Supercharge cybersecurity with Netflow and Yellowbrick Data

Effective cybersecurity hinges on automating risk detection, using predictive analytics to measure the risk of real-time activity against aggregated historical patterns But data volumes are now so massive that historical data is regularly archived to create space for new data, which means that less historical data is being analyzed quickly, so end results are less accurate or less timely. Enterprises are struggling for ways to collect Netflow data in near real time and analyze it against billions or trillions of rows of historical data without making a science project out of it. Enter Yellowbrick Data.

Yellowbrick offers a simple solution to the massive scale and latency requirements associated with production level Netflow analysis. Utilizing the latest hardware and software approaches, Yellowbrick can ingest Netflow and other data sets at line rate while still serving queries and other application needs. Yellowbrick can make the data available immediately and there are no indexes to create and manage. 2PB data sets are easily ingested and managed with a single Yellowbrick cluster occupying only 10U of rack space. Following are some key benefits.

Query IP data quickly and efficiently

The Yellowbrick database treats IP addresses as their own type, similar to a number, and supports order and functions. This enables organizations to query ranges with greater than, less than, or between operators without the need to convert data types. For example:

```
select count(*) from netflow_ip where destination_ip >
‘200.200.200.1’
```

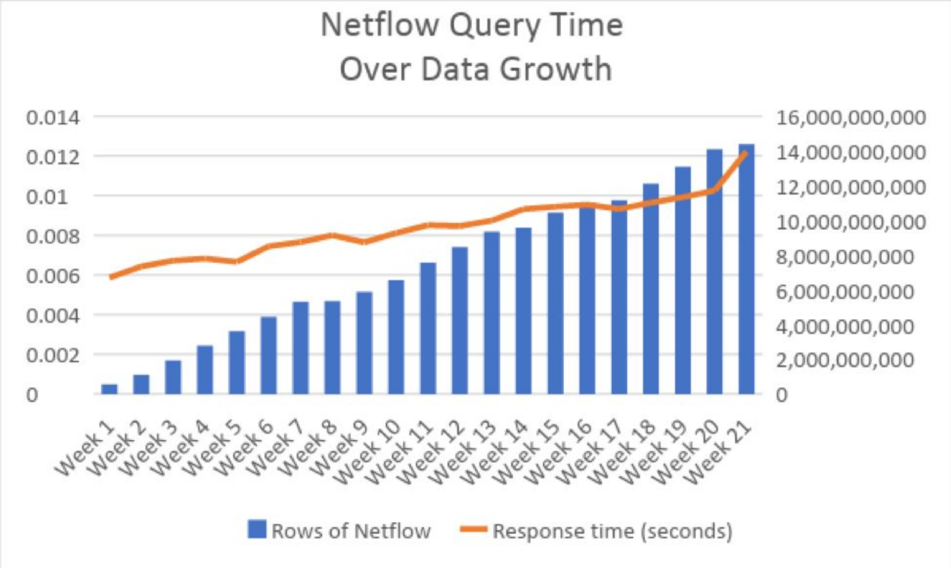
Rapidly ingest, scan, and access data

Yellowbrick can perform table scans at hundreds of gigabytes per second. A million-row table scan is a sub-second activity for Yellowbrick. Some other ways Yellowbrick improves data analytics is by enabling users to sort data into columns. Even after the initial sort, sorting continues as new data is added or inserted. By sorting IP addresses in order, organizations can significantly accelerate IP address lookups within a particular range. Yellowbrick also uses zone mapping to tell the database which blocks of the NVMe flash contain which ranges of data. This significantly accelerates scan times by reducing the amount of data that must be scanned. Yellowbrick even displays read efficiency to users to help with data layout and query planning.

Scale easily to massive data sets

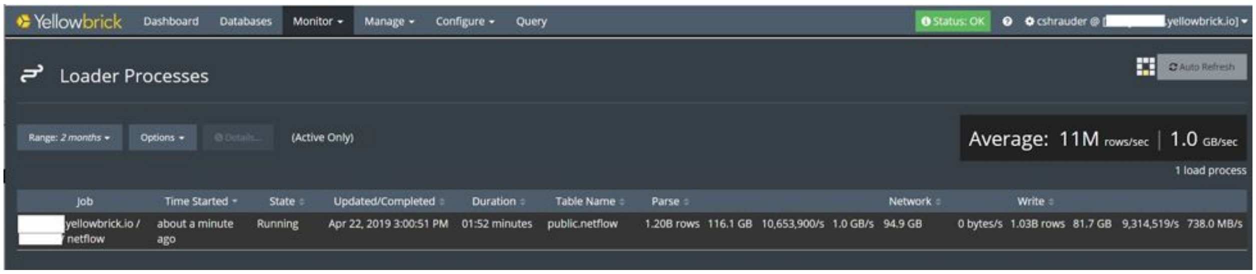
The figure below shows the results of a query response time test as data set size grew from 500 million records up to 14 billion records. Querying 28x more records took just double the time.

Figure 1. Querying a 28x a larger dataset size took just double the time (.0122 seconds).



Load data at 10-million rows per second

Enterprise cybersecurity also struggles with loading massive volumes of data into the system to begin with. With Yellowbrick, organizations can achieve load times of ~150,000 rows a second using standard ODBC, JDBC, and ADO.net connectors. Using the Yellowbrick high-speed tool, ybload, organizations can load data directly over a 10G network, achieving load speeds of 10 million rows per second (about 1 GB/s). In the example below, a 308GB dataset loaded and was available for querying in about 5 minutes and 15 seconds.



Average: 11M rows/sec | 1.0 GB/sec

1 load process

Job	Time Started	State	Updated/Completed	Duration	Table Name	Parse	Network	Write
[redacted] yellowbrick.io / [redacted] netflow	about a minute ago	Running	Apr 22, 2019 3:00:51 PM	01:52 minutes	public.netflow	1.20B rows 116.1 GB 10,653,900/s 1.0 GB/s	94.9 GB 0 bytes/s	1.03B rows 81.7 GB 9,314,519/s 738.0 MB/s

Sound too good to be true? Seeing is believing. Contact a Yellowbrick Data representative to set up a demo in your own environment: federal-info@yellowbrick.com.