Which Metric Types Should We Use?

In this lesson, we will discuss which metric types should we use?

WE'LL COVER THE FOLLOWING

If this is the first time you're using `Prometheus` hooked into metrics from Kube API, the sheer amount might be overwhelming. On top of that, consider that the configuration excluded many of the metrics offered by Kube API and that we could extend the scope even further with additional exporters.

While every situation is different and you are likely to need some metrics specific to your organization and architecture, there are some guidelines that we should follow. In this section, we'll discuss the key metrics. Once you understand them through a few examples, you should be able to extend their use to your specific use-cases.

Key metrics

> 🔍 The four key metrics everyone should utilize are latency, traffic, errors, and saturation.

Those four metrics are being championed by Google Site Reliability Engineers (SREs) as the most fundamental metrics for tracking performance and the health of a system.

Latency

Latency represents the time it takes a service to respond to a request. The focus should not be only on duration but also on distinguishing between the latency of successful requests and the latency of failed requests.

Traffic

Traffic is a measure of demand that is being placed on services. An example would be the number of HTTP requests per second.

Errors

Errors are measured by the rate of requests that fail. Most of the time those failures are explicit (e.g., HTTP 500 errors) but they can be implicit as well (e.g., an HTTP 200 response with the body describing that the query did not return any results).

Saturation

Saturation can be described by the "fullness" of a service or a system. A typical example would be the lack of CPU that results in throttling and, consequently, degrades the performance of the applications.

Over time, different monitoring methods were developed. We have, for example, the **USE** method that states that for every resource, we should check **u**tilization, **s**aturation, and **e**rrors. Another one is the **RED** method that defines **r**ate, **e**rrors, and **d**uration as the key metrics. Those and many others are similar in their essence and do not differ significantly from SREs demand to measure latency, traffic, errors, and saturation.

We'll go through each of the four types of measurements described by SREs and provide a few examples. We might even extend them with metrics that do not necessarily fit into any of the four categories.

Q The number of HTTP requests per second, comes in which Key Metric?

In the next lesson, we will see Latency Related Issues.