微服务和 Kubernetes

极客时间

扫码试看/订阅

《NLP 实战高手课》视频课程

微服务和 Kubernetes

- 微服务
- 微服务的好处
- 微服务的问题
- Kubernetes

微服务

- 一种设计模式
  - Loosely coupled
  - Independently coupled
  - Owned by small teams
- 常常和 Monolithic Architecture 对比

微服务的好处

- 开发周期和维护
- Scalable Deployment
- Fault Tolerance
- 可以选择独立的 Tech Stack

微服务的问题

- 糟糕的设计
- 微服务本身带来的效率损失
- 微服务带来的其他问题

Kubernetes

- Container Orchestration 工具
- 协助实现微服务的设计模式
- 协助实现大规模云端部署
- 功能举例：
  - Service Discovery
  - Auto Scaling
  - Singleton Management

Docker 简介

Docker 简介

- 环境问题
- Docker Container 和传统虚拟机的区别
- Docker 部署的注意事项

环境问题

- 大部分 AI 库都对环境有强烈依赖
- 大部分 AI 库对于环境错误的容忍性都比较差
- 大部分 AI 库都缺少部分运维的功能

Docker Container 与虚拟机

- “A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings”
- 与虚拟机最大的区别：Shared OS Kernel
- 实际构成：由多个 Layer 构成

Docker部署的注意事项

- Docker 本身仍然有很大的 overhead
- 一些老旧的系统不一定支持 Docker
- Docker 对于 immutable 的服务部署支持较好
- Docker 本身有可能不稳定

Kubernetes 基本概念

Kubernetes 基本概念

- Kubernetes Objects
- Pods
- Deployment
- Service

Kubernetes Objects

- Kubernetes objects are persistent entities in the Kubernetes system.
- A Kubernetes object is a record of intent.
- 通常通过 kubectl 和 YAML 文件进行描述

YAML 文件举例

```yaml
apiVersion: apps/v1 # for versions before 1.9.0 use apps/v1beta2
kind: Deployment
metadata:
  name: nginx-deployment
spec:
  selector:
    matchLabels:
      app: nginx
  replicas: 2 # tells deployment to run 2 pods matching the template
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:1.14.2
        ports:
        - containerPort: 80
```

Pods

- A Pod (as in a pod of whales or pea pod) is a group of one or more containers, with shared storage/network resources, and a specification for how to run the containers.
- 一般不会单独创建
- Pods have life cycles.

Deployment

- A Deployment provides declarative updates for Pods *ReplicaSets*.

- A ReplicaSet's purpose is to maintain a stable set of replica Pods running at any given time.

Service

- 如果 Deployment 中一些参数变化（例如 IP 地址），我们如何进行访问呢？
- 定义：An abstract way to expose an application running on a set of Pods as a network service.
- 如何指定对应的资源：Selector

Kubernetes AutoScaling

Kubernetes AutoScaling

- 动态调整资源，以达到期望的资源利用率（例：CPU 利用率）
- 三种方法
  - 基于节点（根据云服务商而定）
  - Vertical AutoScaling
  - Horizontal AutoScaling

实践

Kubernetes Service Discovery

Kubernetes Service Discovery

- 避免手动调整服务的网络地址
- 几种常见方式
  - InterApp Communication
  - Endpoint
  - NodePort
  - Load Balance
  - Ingress

实践

极客时间

扫码试看/订阅

《NLP 实战高手课》视频课程