

Guia de Carreira

Engenheiro

DataOps



Equipe Data Science Academy
www.datascienceacademy.com.br



Conteúdo do Guia

- Acompanhe o descritivo dos conteúdos deste guia:

Página	Descritivo
1	Capa
2	Conteúdo do Guia
3	Bem-vindo(a)! / O Que é DataOps?
4	Quem é o Engenheiro DataOps?
5	Principais Habilidades do Engenheiro DataOps
6 e 7	Como Implementar DataOps / DataOps x MLOps
8	Caso de Uso
9	Ferramentas de DataOps
10	Ferramentas e Frameworks Comuns
11, 12	E Por Onde Começar em 5 Passos?
13	Como se Preparar?
14	Créditos



Bem-vindo(a)!

Ansioso por desbravar o universo da Ciência de Dados e não sabe por onde começar? Nós ajudaremos você. Preparamos um guia que vai ajudá-lo(a) a compreender o que faz um **Engenheiro DataOps** como iniciar sua preparação! Confira.

O Que é DataOps?



DataOps, uma abreviação de “Data Operations”, é uma metodologia automatizada, orientada por processos, que visa melhorar a qualidade e reduzir o ciclo de vida dos sistemas de análise de dados. Inspirada nas filosofias do DevOps e do Agile, o DataOps foca na colaboração e comunicação entre Analistas de Dados, Engenheiros de Dados, Cientistas de Dados e profissionais de operações de TI, com o objetivo de melhorar a coordenação e a agilidade na gestão de dados. Essa abordagem enfatiza a automação de processos de dados, a integração contínua e a entrega contínua (CI/CD) de dados e insights, buscando reduzir os tempos de ciclo, aumentar a eficiência e garantir a qualidade e a confiabilidade dos dados.



Quem é o Engenheiro DataOps?

O Engenheiro DataOps é um profissional especializado na aplicação de práticas de DevOps ao ciclo de vida dos dados. O Engenheiro DataOps é um profissional especializado na interseção entre operações de dados (Data Operations ou DataOps) e engenharia de software.

Essa função é essencial no gerenciamento e otimização de fluxos de dados dentro de uma organização, garantindo que os dados sejam acessíveis, confiáveis e utilizáveis para análise e tomada de decisão.



Com a crescente demanda por tomada de decisão baseada em dados e as aplicações de IA cada vez mais comuns, as empresas buscam otimizar o fluxo de dados, desde a coleta até a análise. O Engenheiro DataOps desempenha seu trabalho exatamente na orquestração desses processos, garantindo que os dados sejam confiáveis, acessíveis e utilizados eficientemente.

O **Engenheiro DataOps** é um profissional indispensável na era moderna da Ciência de Dados. Com um conjunto de habilidades diversificado e a capacidade de otimizar o fluxo de dados em uma organização, eles desempenham um papel fundamental na promoção de uma cultura de dados ágil e eficiente.

À medida que o campo de dados continua a evoluir, a demanda por esses profissionais especializados certamente crescerá, tornando a capacitação em DataOps uma jornada valiosa.



Principais Habilidades do Engenheiro DataOps

- **Entendimento de Ciclo de Vida de Dados:** Conhecimento aprofundado de como os dados são coletados, processados, armazenados e analisados.
- **Automação de Processos:** Habilidade para automatizar fluxos de trabalho de dados e processos de integração.
- **Conhecimento em Infraestrutura de Dados:** Proficiência em tecnologias de armazenamento e processamento de dados.
- **Práticas de DevOps:** Familiaridade com CI/CD, containerização e configuração de infraestrutura como código (IaC).
- **Segurança de Dados e Compliance:** Entendimento das práticas de segurança de dados e regulamentações relevantes.



Como Implementar DataOps?



DataOps é o resultado da aplicação dos princípios do DevOps ao ciclo de vida dos dados. A ideia básica em DataOps é: “se você construir um sistema em torno dos dados – que automatize muito do monitoramento, implantação e colaboração sua produtividade aumenta, seus clientes ficam muito mais felizes e você acaba fazendo um trabalho melhor”.

A implementação de DataOps envolve vários passos, incluindo:

1. Definição de Processos: Defina os processos e fluxos de trabalho para os pipelines de dados, incluindo a integração, validação, teste e implantação.



2. Automação: Automatize tarefas repetitivas para melhorar a eficiência e a precisão. Isso inclui a automação de testes, implantações e atualizações.



3. Colaboração: Crie uma equipe cross-funcional de desenvolvimento de software e de dados para trabalhar juntos na criação, manutenção e monitoramento dos pipelines de dados.



4. Monitoramento: Monitore o desempenho de cada pipeline de dados para identificar problemas e oportunidades de melhoria.





5.Feedback: Implemente um sistema de feedback para permitir que as equipes de desenvolvimento de software e de dados possam compartilhar informações e soluções em tempo real.



6.Cultura: Fomente uma cultura de experimentação, inovação e melhoria contínua para garantir que todos estejam sempre procurando formas de tornar cada pipeline de dados mais eficiente e eficaz.



A implementação de DataOps é um processo contínuo e pode levar algum tempo. No entanto, a implementação correta pode melhorar significativamente a qualidade, confiabilidade e agilidade na entrega de dados.

DataOps x MLOps

DataOps é uma abordagem de gerenciamento de dados que tem como objetivo aumentar a velocidade, qualidade e eficiência do ciclo de vida dos dados. DataOps se concentra em automatizar e otimizar processos de coleta, armazenamento, processamento e distribuição de dados.

Já MLOps é uma extensão do DataOps que se concentra especificamente no gerenciamento do ciclo de vida dos modelos de aprendizado de máquina (Machine Learning). Isso inclui tarefas como o treinamento, o teste e o monitoramento de modelos, bem como a implementação e o gerenciamento de modelos em produção.

Enquanto DataOps pode ser responsabilidade de um Engenheiro de Dados ou de um Engenheiro DataOps, MLOps é responsabilidade do Engenheiro de Machine Learning.



Caso de Uso

Aqui está um exemplo de caso de uso para implementar DataOps na prática:

- **Identificação das necessidades de dados:** A primeira etapa é compreender as necessidades de dados da empresa. Isso pode incluir a definição de KPIs, o entendimento do fluxo de dados e a identificação dos dados críticos para o negócio.
- **Criação de pipelines de dados:** Uma vez que as necessidades de dados são conhecidas, é hora de criar pipelines de dados para coletar, processar e distribuir os dados. Isso pode ser feito usando uma ferramenta de integração de dados, como Apache Airflow ou Talend.
- **Automatização de processos:** A próxima etapa é automatizar processos como a validação de dados, a geração de relatórios e a distribuição de dados. Isso pode ser feito usando scripts ou ferramentas de automação, como Airbyte.
- **Monitoramento e otimização de pipelines de dados:** É importante monitorar continuamente o desempenho dos pipelines de dados para identificar problemas e oportunidades de otimização. Isso pode ser feito usando ferramentas de monitoramento, como Amazon CloudWatch.
- **Colaboração e documentação:** Por fim, é importante promover a colaboração entre equipes e documentar processos para garantir a transparência e a escalabilidade. Isso pode ser feito usando ferramentas de colaboração, como Confluence, e a criação de documentação detalhada dos processos de dados.



Ferramentas de DataOps



Algumas das principais ferramentas de DataOps incluem:

- **Apache Airflow**: um sistema de orquestração de pipelines de dados baseado em tarefas.
- **AWS Glue**: um serviço de ETL da Amazon Web Services que permite a criação, execução e gerenciamento de pipelines de dados.
- **Talend**: uma plataforma de integração de dados que oferece ferramentas para coletar, integrar e distribuir dados.
- **Airbyte**: ferramenta open-source para movimentação de dados entre sistemas heterogêneos.
- **StreamSets**: uma plataforma de gerenciamento de dados que permite a criação, execução e monitoramento de pipelines de dados.
- **DataKitchen**: uma plataforma de automação em DataOps.



Ferramentas e Frameworks Comuns



- **Linguagens de Programação:** Python, Scala, Java.
- **Ferramentas de Orquestração de Dados:** Apache Airflow, Luigi.
- **Plataformas de Big Data:** Databricks, Apache Spark.
- **Ferramentas de Processamento e Integração de Dados:** Airbyte, DBT, SQL.
- **Tecnologias de Containerização:** Docker, Kubernetes.
- **Ferramentas de CI/CD:** Jenkins, GitLab CI.
- **Plataformas de Cloud Computing:** AWS, Azure, Google Cloud Platform.



E Por Onde Começar em 5 Passos?

Passo 1:

Infraestrutura Como Código com Terraform AWS Azure e Databricks:

Desenvolva suas habilidades com o Terraform, uma ferramenta open-source que permite definir e gerenciar infraestrutura como código. Isso inclui a implantação e gerenciamento de infraestrutura em vários provedores de cloud computing, como AWS e Azure, usando uma linguagem simples e declarativa.

Passo 2:

Engenharia de Dados com Airbyte DBT e SQL:

Aprimore suas habilidades em engenharia de dados através de ferramentas essenciais como Airbyte, DBT e SQL. Essas habilidades são fundamentais para extrair, transformar e carregar dados (ETL), bem como para a construção e manutenção de pipelines de dados robustos.

Passo 3:

Orquestração de Fluxos de Dados com Apache Airflow:

Aprenda a orquestrar fluxos de dados com o Apache Airflow, uma ferramenta poderosa para o agendamento e monitoramento de fluxos de dados. Essa competência é fundamental para gerenciar e automatizar complexos pipelines de dados e de Machine Learning.



Passo 4:

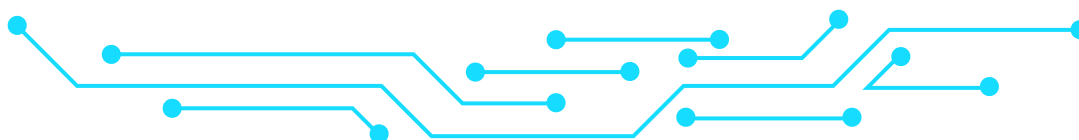
Pipelines de ETL e Machine Learning com Apache Spark

Entrar no universo do Apache Spark é embarcar em uma jornada rumo ao coração das operações de dados mais avançadas de hoje. Construir pipelines de dados e de Machine Learning faz parte do dia a dia do Engenheiro DataOps.

Passo 5:

Pipelines de CI/CD Para Operações de Machine Learning e IA

Desenvolva uma visão profunda das práticas e ferramentas essenciais necessárias para construir, gerenciar e otimizar Pipelines de CI/CD em projetos de Machine Learning e Inteligência Artificial.



Seguir esses passos ajudará a desenvolver as competências técnicas necessárias para uma carreira bem-sucedida em DataOps, permitindo a automação e otimização de processos de dados.



Como Se Preparar?

Formação Engenheiro DataOps

Transformação Digital na Era da Inteligência Artificial



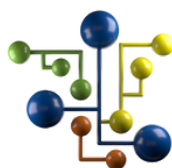
100% Online e 100% em Português



Com Certificado de Conclusão
em Português ou Inglês



www.datascienceacademy.com.br





Créditos

Equipe Data Science Academy

- Equipe DSA, 2024, Guia de Carreira Engenheiro DataOps.
- Versão 4.0
- Site: www.datascienceacademy.com.br