Vertica 8.0.1 の新機能: Outlier Detection（外れ値検出）

原文はこちら

Vertica 8.0.1 では、DETECT_OUTLIERS という関数を追加しました。この関数を使用すると、指定した閾値から外れるデータポイントを識別できます。一般に、データを分析する前に、データセットから外れ値を削除しておく必要があります。外れ値は、他のデータポイントとは大きく異なります。データセットの中に外れ値を残しておくと、結果が歪んでしまう可能性があります。

また、このブログの後半で説明するように、これらの外れ値の影響を z スコアの計算に反映させることもできます。

外れ値の検出方法：

```
DETECT_OUTLIERS ( 'output_view', 'input_table',

                  'input_columns', 'outlier_method'
                  [,
                  [-outlier_threshold=value]
                  [-exclude_columns= " col1, col2, … coln " ]
                  [-key_columns= " col1, col2, … coln " ]
                  ']
                 )
```

複数の属性の外れ値を検出できます。たとえば、時間、ヒット数、給与データの外れ値を検出したいとします。

```
=> CREATE TABLE baseball (id identity, first_name varchar(50), last_name
varchar(50), dob DATE,
team varchar(20), hr INT, hits INT, avg FLOAT, salary FLOAT);

=> SELECT DETECT_OUTLIERS('baseball_hr_hits_salary_outliers', 'baseball',
'hr, hits, salary', 'robust_zscore',
'-outlier_threshold=3.0 -key_columns="id, team"');

DETECT_OUTLIERS
----------
Finished in 1 iteration
(1 row)
```

baseball_hr_hits_salary_outliers テーブルに存在する外れ値のデータがリストされたので、それらの値を除外したビューを作成できます。この「クリーンな」ビューを使用して分析を行います。

```
=> CREATE VIEW clean_baseball AS
SELECT * FROM baseball WHERE id NOT IN (SELECT id FROM
baseball_hr_hits_salary_outliers);
CREATE VIEW
```

ロバスト z スコア

これが以下の小さなサンプルデータセットでどのように機能するかを見てみましょう。Id 列の 1 と 3 のデータには他の値と大きく異なる値が含まれています。

```
VMart=> SELECT * FROM normDataSet;
 id |    c2     |   c3
----+-----------+-------
  2 |  3434673  |    223
  3 |  9994673  |  54367
  4 |  6286743  |  12583
  1 |   343453  |  23423
  5 |  5672345  |   2456
(5 rows)
```

Mean: 5146377.4

Median: 5672345

Standard deviation: 3573550.77577

通常の Z スコア正規化法を使用すると、各列の値を、各列の平均からの観測値である標準偏差の数に正規化することができます。これにより、データを正規分布の確率変数と比較することができます。

正規化された列の任意の値は、その列の平均からの標準偏差の数になります。

ロバスト z スコア法を使用すると、列の値はその列の中央値からの標準偏差の数になります。平均値の代わりに中央値を使用すると、データの外れ値の影響を取り除くのに役立ちます。

```
VMart=> SELECT NORMALIZE('balance_norm_zscore' , 'normDataSet' , 'c2, c3' ,
                         'zscore' , '--key_columns=id');

         normalize
--------------------------
 Finished in 1 iteration
(1 row)

VMart=> SELECT * FROM balance_norm_zscore;
 id |         c2          |         c3
----+---------------------+---------------------
  2 | -0.478992606347343  | -0.835556539176108
  3 |   1.35671658364981  |   1.62484423837543
  4 |   0.3191127457129   | -0.273895900683624
  1 |  -1.34402019206427  |  0.218693202978068
  5 |  0.147183469048895  | -0.734085001493769
(5 rows)
```

Number of standard deviations from the *mean* of column c3

詳細については、Vertica ドキュメントの DETECT_OUTLIERS を参照してください。