Estimating Efficient Sampling Rates of Metrics for Training Accurate Machine Learning Models

Tigran A. Bunarjyan[1], Ashot N. Harutyunyan[1], Arnak V. Poghosyan[1], A.J. Han Vinck[2], Yanling Chen[2], and Narek A. Hovhannisyan[3]

[1]VMware Eastern Europe, [2]University of Duisburg-Essen, [3]TeamViewer Armenia

CODASSCA 2020

AUA, Yerevan, Armenia
September 14-17, 2020

vmware

Problem Motivation

Cloud management solutions provide full real-time visibility into modern SDDCs.

Tuning the monitoring solutions according to "adequate" or "efficient" sampling rates will result into:

- a reduction of data management overheads, noise etc.
- save extra compute and storage resources for various on-demand tasks

Our investigation deals with experimental evaluation of efficient sampling rates of time series data in Wavefront subject to two important criteria:

- preserving the original distribution of the metric with minimum information loss
- preserving the information value of the metric in terms of accuracy of ML models we train on to deliver important analytics features in the product

vmware

Cloud management products, such as see Wavefront and vR Ops, are aimed at designing monitoring solutions of high precision and increasingly wider coverage of data center administration aspects. Often these solutions are enabled with a high frequency of sampling rate of data center indicators that is targeted to acquire a maximum level of information to take actions toward many product.

On the other hand, samples acquisition at maximum possible rate implies various costs that affect efficient resource management and design of data-driven analytics.

As a matter of fact, based on such an analysis, we can categorize our initial data base of time series into classes, where each class is characterized by its own efficient sampling rate (reflecting the nature/dynamism of individual flows).

Application Monitoring Solution by Wavefront

An effective tool for application performance analysis and optimization etc.

- Wavefront is designed to get deep insights from the underlying system with high-frequency sampling rate of time series.
- Wavefront also provides an alternate view to customer chart data that mainly focuses on anomaly detection and forecasting of monitored data.

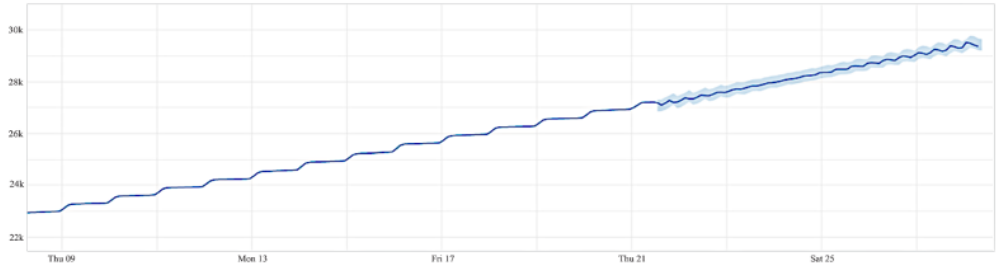

Fig 2. alerting.alerting_period.duration.max data values under 1 p/s (blue) vs. 1 p/10s (yellow) rates

vmware

Wavefront is also an effective tool for application performance analysis and optimization, efficient capacity management and proactive planning, etc., with an intelligent time series query language.

Sampling Rate Reduction

Incremental reduction of sampling rates ranging from 1 sample per 1 to 10 second.

Dataset consists of 1530 metric time series collected by Wavefront.

- Filter out the dataset metrics with some level of consecutive constant behavior.
- Look into only the data values that were collected under lower frequencies.

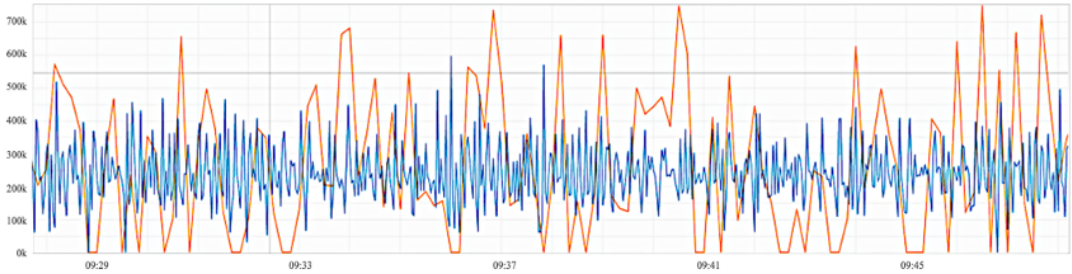

Fig 2. alerting.alerting_period.duration.max data values under 1 p/s (blue) vs. 1 p/10s (yellow) rates

vmware

To analyze the data distribution of the corresponding sampling rate, we specifically look into only the data values that were collected under lower frequencies. Fig. 2 illustrates ten times sample reduction principle to alerting.query.reissue_latency.duration.s metric original data.

Sampling Rate Reduction

Each metric of time series dataset consists of values spread in different range of intervals.

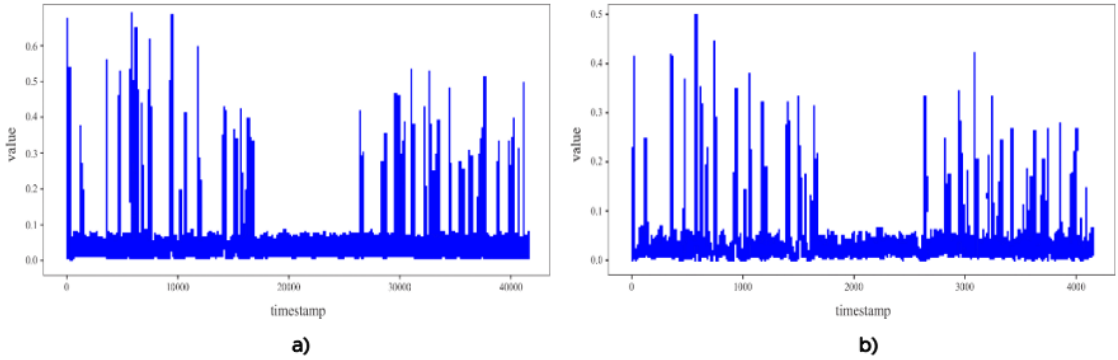

Fig 3. alerting.alerting_period.duration.max metric sampled at (a) 1p/s and (b) 1p/10s.

vmware

The incremental reduction of sampling rates leaves with ten distinct datasets corresponding to data with sampling frequency ranging from 1 sample per second to 1 sample per 10 second. Fig 3a and 3b show an example of the visual difference between ten- and zero-times reduced alerting.alerting_period.duration.max metric data that Wavefront has gathered from monitoring sample cloud application.

Histogram Distributions

Estimation of the probability distribution of the metric.

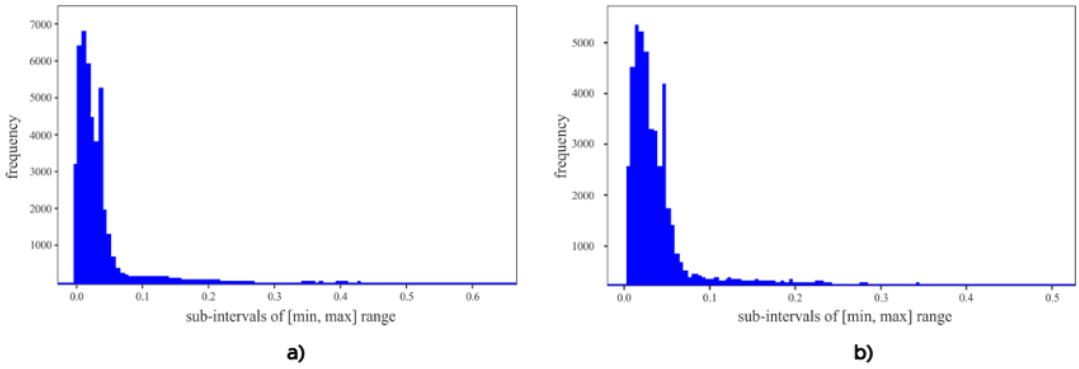

Fig 4. Histogram distributions of (a) original and (b) reduced data of the alerting.alerting_period.duration.max metric.

vmware

Each metric of time series dataset with respect to each sampling rate from 1 to 10 per second, consists of different values spread in different range of intervals. Our approach is to divide the range of each metric into high-granularity (thousand) sub-intervals and compute the relative frequency of data points that fall into those for constructing relevant histogram distributions.

The goal is to get an estimate of the probability distribution/mass function of the metric and see how the sampling rate reduction distorts it. Then the obtained relative frequencies for each metric and its reduced version can be interpreted as probability distributions of those. Fig 4 depicts histogram distributions of the original and 10-times reduced data.

Information Loss Criteria

In what significance does the sampling rate reduction affect the loss of information?

We look into the Jensen-Shannon divergence/distance (also referred to as JSD), to measure the closeness between two probability distributions.

For any probability distributions P and Q, the JSD is defined by the formula:

$$JSD(P,Q) = \frac{1}{2}D(P,M) + \frac{1}{2}D(Q,M),$$

where $M = \frac{P+Q}{2}$ and $D(P,Q)$ is the KL divergence between probability distributions P and Q.

vmware

As our incentive is to measure in what significance does the sampling rate reduction affect the loss of information, we look into JSD, that is a method based on Kullback-Leibler divergence measuring the closeness between two probability distributions. So, To see how similar the probability distributions of original and reduced sampling rated data are, we compute the relevant JSD.

Information Loss Criteria

The significant information content is preserved.

We verified that 1290 out of 1312 metrics experience no more than 4% information loss.

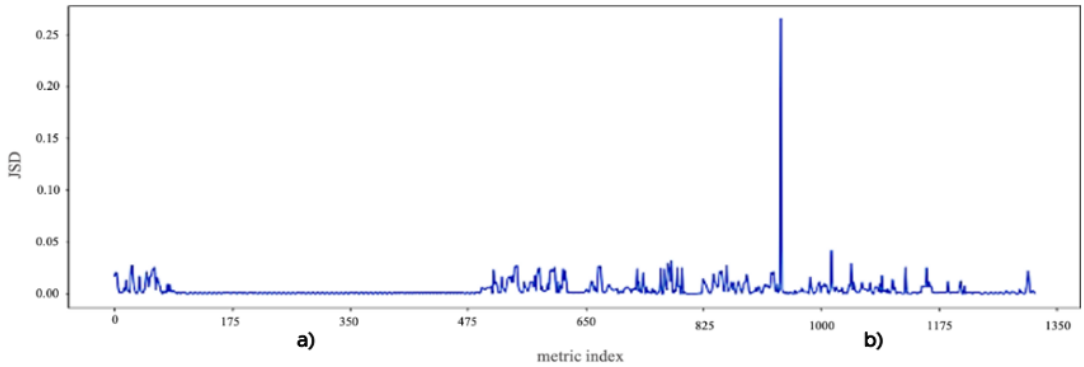

Fig. 5. Information loss experienced by all metrics of sampled at 1p per 10s rate.

vmware

As a result of information loss analysis, we found out that despite the dramatic change in sampling rate, the significant information content is preserved in most of the metrics that are monitored. We verified that 1290 out of 1312 metrics experience no more than 4% information loss (see Fig. 5) while reducing the sampling rate ten times. In this way, we were able to categorize application metrics based on their sensitivity to the sampling rate and refer to ten times lower sampling rate as general sampling rate for 1290 metrics to guarantee the information loss tolerance.

Although most metrics turned to be indifferent to sampling rate reduction in terms of information loss, there is a group of metrics which need preserving the predefined sampling rate. Those metrics express high variability and dynamic changes of the system; hence, high-frequency sampling will be required to detect their abnormality behaviors and other important patterns for reliable management purposes.

Performance Analysis of Algorithms and Related Art

Data reduction in production analytics is an important technology challenge.

- High reduction of sampling rate with low information loss is still tolerable in terms of a quality statistical analysis of the data .
- High-frequency sampling will be required to detect their abnormality behaviors and other important patterns for reliable management purposes.

An interesting empirical evaluation of forecasting algorithms demonstrates that cross-validation approaches can be applied to stationary time series, so:

- the most accurate estimates are produced by out-of-sample methods that preserve the temporal order of observations
- we are interested in validating ML models subject to information loss induced by sampling rate reduction

vmware

The sampling rate reduction may affect the monitored object data from the perspective of change in seasonality and trend. We conducted Seasonal Trend Decomposition (also referred to as STL) filtering procedure for decomposing a seasonal time series. The comparison of STL results applied on the original and ten-times reduced data set demonstrated that none of the metrics in our dataset experience variations in seasonal and other effects being sampled at 1 point per 10 seconds frequency. Thus, we see that such a high reduction of sampling rate with low information loss is still tolerable in terms of a quality statistical analysis of the data.

Performance analysis of algorithms for time series data has been mainly conducted in the context of cross-validation strategies. Compared to such a classical cross-validation setting, we are interested in validating ML models subject to information loss induced by sampling rate reduction.

Time Series Forecasting and Anomaly Detection

Discovery anomalous patterns in the data indicating possible misbehaviors.

The ARMA-based approach of anomaly detection incorporates multiple competitive models using online forecast engine to discover set of anomalies in testing window.

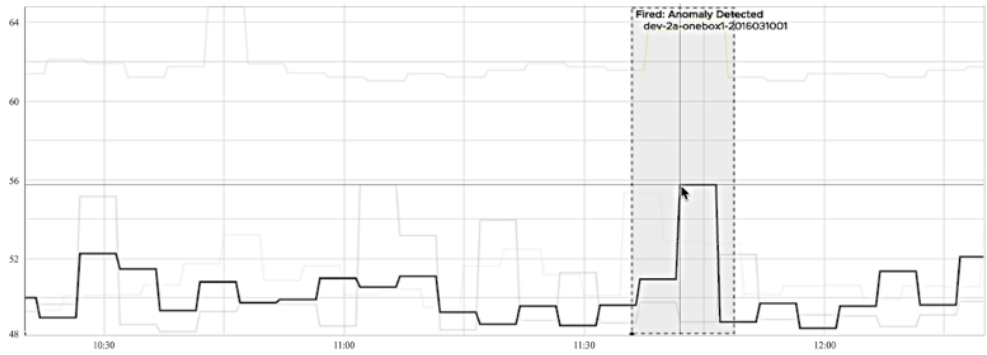

Fig. 6. ARMA-based anomaly detection in Wavefront's AI Genie.

vmware

Wavefront also provides advanced analytical functionality in terms of time series forecasting and anomaly detection - mechanisms that enable its customers to promptly discover anomalous patterns in the data indicating possible misbehaviors within the workflows of the monitored environment. In particular, AI Genie applies two different anomaly detection and forecasting algorithms (ARMA-based and W-TSF.

This algorithm describes the short-term temporal dependent patterns in the time series using autoregressive moving-average (ARMA) model. It calculates the forecast confidence bounds based on the residual error of the selected ARMA model and the user-provided confidence interval parameter. The algorithm performs anomaly detection on complete set of forecasted window values directly.

Time Series Forecasting and Anomaly Detection

Discovery anomalous patterns in the data indicating possible misbehaviors.

W-TSL technology leverages offline pre-trained neural network models and hypothesis testing procedures:

- to achieve confidence bound-assist anomaly detection using transformations of data from non-stationary into a stationary process

Fig. 7. W-TSF anomaly detection in Wavefront's AI Genie.

vmware

W-TSL technology leverages offline pre-trained neural network models and hypothesis testing procedures to achieve confidence bound-assist anomaly detection using transformations of data from non-stationary into a stationary process. In contrast to ARMA-based algorithm, W-TSL starts iterating over the time series with sliding test window principle and calculates anomaly scores for each of the slide window using the set of forecasted values only for candidate slide window itself.

Time Series Forecasting and Anomaly Detection

Calculation of anomaly scores using the set of forecasted values only.

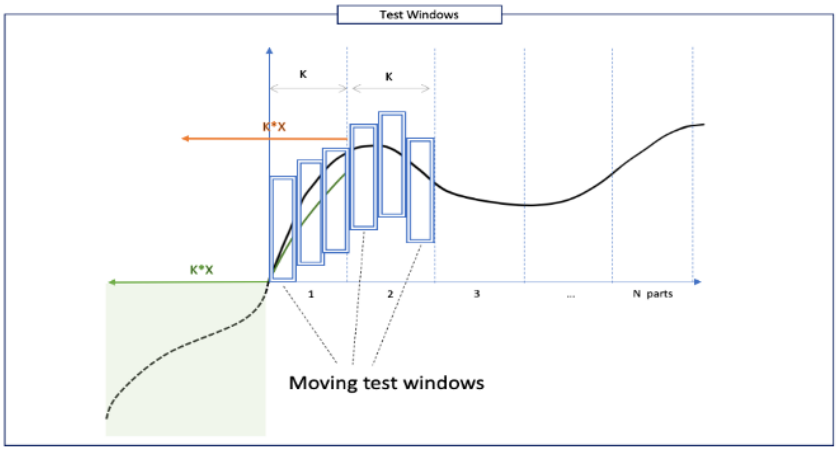

Fig. 7. W-TSF anomaly detection in Wavefront's AI Genie.

If the anomaly score for a time window (e.g., 5 min) crosses a threshold (80% as a default value), then the window is declared to be anomalous (and alert is triggered). The anomaly score is computed by the percentage of data points outlying the confidence bounds of the anomaly detection algorithm, so it varies from [0,1].

Even though the above mentioned two algorithms introduce non-similar architecture designs and approaches to forecasting and anomaly detection task, they still rely on substantial amount of historical data to provide fundamental and accurate analytics on top of that.

Validating Efficient Sampling Rates

Results of anomaly detection performed by two algorithm using AI Genie on production.

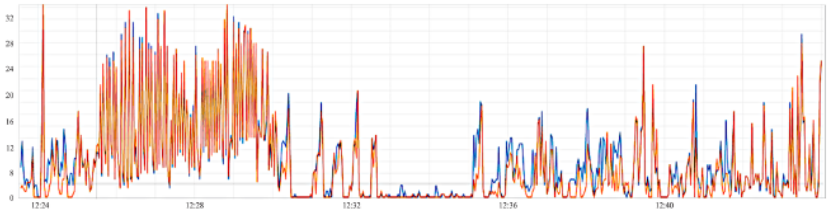

Fig. 9. ARMA-based anomaly count chart (original vs. sampled (ten times reduced))

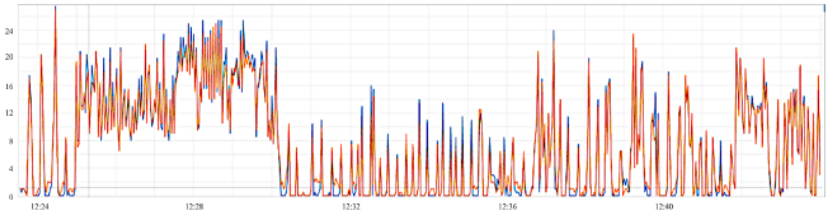

Fig. 9. W-TSF-based anomaly count chart (original vs. sampled (ten times reduced))

To estimate if the reduced time series data set is still preserving its utility for training ML models (such as ARMA-based and W-TSF), we run the following experiment. We use 1290 metrics, each with its original (44392 data points) and reduced sampling rate with the tolerable information loss. Then we compare the results of anomaly detection performed by two algorithm using AI Genie running on production. The algorithms function by taking 20% of historical data at the start for learning and continuously moving forward to produce forecast and anomaly detection for the rest of the data.

Figs 9 and 10 illustrate combination charts of the number of anomalies in the original and sampled datasets provided by the algorithms, respectively, across all metrics, after declaring the total count of anomalies at the end of experiment for each of the metrics.

Validating Efficient Sampling Rates

Results of anomaly detection performed by two algorithm using AI Genie on production.

ML Algorithms	Original data	Ten times reduced data
ARIMA-based	7561	6669
W-TSF	8530	8571

Table 1. Number of anomalies detected by algorithms on test windows of original and ten times sampled datasets.

vmware

As a result of our experiment we discover that W-TSF anomaly detection on ten-times reduced metric dataset provides almost identical distributions of anomalies in the test window. ARMA-based algorithm was more sensitive to the reduction with loss of 11% of anomalies.

Validating Efficient Sampling Rates

The average metric MSE for ARMA-based algorithm is 0.0083 and for W-TSF is 0.0135

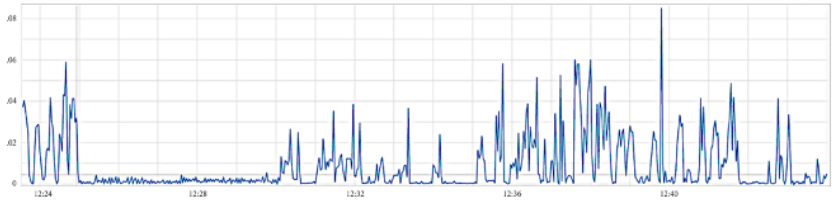

Fig. 11. MSE between anomaly scores in test window of original and sampled (ten times reduced) metrics data using ARMA-based anomaly detection.

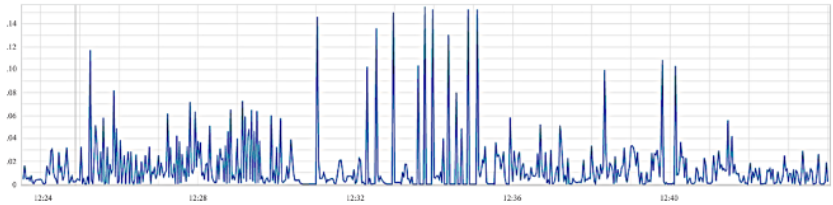

Fig. 12. MSE between anomaly scores in test window of original and sampled (ten times reduced) metrics data using W-TSF anomaly detection.

vmware

Figs 11 and 12 show that the mean squared errors (MSE) of anomaly scores between original and reduced data (that both algorithms adopt as a measure to define anomalies/alerts on). The average metric MSE for ARMA-based algorithm is 0.0083 and the same average for W-TSF is 0.0135. So, for both algorithms we get a low difference in anomaly scores, which implies that anomaly detection algorithms still provide adequate predictions with much sparser data sets. Additionally, with reduced data sets we get significantly (10 times) less memory utilization and Disk IO.

This verifies our intuition that the data reduction subject to tolerable information loss could also provide anomaly predictions enough accurate compared to baselines. Such a reduction implies an essential gain in overall performance for every streaming operation with the data. Based on the proposed analysis and related algorithms, we recommend the product the efficient sampling rates learned for effective data management and analytics. Our algorithms can be regularly run to re-estimate those efficient rates with application evolution/dynamics.

Summary and Conclusion

We described an information-theoretic approach to estimating efficient sampling rates of monitoring flows while preserving their information content.

As a result of experimental research we found out that:

- Significant reduction levels can be achieved
- It is possible to substantially reduce the data management and analytics overhead
- Complex ML models can be still trained with acceptable accuracies
- AI features can improve

vmware

- Our experiments on a large data set measured by Wavefront demonstrate that significant reduction levels can be achieved with very low information loss.
- With such an approach we can substantially reduce the data management and analytics overhead forced by high-frequency monitoring and real-time analysis/representations of data center processes in daily operations.
- Those experiments prove that complex ML models can be still trained within the product with acceptable accuracies on substantially reduced data sets.
- It also improves performance of AI features of a cloud management product in terms of forecasting and anomaly detection.