Cloud Operations

How to Monitor and Alert for AWS

How to Monitor & Alert for AWS

Overview

Maintaining AWS uptime is a required for contiguous business operations and compliance. This paper provides a foundational understanding of Amazon Web Services (AWS) and the most operational and cost-effective means to maintain operational insight for uptime and cost management.

The cost of not monitoring AWS assets is significant. 98% of CIOs estimated the cost to be more than $100,000 per outage or more depending on the type of business. Customers call into support complaining the service is down before you are aware, increasing the churn of the customer base. Employees lose productivity as they are unable to perform their work.

The ability to operate and monitor AWS is a requirement for companies that leverage AWS to perform business. Much of the failure of AWS projects is caused by technical and managerial personnel not understanding AWS costing nor how to set up an efficient infrastructure that alerts.

The result is that businesses are paying for alerting, not monitoring. The driving factor of what to alert on is based on their timeliness to respond. Ultimately, AWS monitoring comes down to how much a business is willing to pay for alerting.

About **Pure Cloud OPs**

Pure Cloud Ops has been operating in the AWS infrastructure for more than twelve (12) years. To scale, maintain availability, and be cost effective, Pure Cloud Ops continually researches, and measures approaches to a healthy and efficient AWS infrastructure. We have developed a hybrid-alerting approach that aligns operational needs to business costs. Our belief is that *quality alerting occurs from a planned approach*.

Pure Cloud Ops is writing this paper as a member of the AWS community. It has been several years since there has been an open discussion on AWS monitoring. Lately, articles are selling opinions that are incorrect. For instance, while AWS Metric Stream is technically a good idea, it is not a good idea for most organizations to implement it. It is too costly and has inconsistent results as a sole monitoring solution and requires multiple interfaces to see collective results.

Content

Having a structural understanding of monitoring and what to monitor directly relates to the ability to get value from monitoring and subsequent alerting.

- **Collection**: Determining what to collect and how to collect it.
- **Alert**: How to articulate an alert that is relevant.
- **Response**: How to close alerts.

Monitoring is not a tool, but a continuous process. Having a plan allows that ability to have results and recognize when monitoring is achieved.

The centerpiece of a monitoring plan is determining what are the situations (alerts) that support an action (response). For this reason, a plan starts with understanding how alerts trigger, and what alerts need to be created. We do not start by looking at what metrics are available. For AWS there are over a thousand metrics that can be collected via API. While AWS seems like it costs little to perform API calls, the volume of metrics and the frequency of calls can generate a substantial invoice. Furthermore, many metrics are not in play for an organization. Later, we will look at the cost of API calls.

Alerting

Alerts and notifications reduce response times and increase the scalability of a system. By setting up rules and thresholds, the system can identify anomalies and trigger alerts to notify personnel to take corrective actions or even connect to automated responses. This process prevents or mitigates potential problems, improving system reliability and uptime, and reducing operational costs.

The basics algorithms for alerting are:

- **Seasonal prediction** considers seasonal trends to predict future values
- **Linear regression** prediction is a base alerting algorithm that uses historical data to make predictions about future values.
- **Threshold** is a simple alerting algorithm that triggers an alert when a metric exceeds a specified threshold.
- **Interruption** detects sudden changes or drops in metric values, and instance detects when a metric value occurs for the first time.

These base algorithms can be customized and combined to create more advanced alerting strategies. Pure Cloud Ops's prediction model ensures that alerts are triggered accurately and in a timely manner, allowing users to quickly respond to potential issues.

Pure Cloud Ops uses dashboards to visual represent data and insights. Dashboards help stakeholders to understand the state of a system, monitor performance, identify trends, and make informed decisions based on the available data.

While dashboards are useful for initial data exploration and understanding, they are not suitable for operational use, where real-time decision-making and intervention are required. In such cases, automated systems that monitor and analyze data in real-time and alert relevant parties when a problem arises is superior.

Pure Cloud Ops believes that while dashboards can be useful for data visualization and exploration, alerting is more effective for operations.

Cost

Part of “How To” should include “How to Pay” for monitoring. How a company monitors determines the cost of monitoring. Monitoring costs are divided into two parts: the “monitoring tool cost” and the “use cost". Monitoring tools normally charge per seat cost. For instance, Data Dog charges $23/month per asset being monitored. The use cost of monitoring is the AWS costs to the user of using the AWS API. This is $0.01at 1,000 calls. The number of calls, and therefore the cost, is based on three factors:

- Number of Metrics being monitored
- Frequency
- Retention window

Businesses that run on AWS quickly learn that brute force solutions equate to high costs. AWS charges by usage. This is simple, but fundamentally different than how costing works for server infrastructures. In a server infrastructure a company purchases the server and can leverage the full power of the server. Usage does not cost more. Most software prior to the cloud considers this cost model.

In AWS, a company does not pay until they use the infrastructure. This makes inefficient processes more expensive. Organizations often throw more computational power to overcome large data. In AWS, the system scales processing power at resolve congestion, however, the cost also increases at the same rate as the processes scales. Solving a technical problem using brute force means it will cost more in AWS.

To be cost effective in the cloud, a monitoring solution needs to:

- Focus on alerting more than dashboard
- Not increase personnel workload as the infrastructure scales
- Reduce the usage per asset/function
- Adjust usage based on business need/criticality

The cost of a monitoring solution, the use cost, is carried by the operator and not the monitoring software. An analogy is buying a car. The cost per mile to drive the car is based off the fuel efficiency of the car you buy. Monitoring solutions can be very expensive based on the solution you buy.

Cost Example

Traditional polling to monitor an infrastructure is inexpensive when operating on a system model. The polling of the infrastructure costs little in processing power or network bandwidth. Systems and processes could be polled aggressively with no change in cost to operate. In the cloud however, the cost is charged per API request.

This means a one (1) minute interval will have 43,800 requests a month. For SOC2 compliance there are seven (7) common metrics to request, resulting in 306,600 requests per system. The infrastructure of **500 assets** is 153,300,000 calls a month. At $0.01at 1k calls, means that this cost the company $**1,533 USD per month** just for API calls.

Lessons Learned in Cost

A long-term strategy of turning on metrics across the entire infrastructure indiscriminately is wasteful. It is essential to have a thoughtful and strategic approach to metrics collection and monitoring.

Start by using low-frequency, common metrics to gain visibility into the system's performance and identify any potential issues or bottlenecks. SOC2 basic-seven metrics, as provided by Pure Cloud Ops, can be a good starting point, especially for compliance requirements. These metrics cover essential areas such as system errors, capacity issues, and high CPU usage. It also provides a complete view without agents.

However, it is also important to recognize that critical systems may require more extensive and detailed metrics that operate at lower frequencies. This is where AWS tagging can be useful in marking assets and matching metrics to their corresponding frequency. By tagging assets, it becomes easier to track

performance across different systems and identify any potential issues or trends that may emerge over time.

Monthly Cost of Monitoring

Let's use the same costing exercise as before with this approach. This means a five (5) minute interval will have 8,760 requests a month. For SOC2 compliance there are seven (7) common metrics to request, resulting in 61,320 requests per system. The infrastructure of **500 assets** is 30,660,000 calls a month. At $0.01at 1k calls, means that this cost the company $**307 USD per month** just for API calls.

The cost of the software, using a Pure Cloud Ops Professional plan, plus AWS use costs is **$705**. This is a tenth of the most common solution. Using the Data Dog Pro plan (their lowest plan), plus AWS use costs is **$7,807**.

In summary, a good metrics strategy should start with common and low-frequency metrics that provide a baseline for visibility and compliance needs. As the system evolves, more extensive and detailed metrics can be added, and AWS tagging can be used to manage and match metrics to their corresponding assets and frequency.

Conclusion

To be successful in monitoring your AWS infrastructure, you need a solution that can provide immediate insight and scale to your growth. Three aspects to consider are cost, visibility, and alerting capabilities. The number of metrics and the interval have the largest impact on cost. While metrics will be the future of monitoring, how it is packaged, and its reliability issues make it a bad choice for operations today.