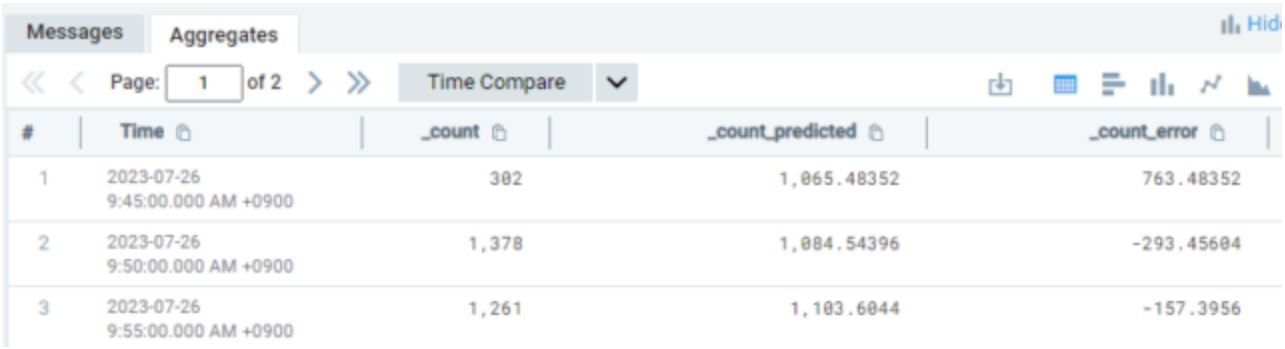
I used Sumo Logic to create a graph that predicts future traffic based on the number of traffic to a web server

dev.classmethod.jp/articles/sumo-logic-20230725-sakumashogo

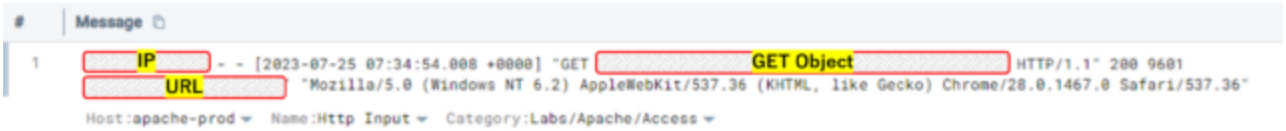
佐久間昇吾

July 26, 2023



Predicting future values

`_sourceCategory=Labs/Apache/Access`Search for the following. The following log was output.



*Although some key points have been masked, the data from the hidden parts will not be used this time.

Now we will process this log and analyze the data.

Use the predict operator to predict future values based on the number of accesses

The predict operator is an operator that predicts future trends in log data volume based on the amount of log data aggregated over a given period.

There are two predictive models. The first is called **Linear Regression** , which uses data from a specific period of time to plot it as a graph based on a linear regression model (described below). The second is called **Auto-Regressive** , which learns the patterns of logs output within a specific period of time in the past and converts them into data. Predictions are made based on this learned data.

Second, the mandatory requirement is that you must use **aggregation operators (count, min, max, sum)** and **the timeslice operator .**

I'll actually write it out so that you can get a concrete idea of what I've achieved so far.

① Prediction model: Linear regression

```
_sourceCategory=Labs/Apache/Access "HTTP/1.1"
| timeslice 5m
| count by _timeslice
| predict _count by 5m
```

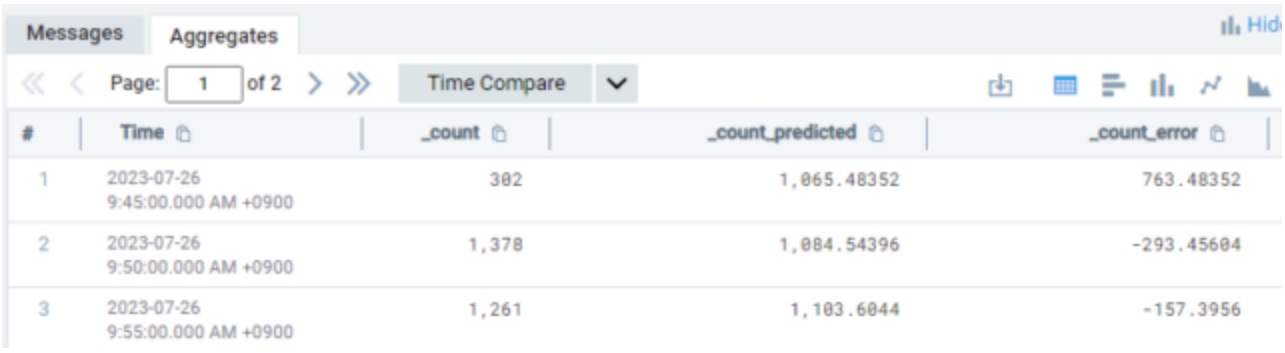
The first line searches for keywords containing the string `"HTTP/1.1"`in the part `HTTP/1.1`. This will exclude logs that do not contain the string HTTP/1.1. In the log output result above, the 200 after HTTP/1.1 is the HTTP status code, so a keyword search was performed to target logs that contain the status.

The second line uses the timeslice operator to divide the search results from the first line into 5-minute intervals.

The third line `_timeslice`counts the number of fields that are separated by 5 minutes on the second line. The underscored fields `_timeslice`are fields that are implicitly generated when using timeslice as on the second line. The contents of the fields are timeslice data.

The prediction operator predict is used in line 4. `_timeslice`The number counted in line 3 is subjected to linear regression analysis at 5-minute intervals.

The output is shown below. _count `_timeslice`shows the number of data points for each 5-minute interval. `_count_predicted`is the value predicted by the linear regression model. `_count_error`is the value obtained by subtracting `_count_predicted`the value of `_count`by the value of . These three values are used to represent the graph.



#	Time	_count	_count_predicted	_count_error
1	2023-07-26 9:45:00.000 AM +0900	302	1,065.48352	763.48352
2	2023-07-26 9:50:00.000 AM +0900	1,378	1,084.54396	-293.45604
3	2023-07-26 9:55:00.000 AM +0900	1,261	1,103.6044	-157.3956

Linear regression is an estimation that minimizes the gap between the calculated data value and the predicted value.

This graph was created at 11:30 on 2023/07/26, so the bar graphs displayed after 11:30 `_count_predicted`are predicted values.

② Prediction model: Auto-regressive

```
_sourceCategory=Labs/Apache/Access "HTTP/1.1"
| timeslice 5m
| count by _timeslice
| predict _count by 5m model=ar, ar.window=150, forecast=5
```

Lines 1 to 3 are the same as Linear regression.

In the fourth line, `model=ar`, `ar.window=150`, `forecast=5`is added.

`model=ar`is a parameter that declares the use of Auto-regressive.

`ar.window=150`is an optional option. Select the number of time series data to use for forecasting. The forecast will be made based on this data. If not set, the number required for the forecast will be automatically assigned implicitly.

`forecast=5`is an optional option. It specifies the range of data to be forecast. `forecast=5`If it is set, it will forecast 5 data points in the future. `forecast=5m`If it is set, it will forecast data for 5 minutes ahead. If it is not set, `forecast=3`it will be set.

The output is shown below. shows the number of data points for `_count`each 5-minute interval . is the value predicted by the autoregression algorithm (described later). is the value predicted by the linear regression model. is the value obtained by subtracting the

value of by the value of . These four values are displayed when graphing. The missing values from here on are because they are in the future. Therefore, only is displayed.`_timeslice_count_predicted_count_linear_count_error_count_linear_count2023-07-26 12:15:00.000 PM +0900_count_predicted_count_linear`

Autoregression means that past values affect the present values and cause future values to occur. This can be graphed as shown below.

The dark blue area on the right, from 12:00 PM onwards, `_count_predicted`is a visualization of the forecast values. This graph allows you to see how values will fluctuate over time in the future.

- **Reference source**
  - predict operator - predict Search Operator | Sumo Logic
  - timeslice operator - timeslice Search Operator | Sumo Logic Docs
  - Count Operator | Sumo Logci Docs

summary

What did you think? This is just an example, so I think it can also be used for things other than Apache servers. For example, you could use this example to calculate the URL cache rate and predict future values, predict disk replacement times based on disk usage counts, or predict DB read/write capacity. You might be able to visualize this data as something to pay particular attention to during busy periods.

In terms of security, it may be possible to graph trends in logins and access to important servers, displaying actual and predicted data, and aiming to strengthen security measures early on based on increasing trends.

I hope this article will be of some help to someone.