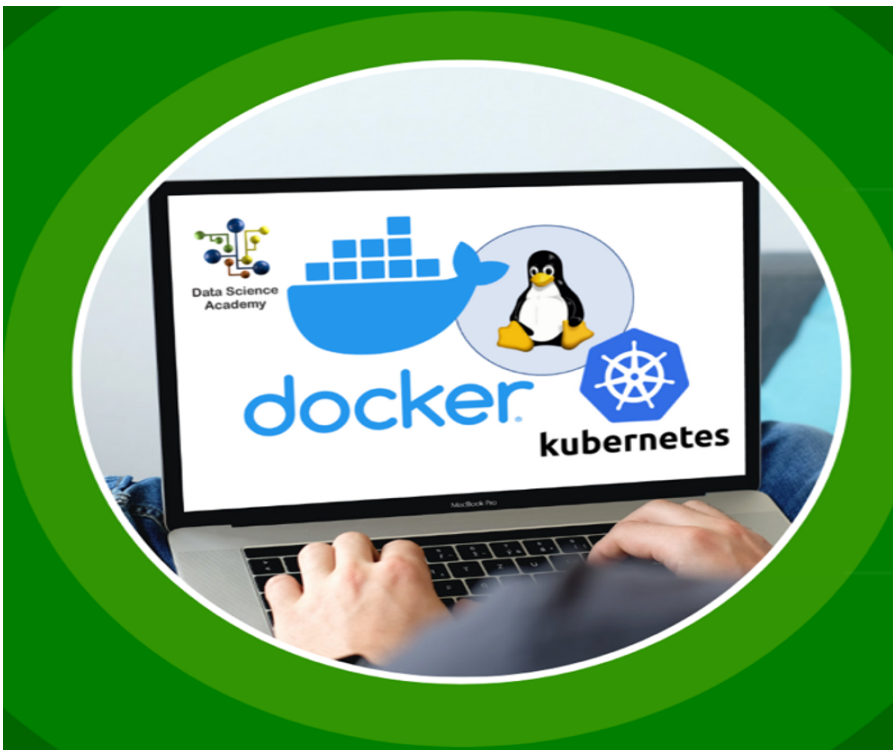
Curso Bônus

Planejando Sua Carreira Para as Profissões do Futuro

Lab 10 - Guia Para Orquestração do Pipeline de Engenharia de Dados



A orquestração de containers Docker para um pipeline de engenharia de dados envolve gerenciar e coordenar vários containers que executam diferentes etapas do pipeline, como extração, transformação, carregamento e processamento de dados.

Aqui está um guia para a orquestração de containers Docker para seu pipeline de engenharia de dados:

1- Divida seu pipeline de engenharia de dados em componentes:

- Identifique as etapas do seu pipeline, como extração de dados (ETL), processamento, armazenamento e análise.
- Para cada etapa, crie um componente independente que possa ser executado em um container.

2- Crie imagens Docker para cada componente:

- Escreva um Dockerfile para cada componente do pipeline, especificando a imagem base, dependências, arquivos de configuração e instruções para executar o componente.
- Construa as imagens Docker para cada componente usando o comando docker build.

3- Escolha uma ferramenta de orquestração:

- Docker Compose: Adequado para desenvolvimento local e pequenos aplicativos em produção.
- Docker Swarm: Bom para aplicações em produção que requerem escalabilidade e alta disponibilidade.
- Kubernetes: Uma escolha popular para aplicativos em produção que exigem escalabilidade, alta disponibilidade e gerenciamento avançado de recursos.
- Apache Airflow: Especialmente adequado para pipelines de engenharia de dados com fluxos de trabalho complexos e dependências de tarefas.

4- Configure sua ferramenta de orquestração:

a) Docker Compose:

- Crie um arquivo docker-compose.yml no diretório raiz do seu projeto.
- Defina os serviços (containers), volumes e redes necessários para seu pipeline.
- Inicie os serviços com o comando docker-compose up e pare-os com docker-compose down.

b) Docker Swarm:

- Inicialize o cluster Docker Swarm com o comando docker swarm init.
- Crie um arquivo docker-stack.yml semelhante ao arquivo docker-compose.yml, mas com recursos específicos do Swarm (como replicação e estratégias de atualização).
- Implante a pilha no Swarm com o comando docker stack deploy -c docker-stack.yml <stack_name>.

c) Kubernetes:

- Instale e configure o Kubernetes ou use um provedor de serviços gerenciado (como Google Kubernetes Engine ou Amazon EKS).
- Crie arquivos de configuração YAML para definir os recursos do Kubernetes necessários para seu pipeline (por exemplo, Deployments, Services, Persistent Volumes).

- Use o comando kubectl apply -f <config_file.yaml> para criar e gerenciar recursos no cluster Kubernetes.

d) Apache Airflow:

- Crie um ambiente Airflow usando Docker Compose ou Kubernetes (ou instale o Airflow em um ambiente não containerizado).
- Escreva DAGs (Directed Acyclic Graphs) para definir as tarefas e dependências do seu pipeline, usando operadores para executar containers Docker ou executar tarefas em outras plataformas (como serviços na nuvem).
- Implante suas DAGs no Airflow e use a interface da web ou a API para gerenciar e monitorar a execução do pipeline.

5- Monitore e gerencie seus containers:

- Use ferramentas e métricas fornecidas pela sua ferramenta de orquestração e monitore seu pipeline executado através de containers.

A orquestração de pipeline de engenharia de dados é assunto estudado na prática na Formação Engenheiro de Dados aqui na DSA.