

03.12-Performance-Eval-and-Query

November 10, 2024

1 High-Performance Pandas: eval and query

As we've already seen in previous chapters, the power of the PyData stack is built upon the ability of NumPy and Pandas to push basic operations into lower-level compiled code via an intuitive higher-level syntax: examples are vectorized/broadcasted operations in NumPy, and grouping-type operations in Pandas. While these abstractions are efficient and effective for many common use cases, they often rely on the creation of temporary intermediate objects, which can cause undue overhead in computational time and memory use.

To address this, Pandas includes some methods that allow you to directly access C-speed operations without costly allocation of intermediate arrays: `eval` and `query`, which rely on the [NumExpr package](#). In this chapter I will walk you through their use and give some rules of thumb about when you might think about using them.

1.1 Motivating query and eval: Compound Expressions

We've seen previously that NumPy and Pandas support fast vectorized operations; for example, when adding the elements of two arrays:

```
[1]: import numpy as np
      rng = np.random.default_rng(42)
      x = rng.random(1000000)
      y = rng.random(1000000)
      %timeit x + y
```

2.21 ms \pm 142 μ s per loop (mean \pm std. dev. of 7 runs, 100 loops each)

As discussed in [Computation on NumPy Arrays: Universal Functions](#), this is much faster than doing the addition via a Python loop or comprehension:

```
[2]: %timeit np.fromiter((xi + yi for xi, yi in zip(x, y)),
                        dtype=x.dtype, count=len(x))
```

263 ms \pm 43.4 ms per loop (mean \pm std. dev. of 7 runs, 1 loop each)

But this abstraction can become less efficient when computing compound expressions. For example, consider the following expression:

```
[3]: mask = (x > 0.5) & (y < 0.5)
```

Because NumPy evaluates each subexpression, this is roughly equivalent to the following:

```
[4]: tmp1 = (x > 0.5)
      tmp2 = (y < 0.5)
      mask = tmp1 & tmp2
```

In other words, *every intermediate step is explicitly allocated in memory*. If the `x` and `y` arrays are very large, this can lead to significant memory and computational overhead. The NumExpr library gives you the ability to compute this type of compound expression element by element, without the need to allocate full intermediate arrays. The [NumExpr documentation](#) has more details, but for the time being it is sufficient to say that the library accepts a *string* giving the NumPy-style expression you'd like to compute:

```
[5]: import numexpr
      mask_numexpr = numexpr.evaluate('(x > 0.5) & (y < 0.5)')
      np.all(mask == mask_numexpr)
```

```
[5]: True
```

The benefit here is that NumExpr evaluates the expression in a way that avoids temporary arrays where possible, and thus can be much more efficient than NumPy, especially for long sequences of computations on large arrays. The Pandas `eval` and `query` tools that we will discuss here are conceptually similar, and are essentially Pandas-specific wrappers of NumExpr functionality.

1.2 pandas.eval for Efficient Operations

The `eval` function in Pandas uses string expressions to efficiently compute operations on `DataFrame` objects. For example, consider the following data:

```
[6]: import pandas as pd
      nrows, ncols = 100000, 100
      df1, df2, df3, df4 = (pd.DataFrame(rng.random((nrows, ncols)))
                             for i in range(4))
```

To compute the sum of all four `DataFrames` using the typical Pandas approach, we can just write the sum:

```
[7]: %timeit df1 + df2 + df3 + df4
```

73.2 ms ± 6.72 ms per loop (mean ± std. dev. of 7 runs, 10 loops each)

The same result can be computed via `pd.eval` by constructing the expression as a string:

```
[8]: %timeit pd.eval('df1 + df2 + df3 + df4')
```

34 ms ± 4.2 ms per loop (mean ± std. dev. of 7 runs, 10 loops each)

The `eval` version of this expression is about 50% faster (and uses much less memory), while giving the same result:

```
[9]: np.allclose(df1 + df2 + df3 + df4,
                  pd.eval('df1 + df2 + df3 + df4'))
```

[9]: True

`pd.eval` supports a wide range of operations. To demonstrate these, we'll use the following integer data:

```
[10]: df1, df2, df3, df4, df5 = (pd.DataFrame(rng.integers(0, 1000, (100, 3)))
                                for i in range(5))
```

Arithmetic operators `pd.eval` supports all arithmetic operators. For example:

```
[11]: result1 = -df1 * df2 / (df3 + df4) - df5
      result2 = pd.eval('-df1 * df2 / (df3 + df4) - df5')
      np.allclose(result1, result2)
```

[11]: True

Comparison operators `pd.eval` supports all comparison operators, including chained expressions:

```
[12]: result1 = (df1 < df2) & (df2 <= df3) & (df3 != df4)
      result2 = pd.eval('(df1 < df2 <= df3 != df4)')
      np.allclose(result1, result2)
```

[12]: True

Bitwise operators `pd.eval` supports the `&` and `|` bitwise operators:

```
[13]: result1 = (df1 < 0.5) & (df2 < 0.5) | (df3 < df4)
      result2 = pd.eval('(df1 < 0.5) & (df2 < 0.5) | (df3 < df4)')
      np.allclose(result1, result2)
```

[13]: True

In addition, it supports the use of the literal `and` and `or` in Boolean expressions:

```
[14]: result3 = pd.eval('(df1 < 0.5) and (df2 < 0.5) or (df3 < df4)')
      np.allclose(result1, result3)
```

[14]: True

Object attributes and indices `pd.eval` supports access to object attributes via the `obj.attr` syntax and indexes via the `obj[index]` syntax:

```
[15]: result1 = df2.T[0] + df3.iloc[1]
      result2 = pd.eval('df2.T[0] + df3.iloc[1]')
      np.allclose(result1, result2)
```

[15]: True

Other operations Other operations, such as function calls, conditional statements, loops, and other more involved constructs are currently *not* implemented in `pd.eval`. If you'd like to execute these more complicated types of expressions, you can use the NumExpr library itself.

1.3 DataFrame.eval for Column-Wise Operations

Just as Pandas has a top-level `pd.eval` function, `DataFrame` objects have an `eval` method that works in similar ways. The benefit of the `eval` method is that columns can be referred to by name. We'll use this labeled array as an example:

```
[16]: df = pd.DataFrame(rng.random((1000, 3)), columns=['A', 'B', 'C'])
      df.head()
```

```
[16]:
```

	A	B	C
0	0.850888	0.966709	0.958690
1	0.820126	0.385686	0.061402
2	0.059729	0.831768	0.652259
3	0.244774	0.140322	0.041711
4	0.818205	0.753384	0.578851

Using `pd.eval` as in the previous section, we can compute expressions with the three columns like this:

```
[17]: result1 = (df['A'] + df['B']) / (df['C'] - 1)
      result2 = pd.eval("(df.A + df.B) / (df.C - 1)")
      np.allclose(result1, result2)
```

```
[17]: True
```

The `DataFrame.eval` method allows much more succinct evaluation of expressions with the columns:

```
[18]: result3 = df.eval('(A + B) / (C - 1)')
      np.allclose(result1, result3)
```

```
[18]: True
```

Notice here that we treat *column names as variables* within the evaluated expression, and the result is what we would wish.

1.3.1 Assignment in DataFrame.eval

In addition to the options just discussed, `DataFrame.eval` also allows assignment to any column. Let's use the `DataFrame` from before, which has columns 'A', 'B', and 'C':

```
[19]: df.head()
```

```
[19]:
```

	A	B	C
0	0.850888	0.966709	0.958690
1	0.820126	0.385686	0.061402

```

2  0.059729  0.831768  0.652259
3  0.244774  0.140322  0.041711
4  0.818205  0.753384  0.578851

```

We can use `df.eval` to create a new column 'D' and assign to it a value computed from the other columns:

```
[20]: df.eval('D = (A + B) / C', inplace=True)
      df.head()
```

```
[20]:
```

	A	B	C	D
0	0.850888	0.966709	0.958690	1.895916
1	0.820126	0.385686	0.061402	19.638139
2	0.059729	0.831768	0.652259	1.366782
3	0.244774	0.140322	0.041711	9.232370
4	0.818205	0.753384	0.578851	2.715013

In the same way, any existing column can be modified:

```
[21]: df.eval('D = (A - B) / C', inplace=True)
      df.head()
```

```
[21]:
```

	A	B	C	D
0	0.850888	0.966709	0.958690	-0.120812
1	0.820126	0.385686	0.061402	7.075399
2	0.059729	0.831768	0.652259	-1.183638
3	0.244774	0.140322	0.041711	2.504142
4	0.818205	0.753384	0.578851	0.111982

1.3.2 Local Variables in DataFrame.eval

The `DataFrame.eval` method supports an additional syntax that lets it work with local Python variables. Consider the following:

```
[22]: column_mean = df.mean(1)
      result1 = df['A'] + column_mean
      result2 = df.eval('A + @column_mean')
      np.allclose(result1, result2)
```

```
[22]: True
```

The `@` character here marks a *variable name* rather than a *column name*, and lets you efficiently evaluate expressions involving the two “namespaces”: the namespace of columns, and the namespace of Python objects. Notice that this `@` character is only supported by the `DataFrame.eval` method, not by the `pandas.eval` function, because the `pandas.eval` function only has access to the one (Python) namespace.

1.4 The DataFrame.query Method

The `DataFrame` has another method based on evaluated strings, called `query`. Consider the following:

```
[23]: result1 = df[(df.A < 0.5) & (df.B < 0.5)]
      result2 = pd.eval('df[(df.A < 0.5) & (df.B < 0.5)]')
      np.allclose(result1, result2)
```

[23]: True

As with the example used in our discussion of `DataFrame.eval`, this is an expression involving columns of the `DataFrame`. However, it cannot be expressed using the `DataFrame.eval` syntax! Instead, for this type of filtering operation, you can use the `query` method:

```
[24]: result2 = df.query('A < 0.5 and B < 0.5')
      np.allclose(result1, result2)
```

[24]: True

In addition to being a more efficient computation, compared to the masking expression this is much easier to read and understand. Note that the `query` method also accepts the `@` flag to mark local variables:

```
[25]: Cmean = df['C'].mean()
      result1 = df[(df.A < Cmean) & (df.B < Cmean)]
      result2 = df.query('A < @Cmean and B < @Cmean')
      np.allclose(result1, result2)
```

[25]: True

1.5 Performance: When to Use These Functions

When considering whether to use `eval` and `query`, there are two considerations: *computation time* and *memory use*. Memory use is the most predictable aspect. As already mentioned, every compound expression involving NumPy arrays or Pandas `DataFrames` will result in implicit creation of temporary arrays. For example, this:

```
[26]: x = df[(df.A < 0.5) & (df.B < 0.5)]
```

is roughly equivalent to this:

```
[27]: tmp1 = df.A < 0.5
      tmp2 = df.B < 0.5
      tmp3 = tmp1 & tmp2
      x = df[tmp3]
```

If the size of the temporary `DataFrames` is significant compared to your available system memory (typically several gigabytes), then it's a good idea to use an `eval` or `query` expression. You can check the approximate size of your array in bytes using this:

```
[28]: df.values.nbytes
```

```
[28]: 32000
```

On the performance side, `eval` can be faster even when you are not maxing out your system memory. The issue is how your temporary objects compare to the size of the L1 or L2 CPU cache on your system (typically a few megabytes); if they are much bigger, then `eval` can avoid some potentially slow movement of values between the different memory caches. In practice, I find that the difference in computation time between the traditional methods and the `eval/query` method is usually not significant—if anything, the traditional method is faster for smaller arrays! The benefit of `eval/query` is mainly in the saved memory, and the sometimes cleaner syntax they offer.

We’ve covered most of the details of `eval` and `query` here; for more information on these, you can refer to the Pandas documentation. In particular, different parsers and engines can be specified for running these queries; for details on this, see the discussion within the [“Enhancing Performance”](#) section of the documentation.