使用离群值选择性。

- 如果离群值是一个数据值，查询优化器会假定选择的字段值不是离群值。
  例如，总部位于马萨诸塞州的公司的员工记录可能有OfficeState字段离群值MA (Massachusetts)。
  优化器假设查询不会选择' MA '，因为这将返回数据库中的大多数记录。
  但是，如果正在编写一个查询来选择离群值，可以通过将离群值封装在双括号中来通知优化器。
  在该字段上查询时，指定一个WHERE子句，如下所示:WHERE OfficeState=(('MA'))。
  这种技术抑制了文字替换，并迫使查询优化器在构建查询计划时使用离群值选择性。
  对于动态SQL查询，以及在使用ODBC/JDBC提供的InterSystems
  IRIS之外编写的查询，这种语法是必需的。
  对于类查询、嵌入式SQL查询或通过视图访问的查询，则不需要这样做。

- 根据参数值SQL设置配置系统范围的优化查询。
  该选项为离群值设置了运行时计划选择(RTPC)优化和作为离群值(BQO)优化的偏差查询的适当组合。
  注意，更改此配置设置将清除所有名称空间中的所有缓存查询。
  使用管理门户，选择System Administration、Configuration、SQL和Object
  Settings、SQL来查看和更改此选项。
  可用的选择有:

  - 假设查询参数值不是字段离群值(BQO=OFF, RTPC=OFF，初始默认值)
  - 假设查询参数值经常匹配字段离群值(BQO=ON, RTPC=OFF)
  - 在运行时优化实际查询参数值(BQO=OFF, RTPC=ON)

要确定当前设置，调用$SYSTEM.SQL.CurrentSettings()。

- 覆盖查询的系统范围配置设置。

通过指定%NORUNTIME restrict关键字，可以覆盖单个查询的RTPC。
如果查询SELECT Name,HaveContactInfo FROM t1 WHERE HaveContactInfo=?
将导致RTPC处理，查询SELECT %NORUNTIME Name,HaveContactInfo FROM t1 WHERE HaveContactInfo=?
将覆盖RTPC，从而产生一个标准的查询计划。

通过指定注释选项/*#OPTIONS {"BiasAsOutlier":1} */，可以覆盖偏见查询作为单个查询的离群值。

“备注”列

管理门户优化表信息选项为每个字段显示一个备注列。此字段中的值是系统定义的，不可修改。它们包括以下内容：
- RowID字段：一个表有一个RowID，由系统定义。它的名称通常是ID，但可以有不同的系统分配的名称。由于其所有值(根据定义)都是唯一的，因此其选择性始终为1。如果类定义包括SqlRowIdPrivate，则Notes列值为RowID字段、Hidden字段。
- 隐藏字段：隐藏字段定义为私有，SELECT*不显示。默认情况下，CREATE TABLE将RowID字段定义为隐藏；可以指定%PUBLICROWID关键字以使RowID不隐藏和公开。默认情况下，由持久化类定义定义的表将RowID定义为非隐藏；可以指定SqlRowIdPrivate将RowID定义为隐藏和私有。容器字段定义为隐藏。
- 流字段：表示使用流数据类型定义的字段，可以是字符流(CLOB)，也可以是二进制流(BLOB)。流文件没有平均字段大小。
- 父引用字段：引用父表的字段。

注释列中未标识标识字段、ROWVERSION字段、序列字段或UNIQUEIDENTIFIER(GUID)字段。

平均字段大小

运行调谐表根据当前表格数据集计算所有非流字段的平均字段大小(以字符为单位)。这与AVG($length(Field))相同(除非另有说明)，四舍五入到小数点后两位。可以更改各个字段的平均字段大小，以反映字段数据的预期平均大小。
- NULL：因为$LENGTH函数将NULL字段视为长度为0，所以将长度为0的NULL字段取平均值。这可能会导致平均字段大小小于一个字符。
- 空列：如果列不包含数据(所有行都没有字段值)，则平均字段大小值为1，而不是0。对于不包含数据的列，AVG($length(Field))为0。
- ExtentSize=0：将ExtentSize设置为0时，所有字段的平均字段大小将重置为0。
- 逻辑字段值：平均字段大小始终根据字段的逻辑(内部)值计算。
- 列表字段：InterSystems IRIS列表字段根据其逻辑(内部)编码值计算。此编码长度大于列表中元素的总长度。
- 容器字段：集合的容器字段大于其集合对象的总长度。例如，在Sample.Person中，Home容器字段的Average Field Size大于HomeStreet、HomeCity、HomeState和HomeZip平均字段大小的总和。
- 流字段：流字段没有平均字段大小。

如果特性/字段的特性参数CALCSELECTIVITY设置为0，则调谐表不会计算该特性/字段的平均字段大小。

可以通过从调谐表显示中选择单个字段来修改平均字段大小计算值。这将在显示屏右侧的详细信息区域中显示该字段的值。可以将“ 平均字段大小 ”修改为更适合预期的完整数据集的值。由于设置此值时优化表不执行验证，因此应确保该字段不是流字段，并且指定的值不大于最大字段大小(MaxLen)。

平均字段大小还显示在管理门户目录详细信息选项卡字段选项表中。必须已为字段选项表运行了调整表，才能显示平均字段大小值。

map BlockCount选项卡

调优表Map BlockCount选项卡显示SQL映射名称、BlockCount(作为正整数)和BlockCount的来源。
块计数的来源可以在类定义中定义、由类编译器估计或由TuneTable度量。
将类编译器估计的调优表更改运行到TuneTable测量;
它不影响在类定义中定义的值。

通过从调优表显示中选择单个SQL映射名称，可以修改BlockCount计算值。
这将在显示器右侧的详细信息区域中显示该地图名称的块计数。
可以将块计数修改为一个更适合预期的完整数据集的值。
因为在设置该值时，Tune Table不执行验证，所以应该确保块计数是一个有效值。
修改BlockCount会将BlockCount的来源更改为类定义中定义的。

导出和重新导入调优表统计信息

可以从一个表或一组表导出调优表统计信息，然后将这些调优表统计信息导入一个表或一组表。
以下是可能希望执行此导出/导入的三种情况。
(为简单起见，这些描述了从单个表导出/导入统计数据;
在实际使用中，通常会从多个相互关联的表中导出/导入统计数据):
- 为生产系统建模:生产表完全填充了实际数据，并使用Tune table进行优化。
在测试环境中，创建的表具有相同的表定义，但数据少得多。
通过从生产表导出调优表统计信息并将它们导入测试表，可以在测试表上对生产表优化建模。
- 要复制生产系统：生产表完全填充了实际数据，并使用tune Table进行了优化。将创建具有相同表定义的第二个生产表。(例如，生产环境及其备份环境，或者多个相同的表定义，每个表包含不同医院的患者记录。)。通过从第一个表导出调优表统计信息并将其导入第二个表，您可以为第二个表提供与第一个表相同的优化，而无需第二次运行调优表或等待第二个表填充有代表性的数据。
- 要恢复到以前的统计信息集：可以通过运行tune Table或显式设置统计信息来创建表的优化统计信息。通过导出这些统计信息，可以在尝试其他统计信息设置时保留它们。一旦确定了最佳统计信息集，就可以将它们重新导入到表中

。

可以使用$SYSTEM.SQL.Stats.Table.Export()方法将调优表统计信息导出到XML文件。此方法可以导出名称空间中一个、多个或所有表的优化表统计信息，如以下示例所示：

```
DO $SYSTEM.SQL.Stats.Table.Export("C:\AllStats.xml")
/* ?????????????/??TuneTable???? */


DO $SYSTEM.SQL.Stats.Table.Export("C:\SampleStats.xml","Sample")
/*  ??Sample???????????? */


DO $SYSTEM.SQL.Stats.Table.Export("C:\SamplePStats.xml","Sample","P*")
/* ??Sample????????“P”??????????? */


DO $SYSTEM.SQL.Stats.Table.Export("C:\SamplePersonStats.xml","Sample","Person")
/*  ??Sample???????Person */
```

可以使用$SYSTEM.SQL.Stats.Table.Import()方法重新导入使用$SYSTEM.SQL.Stats.Table.Import()方法导出的调优表统计信息。

$SYSTEM.SQL.Stats.Table.Import()有一个KeepClassUpToDate boolean选项。
如果为真(并且update为真)，$SYSTEM.SQL.Stats.Table.Import()将用新的EXTENTSIZE和选择性值更新类定义，但类定义将保持最新。
但是，在许多情况下，最好在调优了类表之后重新编译类，这样类定义中的查询就可以重新编译，SQL查询优化器就可以使用更新后的数据统计信息。
默认值为FALSE(0)。请注意，如果该类已部署，则类定义不会更新。

$SYSTEM.SQL.Stats.Table.Import()有一个ClearCurrentStats boolean选项。
如果为TRUE， $SYSTEM.SQL.Stats.Table.Import()将在导入统计信息之前从现有表中清除所有先前的区段大小、选择性、块计数和其他调优表统计信息。
如果您想要完全清除导入文件中没有指定的那些表状态，而不是让它们在表的persistent类中定义，则可以使用此方法。
默认值是FALSE(0)。

如果$SYSTEM.SQL.Stats.Table.Import()没有找到相应的表，它将跳过该表并继续导入文件中指定的下一个表。
如果找到了一个表，但是没有找到一些字段，那么这些字段将被跳过。

无法继承类存储定义中映射的BlockCount。BlockCount只能出现在映射起源的类的存储定义中。如果映射源自超类，则$SYSTEM.SQL.Stats.Table.Import()仅设置投影表的BlockCount元数据，而不设置类存储BlockCount元数据。

#SQL #Caché #InterSystems IRIS #InterSystems IRIS for Health

源URL:
https://cn.community.intersystems.com/post/%E7%AC%AC%E4%B8%89%E7%AB%A0-%E4%BC%98%E5%8C%96%E8%A1%A8%EF%BC%88%E4%BA%8C%EF%BC%89