

Introdução a Ciência de Dados

Universidade Federal de Pernambuco - UFPE
Centro de Informática - CIn

Professor: Renato Vimieiro

Alunos: Pedro Jose de Souza Neto e Rafael Felipe Pedroza Jordão

Proposta

Analisar os repositórios de projetos open source disponíveis no github que utilizem o Docker. A ideia é obter dados quanto a utilização da ferramenta e tracejar com as boas práticas DevOps, angariando informações contidas no arquivo Dockerfile, tais como: quais outras imagens utilizam? qual o sistema operacional? se utilizam configurações personalizadas de rede, quais pacotes são instalados em conjunto com o projeto? quais as linguagens de computação mais utilizadas nesses projetos? quais comandos comuns executados antes e durante a inicialização do container? quais as boas práticas de continuidade do negócio adotadas (escalabilidade, resiliência, backup, replicação dos dados)? quais são as aplicações mais (re)utilizadas em containers? entre outras.

Etapas de Execução

1. Estudar boas práticas para criação do Dockerfile
 - a. Entender a estruturação deste documento para identificar e capturar todas as informações disponíveis que sejam úteis para uma possível análise
2. Estudar a API de Github
 - a. Fazer uma análise dos métodos disponibilizados pelo Github que possa nos auxiliar na extração de dados dessa ferramenta
3. Filtrar os repositórios

- a. Identificar os repositórios que contém o Dockerfile e que sua última atualização tenha sido a partir do Janeiro de 2015
4. Extrair os dados do Dockerfile
 - a. Capturar as informações que serão utilizadas para análise
5. Realizar a limpeza dos dados
6. Identificar as tecnologias que estão sendo mais utilizadas no ambiente do Docker
7. Inferir padrões de relacionamento entre linguagens de programação e seu ambiente de configuração
8. Inferir padrões de relacionamento entre as aplicações/containers mais utilizados e boas práticas de continuidade de negócio
9. Inferir arquiteturas de "containerização"
10. Inferir boas práticas DevOps