記事
Toshihiko Minamoto · 2022年3月2日 5m read

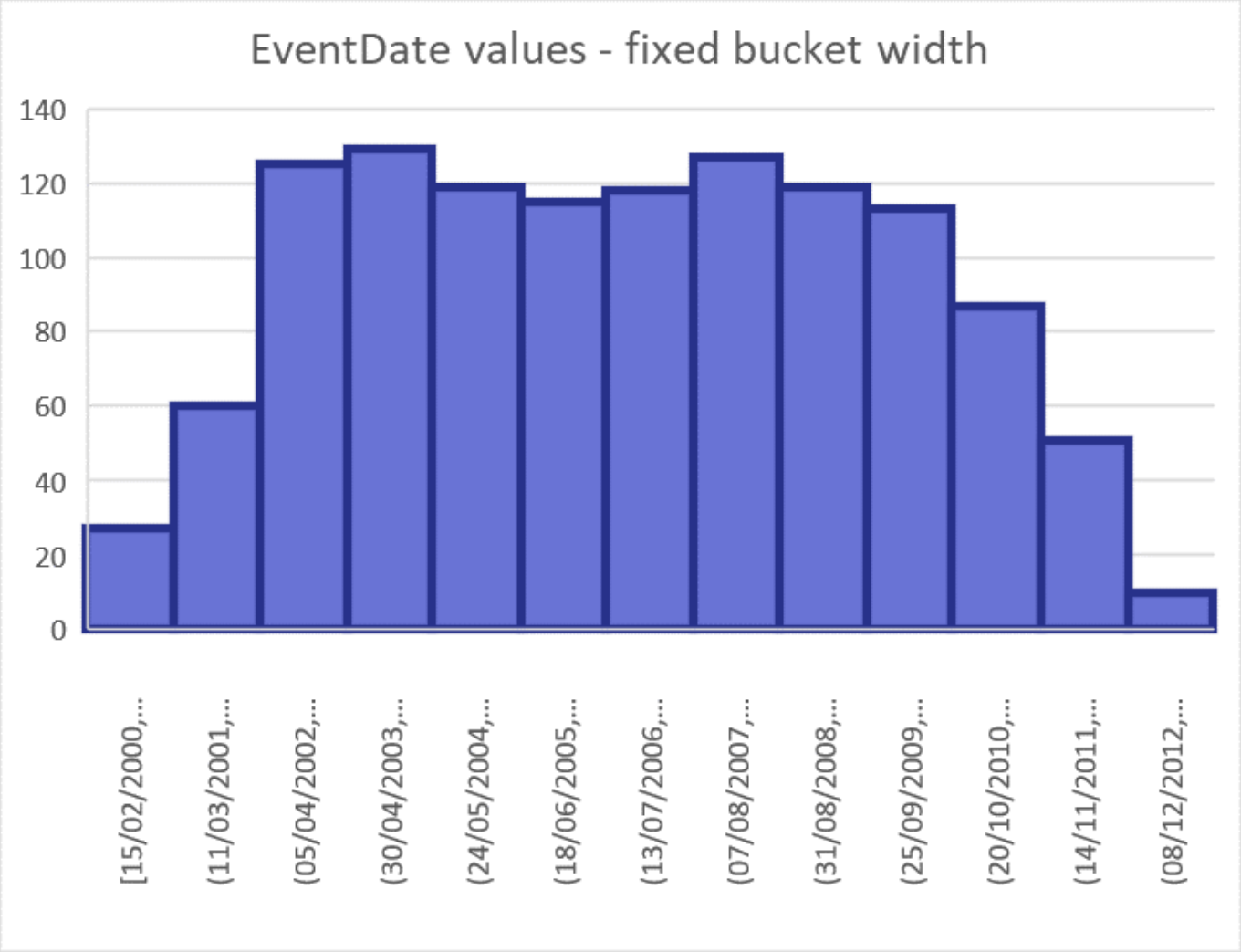
2021.2 SQL 機能スポットライト - 高度なテーブル統計

これは、IRIS でリレーショナルデータをクエリするアナリストとアプリケーションに、さらに優れた適応性とパフォーマンスによるエクスペリエンスを提供する IRIS SQL のイノベーションをトピックとした短い連載の 3 つ目の記事です。 2021.2
では連載の最後の記事になるかもしれませんが、この分野ではさらにいくつかの機能強化が行われています。
この記事では、このリリースで収集し始めた**ヒストグラム**
という追加のテーブル統計について、もう少し詳しく説明します。

ヒストグラムとは？

ヒストグラムは数値フィールド（またはより広範には、厳密な順序を持つデータ）のデータ分布の近似表現です。
このようなフィールドの最小値、最大値、および平均値がわかれば役立ちますが、データが 3
つのポイント間でどのように分布しているかはほとんどわかりません。 ここで役立つのがヒストグラムです。値の範囲をバケットに分割し、バケットごとに出現するフィールド値の数をカウントします。

これは非常に柔軟な定義であるため、バケットがフィールド値に関して同じ「幅」になるように、またはカバーされるサンプル値の数に関して同じ「大きさ」になるように、バケットのサイズを選択することができます。
後者の場合、各バケットには同じパーセンテージの値が含まれるため、バケットはパーセンタイルを表します。
以下のグラフは、日数で表現された同じバケット幅を使用して、Aviation Demo データセットの EventData
フィールドのヒストグラムをプロットしています。



ヒストグラムが必要な理由

カリフォルニア州で 2004
年より前のすべてのイベントについて、このデータセットのクエリを実行しているとします。

```
SELECT * FROM Aviation.Event WHERE EventDate &lt; '2004-05-01' AND LocationCountry =
'California'
```

「ランタイムプランの選択」という前の記事では、テーブル統計で LocationCounty
のようなフィールドの選択性と潜在的な外れ値をキャプチャする方法についてすでに説明しています。
しかし、そのような個別のフィールド値の統計は、EventDate での < 条件ではあまり実用的ではありません。
この条件の選択制を計算するには、2004 年 5 月 1 日までのすべての潜在的な EventDate 値の選択制を集計する必要があり、クエリのプランニング時に行えるような手っ取り早い見積もりではなく、それだけで非常に厳しいクエリとなる可能性があります。 ここで使用できるのがヒストグラムです。

EventDate 値の分布のヒストグラムデータを見てみましょう。今回は、データを同じサイズの 16
個のバケットに分割し、各バケットには 6.667% のデータが保持されています。
このようにすると、クエリコストの見積もりに使用できるパーセンタイルと選択制の数値に簡単に変換できます。
このテーブルを読み取るために、4 行目を見てみましょう。値の 20%（各 6.667% の 3
つのバケット）がこのバケットの下限である 2003 年 6 月 22 日より前にあり、さらに 6.667% の値が 2003 年 9 月 19 日まで保持されています。
<colgroup><col style="width:48pt" width="64"><col style="width:61pt" width="81"><col style="width:64pt" width="85"></colgroup>

Bucket	Percentile	Value
	0%	21/12/2001

1	7%	02/07/2002
2	13%	19/01/2003
3	20%	22/06/2003
4	27%	19/09/2003
5	33%	30/12/2003
6	40%	01/10/2004
7	47%	01/10/2005
8	53%	20/08/2006
9	60%	14/01/2007
10	67%	02/04/2008
11	73%	14/05/2008
12	80%	29/11/2008
13	87%	01/06/2010
14	93%	30/10/2011
15	100%	26/09/2012

上記のクエリ例で使用されているカットオフ日（2004 年 5 月 1 日）は、5
番目のバケットにあり、その日付より前には 33% から 40% の値があります。
バケットが小さくなるにつれ、その中
の分布はほぼ均一であると見なすことができ、下限と上限の間を単に補完することができます。つまり、この場合、選択性は約 37% となり、これをクエリコストの見積もりに使用することができます。

ヒストグラムの使用を可視化するには、もう一つ、累積分布グラフとしてプロットする方法があります。 X 軸で 2004 年 5 月 1 日の線（値）がどのように描かれるかを確認すれば、Y 軸で 約 37% と解釈できます。

上記の例では、わかりやすくするために上限のみの範囲条件を使用していますが、このアプローチは、下限または間隔条件（BETWEEN 句を使用するなど）を使用しても当然動作します。

2021.2 より、文字列を含むすべての照合フィールドのテーブル統計の一環としてヒストグラムを収集しており、それを使用して RTPC の一部として範囲選択性を推定することができるようになっています。
実世界での多くのクエリには日付（およびその他の）フィールドでの範囲条件が伴うため、この IRIS SQL の機能強化によって、多くのお客様のクエリプランに役立つと信じています。いつものように、皆さんの体験をお聞かせください。

#SQL #リレーショナルテーブル #InterSystems IRIS

ソースURL:
https://jp.community.intersystems.com/post/20212-sql-%E6%A9%9F%E8%83%BD%E3%82%B9%E3%83%9D%E3%83%83%E3%83%88%E3%83%A9%E3%82%A4%E3%83%88-%E9%AB%98%E5%BA%A6%E3%81%AA%E3%83%86%E3%83%BC%E3%83%96%E3%83%AB%E7%B5%B1%E8%A8%88