Lesson 8: Monitoring and Scaling Kubernetes Applications

Lữ Thanh Tùng

Monitor health and performance of Kubernetes cluster and applications

Monitor Kubernetes Cluster

- Khám phá các node trong cluster: tình trạng hoạt động, tài nguyên sử dụng, số ứng dụng đang chạy và mức sử dụng tài nguyên của toàn bộ cluster.
- Một số chỉ số đo được :
- Các tài nguyên đang sử dụng: Băng thông mạng, sử dụng disk, CPU, memory,.... Sử dụng các số liệu này, ta có thể quyết định tăng hay giảm số lượng và kích thước của các nút trong cluster.
- Số lượng node: Cho phép tìm ra những gì đang trả tiền (nếu sử dụng các dịch vụ) và khám phá mục đích sử dụng của cluster
- Các pod đang chạy: Cho biết số lượng nút khả dụng có đủ hay không và liệu chúng có thể xử lý toàn bộ khối lượng công việc trong trường hợp một nút bị lỗi hay không.

Monitor health and performance of Kubernetes cluster and applications

Kubernetes Pod Status Monitoring

- Chia làm 3 loại: Kubernetes metrics, pod container metrics, application metrics.
- Bằng cách sử dụng Kubernetes metric, theo dõi được cách pod và deployment của nó đang được xử lý bởi bộ điều phối. Sử dụng để theo dõi
  - Số lượng phiên bản mà một pod có tại thời điểm này và số lượng dự kiến (nếu số lượng thấp, cụm của bạn có thể hết tài nguyên),
  - Quá trình deployment đang diễn ra như thế nào (có bao nhiêu phiên bản đã được thay đổi từ phiên bản cũ sang phiên bản mới),
  - Kiểm tra tình trạng và một số dữ liệu mạng có sẵn thông qua các dịch vụ mạng.
- Container metric có sẵn thông qua cAdvisor và được hiển thị bởi Heapster, truy vấn mọi nút về các vùng chứa đang chạy. Gồm các số liệu như : mức sử dụng CPU, mạng và bộ nhớ so với mức tối đa cho phép.
- Application metrics: Do chính ứng dụng phát triển và có liên quan đến các quy tắc kinh doanh mà ứng dụng giải quyết.

Metrics, Logging, and Tracing in Kubernetes

- Metric cấp dữ liệu về tình trạng, hiệu suất và việc sử dụng tài nguyên của Kubernetes cluster: node, pod và volume. Giúp ta hiểu trạng thái và hiệu quả hoạt động của hệ thống.
- Logging liên quan đến việc lưu trữ thông báo được tạo bởi các ứng dụng và thành phần Kubernetes gồm lỗi, cảnh báo và các thông tin liên quan khác.
- Tracing cho phép bạn theo dõi luồng yêu cầu trên các dịch vụ và thành phần khác nhau, cung cấp thông tin chi tiết về hiệu suất và hành vi của các ứng dụng phân tán của bạn. Tracing đặc biệt hữu ích trong kiến trúc microservices.

Horizontal and Vertical scaling trong Kubernetes

- **Horizontal scaling:**
  - Pod Horizontal scale: Tăng số lượng replica pod khi tải cao và giảm số lượng khi ít tải
  - Cluster scale: Tăng số lượng node khi cluster quá tải
- **Vertical pod scaling:** Tăng cấu hình của pod (RAM, CPU, GPU) để đáp ứng khi tải cao và giảm bớt cấu hình khi tải thấp. Mở rộng bị giới hạn do node có giới hạn tài nguyên.