TencentCloud Managed Service for Prometheus

Product Introduction

Product Documentation

Contents

Product Introduction

Overview

Last updated：2024-01-29 15:55:07

TencentCloud Managed Service for Prometheus (TMP) provides the highly available Prometheus service as well as the open-source visualization tool Grafana and Cloud Monitor alarms while inheriting the monitoring capabilities of the open-source Prometheus, which reduce your development and OPS costs.

Prometheus Overview

Prometheus is an open-source monitoring system. Similar to Kubernetes inspired by Google's Borgman monitoring system, it was inspired by Google's Borgman monitoring system. It was created and developed in 2012 by SoundCloud's internal engineers and released in January 2015. In May 2016, it became the second project after Kubernetes to officially join Cloud Native Computing Foundation (CNCF). Nowadays, it is usually used for monitoring in the most common Kubernetes container management systems.

Prometheus has the following features:

Custom multidimensional data models (a time series data entry is composed of a metric and a key-value pair called label).

Flexible and powerful query language PromQL, which can use multidimensional data to complete complex monitoring queries.

Independence from distributed storage and support for operations based on one single master node.

Time series data collection through HTTP pull.

Data push through Pushgateway.

Acquisition of collection target servers through dynamic scrape configuration or static configuration.

Integration with Grafana to easily support various visual charts and dashboards.

Features

According to the layering of monitoring, TMP covers business monitoring, application layer monitoring, middleware monitoring, and system layer monitoring. Together with Cloud Monitor's alarming capabilities and open-source Grafana, it can provide a one-stop all-round monitoring system to help your quickly identify and locate business problems and reduce the impact of various faults on your business.

System layer monitoring: CPU, memory, disk, network, etc.

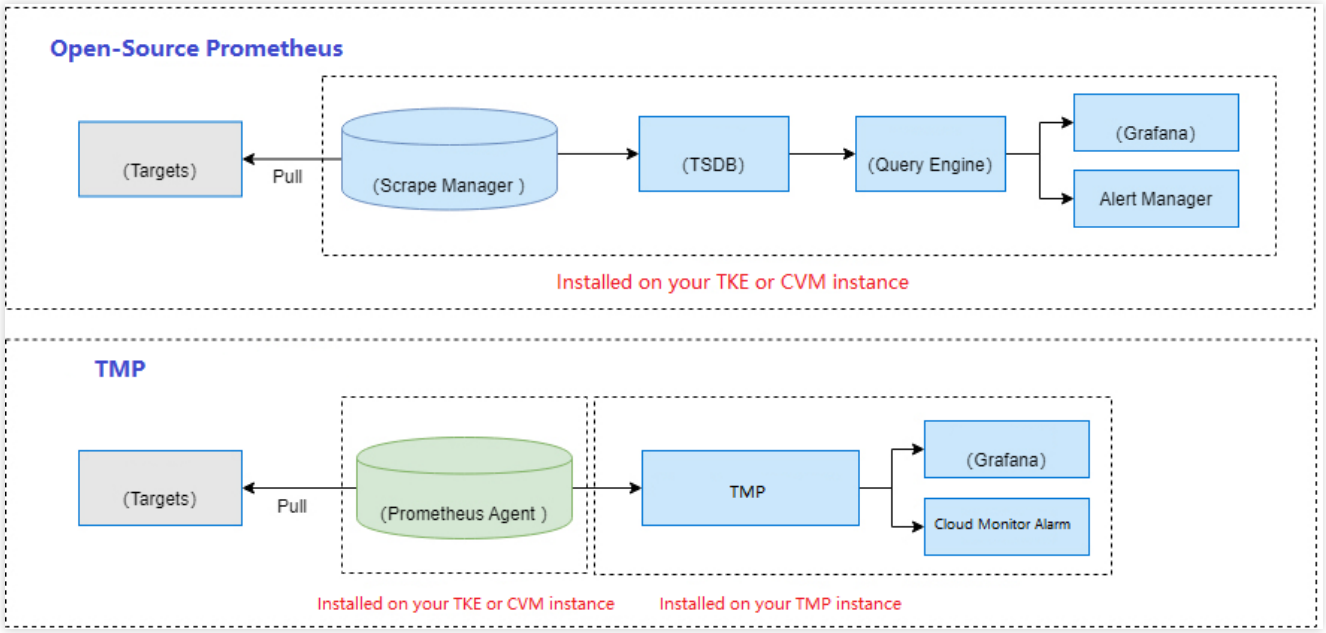
Intermediate component layer monitoring: Kafka, MySQL, Redis, etc.

Application layer monitoring: application services such as JVM, HTTP, and RPC.

Business monitoring: golden business metrics such as the number of logins and the number of orders.

Strengths

TMP has the following strengths over the open-source Prometheus:



Lighter, more stable, and more available.

Fully compatible with the open-source Prometheus ecosystem.

Free of manual setup, saving the development costs.

Highly integrated with TKE, saving the development costs of integrating with Kubernetes.

Integrated with Cloud Monitor's alarming system, saving the costs of developing alarm notifications.

Integrated with commonly used Grafana dashboards and alerting rule templates.

Strengths

Last updated：2024-01-29 15:55:07

Lightweight Service

Compared with the open-source Prometheus monitoring service, TMP features a more lightweight overall structure. You can use a TMP instance simply after creating it in Cloud Monitor, where an agent can complete data scrape by using less than 1 GB of memory.

High Stability and Reliability

TMP only occupies megabytes of resources, much fewer than the open-source Prometheus does. It also integrates Tencent Cloud storage services and its own replica capabilities to reduce the number of system interruptions and provide services with a higher availability.

Great Openness

TMP offers the out-of-the-box Grafana service. It also integrates a wealth of Kubernetes basic monitoring services and dashboards for common service monitoring, which can be quickly used after activation.

Low Costs

TMP provides the native Prometheus monitoring service. After you purchase a TMP instance, you can quickly integrate it with TKE to monitor services running on Kubernetes, which eliminates the costs of setup, OPS, and development.

High Scalability

TMP features an unlimited data storage capacity not restricted to local disks. It can be dynamically scaled based on Tencent Cloud's proprietary sharding and scheduling technologies to meet your elastic needs. It also supports CLB for better load balancing. This helps solve the pain points that the open-source Prometheus cannot be scaled horizontally.

High Compatibility

TMP is fully compatible with Prometheus protocols, support core APIs, custom multidimensional data models, flexible query language PromQL, and collection target discovery through dynamic scrape configuration or static configuration, so you can easily migrate to it for smooth access.

Use Cases

Last updated：2024-01-29 15:55:07

Integrated Monitoring

TMP provides the one-stop out-of-the-box Prometheus monitoring service, which is natively integrated with Grafana dashboards and Cloud Monitor alarms for various monitoring scenarios such as basic services, application layer, and container services.

Application Service Monitoring

Scenario 1

An application provides external API services but their quality information cannot be secured. TMP can be integrated according to the development language to monitor the access request volume, delay, and success rate of APIs in real time.

Scenario 2

TMP also detects service exceptions to help you understand which APIs an exception responds to, which servers an exception occurs on, and whether an exception occurs on individual servers or in the entire cluster.

Scenario 3

For Java applications, TMP can monitor the GC, memory, and thread status of individual servers to help you fully understand the internal status of JVM.

CVM Monitoring

If your service is deployed in CVM, you must modify the Prometheus scrape configuration almost every time the service is scaled. For this kind of scenario, with the aid of the tagging capability of the Tencent Cloud platform and the tag discovery capability of the Prometheus agent, you only need to properly associate tags with CVM instances to easily manage and monitor target objects, which helps you get rid of continuously updating the configuration manually; for example:
1. Service A is deployed on 2 CVM instances at the same time, and the instances are associated with the tag "service name: A";
2. The original number of CVM instances cannot meet the requirements of a business campaign, and 3 more instances should be added. At this time, you only need to associate the tag "service name: A" with these new instances, and then the agent will automatically discover them and actively scrape their monitoring metrics;
3. After the campaign is over, 3 CVM instances are removed, and the agent can automatically discover the instance deactivation and stop scraping their monitoring metrics.

Custom Monitoring

You can use TMP to customize the reported metric monitoring data so as to monitor internal status of applications or services, such as the number of processed requests and the number of orders. You can also monitor the processing duration of some core logic, such as requesting external services.

Concepts

Last updated：2024-01-29 15:55:07

This document describes the basic concepts involved in using the TMP service, so that you can check and understand related concepts.

Concept	Description
Exporter	Exporter is a component that collects monitoring data and provides data externally based on Prometheus monitoring specifications. There are currently hundreds of official or third-party tools available, see Exporters and Integrations.
Job	A collection of configurations for a set of targets. After the scrape interval is defined, access restrictions will will be applied to the scraping job for a given set of targets.
TMP instance	The logical unit for managing monitoring data collection and data storage analysis provided by TMP.
TMP probe	Kubernetes cluster deployed on either the user or Tencent Cloud side. It is responsible for automatic discovery of collection targets, collection metrics and remote writing to other libraries.
PromQL	The query language for the TMP service, which supports instant query and timespan query. It has a variety of built-in functions and operators, so raw data can be aggregated, sliced, predicted, and federated.
Target	The collection target to be scraped by the Prometheus Agent. The collection target either exposes its own operation and business metrics or serves as a proxy for exposing the operation and business metrics of a monitored object.
Alerting rule	The alert configuration of alerting rule formats in TMP, Which can be described by PromQL.
Label	An pair of key-value used to describe the metric.
Scrape configuration	One of the features in TMP, which can automatically discover collection targets without static configuration. It supports Kubernetes SD, Consul, Eureka, and other ways of service discovery. It also allows collection targets to be exposed via Service Monitor and Pod Monitor.
Recording rule	The recording rule capacity in TMP. With recording rule, raw data can be processed into new metrics through PromQL to improve query efficiency.
Integration Center	It integrates all the services supported by TMP. You can install the corresponding services as instructed on the page. After successful installation, you can view the monitoring data on the monitoring panel.
Alerting rule	It is used to define how to trigger and send an alert.

Cloud product monitor	TMP integrates the monitoring data of Tencent Cloud products. You can quickly install the Agent to view the monitoring data.
Metrics	Metrics are used to collect a series of labeled data exposed by the target, which can fully reflect the operation or business status of the monitored object.
TPS	The total number of reported data points per second. It is an important metric to measure the processing power of the system.
Series limit	The maximum number of metrics. The upper limit of series = (single metric × the dimension combination of the metric) × the number of metrics.

Use Limits

Last updated：2024-07-30 18:14:31

Instance Limits

Each instance can have up to 4.5 million series. Free trial instance is limited to 2 million series. If you need to adjust the limit for a paid instance, contact us. In the case of adequate resources, we will adjust the related limit for you properly.

Note:

A series consists of a metric name and label. The same metric name and labels form a unique series.

Custom Reporting Limits

If you use TMP's custom monitoring feature to monitor data, there will be the following limits on metrics (series with a unique `__name__`).

Data reporting must carry a metric name, i.e., the `__name__` label, which can contain only ASCII letters, characters, digits, underscores, and colons and must start with a letter and match the regex `[a-zA-Z_:][a-zA-Z0-9_:]*` . For more information, see Metric names and labels.

Each metric can have up to 32 labels.

The label name can contain only ASCII letters, digits, and underscores. It must match the regex `[a-Za-Z_][a-Za-Z0-9_]*` . Labels beginning with `__` are reserved for internal use.

The label name and label value can contain up to 1,024 and 2,048 characters respectively.

Under the same metric, the dimension combinations of labels cannot exceed 100,000. When the histogram has many buckets, the histogram-type metrics cannot be adjusted.

Total number of data points reported per second: A paid instance cannot exceed 300,000, and a free trial instance cannot exceed 100,000.

Note

The role of labels: In Prometheus, data is stored as time series, which are uniquely identified by the metric name and a series of labels (key-value pairs). Different labels represent different time series, so you can query the specified data by label. The more labels you add, the finer the query dimension.

Prometheus Query Limits

To ensure the query efficiency and better user experience, Prometheus query has the following limits (which don't apply to metadata such as queries about labels and don't affect the Grafana metrics browser feature).
The number of time series involved in a single query cannot exceed 100,000.
The amount of data involved in a single query cannot exceed 100 MB.
There is no limit on the query frequency, but if the concurrency exceeds 15, there may be a certain queuing delay in slower large queries (the probability is low though). Large queries with a time span of more than two weeks will have a higher delay.
The above limits also apply to alarm rules and recording rules. We recommend that you limit the query scope based on your business scenario or appropriately split queries in other ways. You can also use the method of splitting first and then aggregating, such as aggregating recorded results again.

Other Configurations Limits

Configuration limits:
The maximum of alarm rules can be configured for each instance: 150.
The maximum of recording rules can be configured for each instance: 150.

Features

Last updated：2024-01-29 15:55:08

Monitoring Object Access

Feature	Description
Instance creation	You can create a TMP instance in multiple region.
Integration center	The integration center supports quick installation and custom access of various components. After successful installation, you can view monitoring data in Grafana.
Health check	Health check detects the service connectivity on a regular basis to monitor the service health, helping you stay up to date with the service health in real time and promptly discover exceptions to improve the SLA.
Custom monitoring	You can customize monitoring data reporting and monitor key business metrics, such as the number of requests, orders placed, and time consumed requesting external services.

Monitoring Metrics for Collection

Feature	Description
Scrape configuration	ServiceMonitor: A built-in scrape configuration feature in TMP, which is automatically enabled when connected. Currently, the default objects for scrape metrics collection are pods contained in all namespaces under the Kubernetes cluster. ServiceMonitor: It collects the monitoring data in the corresponding endpoints of services based on Prometheus Operator in the K8s ecosystem. PodMonitor: It collects the corresponding monitoring data in pods based on Prometheus Operator in the K8s ecosystem.
Targets	You can visually view the target being scraped through Targets and check whether the scraping status is normal. You can also view the metrics exposed in the target.

Monitoring Data

Feature	Description

Getting Remote Write address	The Remote Write feature can store TMP data as a remote database. You can use the Remote Write address to store the monitoring data of self-built Prometheus in the TMP instance to achieve remote storage, and visualize it in the same Grafana system.
Recording rule	A recording rule allows you to calculate some commonly used or complex metrics in advance and then store the calculated data in new data metrics. In this way, querying the calculated data will be faster than querying the original data. This can solve the problems of complicated user configuration and slow query.

Monitoring Data Display

Feature	Description
Grafana	There are numerous embedded Grafana dashboards available. You can also customize one if necessary. Plugins that are often used on the Grafana official website are also preset, and you can install them quickly in the console.
Instance monitoring	It supports the monitoring of TMP instance status and usage, including TMP instance storage, alert sending, Grafana requests and the number of dashboards, helping you check the usage of TMP instances in real time.
HTTP API	This API is used to get the TMP data address. You can use this address to connect the monitoring data of the TMP instance to the self-built Grafana dashboard for data display, or perform secondary development of the TMP monitoring data.

Alerts

Feature	Description
Creating alerting rules	There are numerous preset alerting rules available. You can also customize one for a specific monitoring object. TMP integrates the alarm notification template of Tencent Cloud Observability Platform (TCOP), which notifies you to take measures in time when certain metrics are abnormal.
Managing alerting rules	You can perform operations such as enabling, disabling, editing, and deleting alerting rules. You can also quickly import other instance alerts.

Service Regions

Last updated：2024-08-27 16:44:41

Note:

Service regions refer to locations of physical data centers or servers. You can regard them as the regions or countries where a service or resource is available. Once a resource is successfully created in a data center in a specific region, the region of the resource usually cannot be changed.

Regions and AZs supported by TMP are as follows:

Region	AZ
South China (Guangzhou) ap-guangzhou	Guangzhou Zone 3 ap-guangzhou-3
	Guangzhou Zone 4 ap-guangzhou-4
East China (Shanghai) ap-shanghai	Shanghai Zone 2 ap-shanghai-2
	Shanghai Zone 3 ap-shanghai-3
	Shanghai Zone 4 ap-shanghai-4
	Shanghai Zone 5 ap-shanghai-5
Hong Kong (China), Macao (China), and Taiwan (China) (Hong Kong (China)) ap-hongkong	Hong Kong (China) Zone 1 ap-hongkong-1
	Hong Kong (China) Zone 2 ap-hongkong-2
	Hong Kong (China) Zone 3 ap-hongkong-3
North China (Beijing) ap-beijing	Beijing Zone 3 ap-beijing-3
	Beijing Zone 4 ap-beijing-4
	Beijing Zone 5 ap-beijing-5
	Beijing Zone 6 ap-beijing-6
	Beijing Zone 7 ap-beijing-7
Southwest China (Chengdu) ap-chengdu	Chengdu Zone 1 ap-chengdu-1
	Chengdu Zone 2 ap-chengdu-2
Southwest China (Chongqing) ap-chongqing	Chongqing Zone 1 ap-chongqing-1
East China (Nanjing) ap-nanjing	Nanjing Zone 1 ap-nanjing-1

	Nanjing Zone 2 ap-nanjing-2
West US (Silicon Valley) na-siliconvalley	Silicon Valley Zone 1 na-siliconvalley-1
	Silicon Valley Zone 2 na-siliconvalley-2
Europe (Frankfurt) eu-frankfurt	Frankfurt Zone 1 eu-frankfurt-1
East US (Virginia) na-ashburn	Virginia Zone 1 na-ashburn-1
	Virginia Zone 2 na-ashburn-2
South America (Sao Paulo) sa-saopaulo	Sao Paulo Zone 1 sa-saopaulo-1
Southeast Asia (Singapore) ap-singapore	Singapore Zone 1 ap-singapore-1
	Singapore Zone 2 ap-singapore-2
	Singapore Zone 3 ap-singapore-3
	Singapore Zone 4 ap-singapore-4
Asia Pacific (Seoul) ap-seoul	Seoul Zone 1 ap-seoul-1
Asia Pacific (Mumbai) ap-bombay	Mumbai Zone 1 ap-bombay-1
	Mumbai Zone 2 ap-bombay-2
Asia Pacific (Bangkok) ap-bangkok	Bangkok Zone 2 ap-Bangkok-2
Asia Pacific (Tokyo) ap-tokyo	Tokyo Zone 1 ap-tokyo-1
	Tokyo Zone 2 ap-tokyo-2
Southeast Asia Pacific (Jakarta) spa-jakarta	Jakarta Zone 1 spa-Jakarta-1
	Jakarta Zone 2 spa-Jakarta-2