

МЕТОДЫ МОНИТОРИНГА КЛАСТЕРНЫХ СЕРВИСОВ В ОБЛАСТИ ЭЛЕКТРОННОЙ КОМЕРЦИИ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Жук П.Б., Бобов М.Н.

В этой работе рассмотрены основные метрики и методы, применяемые для мониторинга кластерных сервисов в области электронной коммерции.

В настоящее время для обеспечения высокой доступности веб-сервисов, масштабирования, балансирования трафика, данных между несколькими серверами широко используется подход размещения одного сервиса в кластере серверов.

При таком подходе понимание состояния инфраструктуры и систем важно для стабильной работы сервисов. Информация о работоспособности и производительности развертываний не только помогает вовремя реагировать на проблемы, но и дает возможность уверенно вносить все требуемые изменения. Один из способов получить эту информацию – это система мониторинга, позволяющая осуществлять сбор метрик, визуализацию данных и нотификацию в случае неправильной работы кластерных сервисов.

Мониторинг – это процесс сбора, агрегирования и анализа этих данных для улучшения понимания характеристик и поведения компонентов системы. Данные из разных точек среды собираются системой мониторинга, которая отвечает за хранение, агрегацию, визуализацию данных и автоматически реагирует на изменения, когда значения соответствуют заданным условиям.

Метрики, мониторинг и система оповещений составляют основу системы мониторинга и позволяют отразить состояние системы, отследить тенденции в потреблении ресурсов или поведении, а также влияние вносимых изменений.

Одной из функций систем мониторинга является организация и корреляция данных из различных источников. Эффективность показателей мониторинга можно оценить возможностью администратора шаблоны поведения между разными ресурсами и группами серверов.

В основном, выделяют четыре вида метрики для осуществления мониторинга: задержка ответа сервиса, уровень входящего трафика, ошибки, занятость ресурсов.

Задержка – это время, необходимое для завершения действия. Специфика измерения этой метрики зависит от компонента, ее общие аналоги – время обработки, время отклика.

Задержка показывает, как долго будет выполняться конкретная задача или действие. Измерение задержки различных компонентов позволяет построить целостную модель различных характеристик системы. Это может помочь найти узкие места и определить каким ресурсам нужно больше всего времени, и своевременно обратить внимание на замедление работы системы. Следует подчеркнуть, что при расчете задержек важно учитывать как успешные, так и неуспешные запросы, поскольку они могут исказить средние значения сервиса.

Уровень трафика обозначает занятость ресурсов системы. Это нагрузка на сервисы, которая позволяет определить количество входящего и исходящего трафика, обрабатываемого системой в настоящее время.

Ошибки и их количество позволяют иметь более полную картину состояния компонентов и их реакции на запросы. Разделяя различные типы ошибок, возможно более точно определить проблемы, влияющие на приложения. Это также позволяет настроить гибкую систему оповещений: об отдельных типах ошибок система может оповещать немедленно, а другие игнорировать, пока они не превышают определенный порог.

Данные **использования ресурсов** предоставляют информацию о ресурсах, от которых зависит эффективность сервиса. Поскольку работа сервиса, который предоставлен одним компонентом, может требоваться для работы другого сервиса, использование ресурсов является одним из важнейших показателей для определения проблем с пропускной способностью. Проблемы использования ресурсов и задержки в одном слое могут отображать существенный скачок трафика или наличие ошибок в нижнем слое.

На основании вышеперечисленных метрик существует 2 метода мониторинга кластерных сервисов: USE и RED.

USE (utilization, saturation, errors) метод предназначен для выявления проблем в производительности ресурсов и основан на измерении трех основных метрик использования ресурсов:

- 1) Использование (англ. utilization) – время, в течение которого ресурс был занят обработкой полезного трафика.
- 2) Насыщение (англ. saturation) – степень загруженности ресурса, т.е. отношение необработанного трафика к обработанному.

3) Ошибки (англ. errors) – количество ошибок при обработке.

RED (rate, errors, duration) метод сосредоточен на выявлении ошибок, не связанных в большинстве с производительности (ошибки логики программы, неправильной конфигурации) и основан на трех метриках.

1) Темп (англ. rate) – количество успешно обработанных запросов за единицу времени.

2) Ошибки (англ. errors) – количество неудачно обработанных запросов за единицу времени.

3) Длительность (англ. duration) – интервал времени, необходимый для обработки запроса.

Оба метода обеспечивают оценку работы кластерных сервисов, однако только совместное использование данных методов может обеспечить более высокий уровень качества мониторинга сервисов.

Список использованных источников:

1. Newman, S. Building Microservices // O'Reilly Media, Inc. – 2016 – Piter Press Ltd.– P. 197 - 205

2. Beyer, B. Site Reliability Engineering. / C. Jones, J. Petoff, N. R. Murphy // O'Reilly Media, Inc. – 2016 – P. 50 – 67