User Defined Functions for Amazon Redshift

by Jeff Barr | on 11 SEP 2015 | in Amazon Redshift* | Permalink | Share

The Amazon Redshift team is on a tear. They are listening to customer feedback and rolling out new features all the time! Below you will find an announcement of another powerful and highly anticipated new feature.

— Jeff;

https://aws.amazon.com/blogs/aws/user-defined-functions-for-amazon-redshift/

Amazon Redshift makes it easy to launch a petabyte-scale data warehouse. For less than $1,000/Terabyte/year, you can focus on your analytics, while Amazon Redshift manages the infrastructure for you. Amazon Redshift's price and performance has allowed customers to unlock diverse analytical use cases to help them understand their business. As you can see from blog posts by Yelp, Amplitude and Cake, our customers are constantly pushing the boundaries of what's possible with data warehousing at scale.

To extend Amazon Redshift's capabilities even further and make it easier for our customers to drive new insights, I am happy to announce that Amazon Redshift has added scalar user-defined functions (UDFs). Using PostgreSQL syntax, you can now create scalar functions in Python 2.7 custom-built for your use case, and execute them in parallel across your cluster.

Here's a template that you can use to create your own functions:

```
CREATE [ OR REPLACE ] FUNCTION f_function_name
( [ argument_name arg_type, ... ] )
RETURNS data_type
{ VOLATILE | STABLE | IMMUTABLE }
AS $$
  python_program
$$ LANGUAGE plpythonu;
```

Scalar UDFs return a single result value for each input value, similar to built-in scalar functions such as `ROUND` and `SUBSTRING`. Once defined, you can use UDFs in any SQL statement, just as you would use our built-in functions.

In addition to creating your own functions, you can take advantage of thousands of functions available through Python libraries to perform operations not easily expressed in SQL. You can even add custom libraries directly from S3 and the web. Out of the box, Amazon Redshift UDFs come integrated with the Python Standard Library and a number of other libraries, including:

- NumPy and SciPy, which provide mathematical tools you can use to create multi-dimensional objects, do matrix operations, build optimization algorithms, and run statistical analyses.
- Pandas, which offers high level data manipulation tools built on top of NumPy and SciPy, and that enables you to perform data analysis or an end-to-end modeling workflow.
- Dateutil and Pytz, which make it easy to manipulate dates and time zones (such as figuring out how many months are left before the next Easter that occurs in a leap year).

UDFs can be used to simplify complex operations. For example, if you wanted to extract the hostname out of a URL, you could use a regular expression such as:

```
SELECT REGEXP_REPLACE(url, '(https?)://([^@]*@)?([^:/]*)([/:].*|$)', '\3') FROM
table;
```

Or, you could import a Python URL parsing library, URLParse, and create a function that extracts hostnames:

```
CREATE FUNCTION f_hostname(url VARCHAR)
RETURNS varchar
IMMUTABLE AS $$
import urlparse
return urlparse.urlparse(url).hostname
$$ LANGUAGE plpythonu;
```

Now, in SQL all you have to do is:

```
SELECT f_hostname(url)
FROM table;
```

As our customers know, Amazon Redshift obsesses about security. We run UDFs inside a restricted container that is fully isolated. This means UDFs cannot corrupt your cluster or negatively impact its performance. Also, functions that write files or access the network are not supported. Despite being tightly managed, UDFs leverage Amazon Redshift's MPP capabilities, including being executed in parallel on each node of your cluster for optimal performance.

To learn more about creating and using UDFs, please see our documentation and a detailed post on the AWS Big Data blog. Also, check out this how-to guide from APN Partner Looker. If you'd like to share the UDFs you've created with other Amazon Redshift customers, please

reach out to us at redshift-feedback@amazon.com. APN Partner Periscope has already created a number of useful scalar UDFs and published them here.

We will be patching your cluster with UDFs over the next two weeks, depending on your region and maintenance window setting. The new cluster version will be 1.0.991. We know you've been asking for UDFs for some time and would like to thank you for your patience. We look forward to hearing from you about your experience at redshift-feedback@amazon.com.

— Tina Adams, Senior Product Manager

Jeff Barr

Jeff Barr is Chief Evangelist for AWS. H