

Data Science Academy



Guia de Carreira

ENGENHEIRO DE DADOS

Equipe Data Science Academy

WWW.DATASCIENCEACADEMY.COM.BR

Conteúdo do Guia

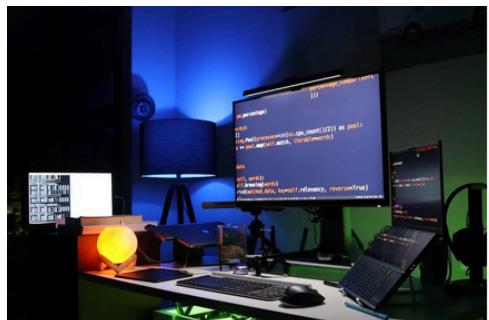
► Acompanhe o descritivo dos conteúdos deste guia:

Página	Descriutivo
1	Capa
2	Conteúdo do Guia
3	Bem-vindo / Quer se Tornar um Engenheiro de Dados?
4, 5 e 6	Mas afinal, o que é Engenharia de Dados?
7	Cientista de Dados x Engenheiro de Dados
8	Perfil Profissional do Engenheiro de Dados
9	Por Onde Começar em 7 Passos
10	1 - Faça uma Auto-Análise
11	2 - Aprenda a Trabalhar com Banco de Dados Relacionais e Linguagem SQL
12	3 - Aprenda a Projetar, Implementar e Manter Data Lakes
13	4 - Desenvolva Suas Habilidades em Cloud Computing e Computação Distribuída
14	5 - Desenvolva Suas Habilidades Sobre Governança e Segurança de Dados
15	6 - Domine Pelo Menos Uma Linguagem de Programação
17	7 - Domine o Sistema Operacional Linux
18	Quer Iniciar Sua Carreira como Engenheiro de Dados?
18	Como se Preparar?
19	Créditos

Bem-vindo(a)!

Ansioso por desbravar o universo da Engenharia 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 de Dados e como iniciar sua preparação! Confira.

Quer se Tornar um Engenheiro de Dados?



Neste Guia, nós definimos o que é Engenharia de Dados e quais as habilidades necessárias para praticá-la em um alto nível. Se você está interessado em projetar sistemas de grande escala ou trabalhar com enormes quantidades de dados, a Engenharia de Dados é um excelente campo de trabalho para você.

A Engenharia de Dados lida com escala e eficiência, e pode ser difícil encontrar material de boas práticas por conta própria. Mas não desista – é possível aprender Engenharia de Dados aqui na DSA com a Formação Engenheiro de Dados. Leia este guia e nós te mostraremos como!

Se você prefere trabalhar com infraestrutura de dados ao invés de análise de dados, a função de Engenheiro de Dados é a ideal para você.

Mas afinal, o que é Engenharia de Dados?

O campo da Ciência de Dados é incrivelmente amplo, abrangendo tudo, desde a limpeza de dados até a implantação de modelos preditivos. No entanto, é raro um único Cientista de Dados trabalhar em todas as áreas da Ciência de Dados. Os Cientistas de Dados geralmente se concentram em algumas áreas e são complementados por uma equipe de outros Cientistas e Analistas.

A Engenharia de Dados também é um campo amplo. Vejamos primeiro o que é a Engenharia de Dados e depois caminhemos através de descrições mais específicas que ilustram as funções nesta área.

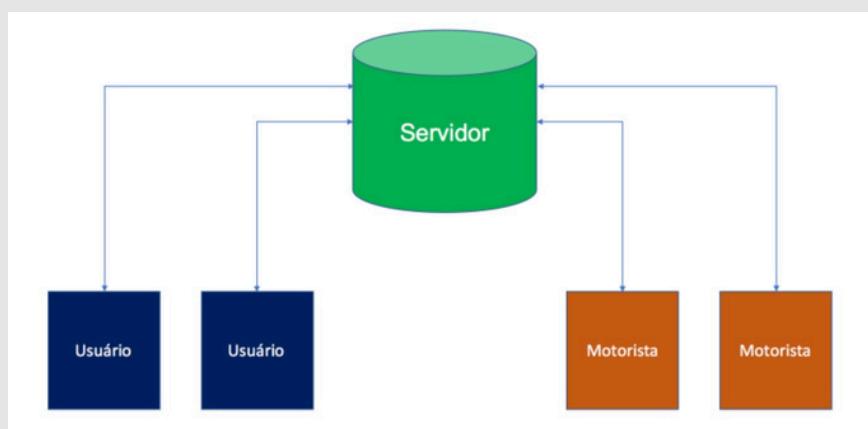
Um Engenheiro de Dados transforma dados em um formato útil para análise. Imagine que você é um Engenheiro de Dados que irá trabalhar em um concorrente do Uber. Seus usuários têm um aplicativo através do qual eles acessam o serviço. Eles pedem uma corrida para um destino através do aplicativo, que é encaminhado para um motorista, que os busca e os leva para onde foi solicitado. Após a corrida, eles são cobrados e têm a opção de avaliar como o motorista dirigiu.

Como Manter Este Serviço?

Para manter um serviço como este, você precisa:

- Um aplicativo para dispositivos móveis, para os usuários.
- Um aplicativo para dispositivos móveis, para os motoristas.
- Um servidor que pode passar pedidos de usuários para motoristas e lidar com outros detalhes, como atualizar informações de pagamento.

Aqui está um diagrama que ilustra esta comunicação:



Como você pode imaginar, esse tipo de sistema vai gerar enormes quantidades de dados. Você terá diferentes tipos de dados para serem armazenados:

- O banco de dados do seu aplicativo principal. Ele deve conter informações sobre o usuário e o motorista.
- Logs de análise de servidor:
 - Registros de acesso ao servidor. Estes contêm uma linha por solicitação feita ao servidor a partir do aplicativo.
 - Logs de erro do servidor. Estes contêm todos os erros do lado do servidor os quais foram gerados pelo seu aplicativo.
- Registros de análise de aplicativos:
 - Registros de eventos da aplicação.
 - Logs de erro da aplicação. Estes contêm informações sobre erros no aplicativo.
- Registros de corridas. Estes contêm informações sobre cada corrida para os usuários / motoristas e contêm informações sobre o status da corrida.
- Registros de Serviços ao Cliente. Estes contêm informações sobre interações com clientes por agentes de atendimento ao cliente. Podem incluir transcrições de voz e registros de e-mail.

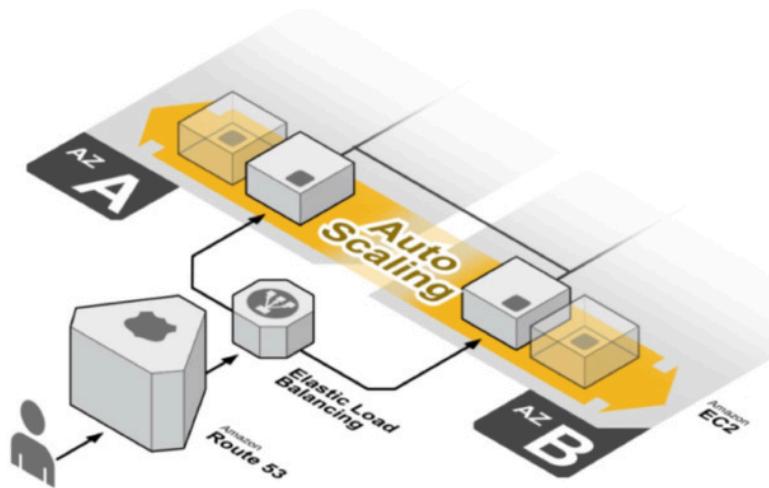
Digamos que um Cientista de Dados deseja analisar o histórico de ação de um usuário com seu serviço e ver quais ações se correlacionam com os usuários que gastam mais. Para isso, será necessário combinar informações dos logs de acesso do servidor e dos logs de eventos do aplicativo. Será necessário:

- Reunir regularmente logs de análise de aplicativos de dispositivos do usuário.
- Combinar os logs da análise do aplicativo com todas as entradas de log do servidor que fazem referência ao usuário.
- Criar uma API que retorna o histórico de eventos de qualquer usuário.

Para resolver isso, você precisará criar um pipeline que possa registrar o uso dos aplicativos móveis e logs do servidor em tempo real, analisá-los e anexá-los a um usuário específico. Em seguida, você precisará armazenar os registros analisados em um banco de dados, para que possam ser facilmente consultados pela API. Você precisará configurar vários servidores em modo de平衡amento de carga para processar os logs recebidos.

A maioria dos problemas que você vai encontrar serão relativos à confiabilidade dos sistemas distribuídos. Por exemplo, se você tiver milhões de dispositivos para reunir logs e demanda variável (de manhã, você recebe uma tonelada de logs, e quase nenhum à meia-noite), você precisará de um sistema que possa escalar automaticamente a contagem de servidores para cima e para baixo.

Neste momento entra em cena a necessidade de ter um ambiente em nuvem, robusto e seguro, que possa processar esse grande volume de dados. Uma estrutura em nuvem, usando um provedor como a AWS (Amazon Web Services), pode ser a melhor alternativa custo/benefício. Os servidores são registrados com um平衡ador de carga e o平衡ador de carga envia tráfego para eles com base em como eles estão ocupados. Isso significa que os servidores podem ser adicionados ou removidos conforme necessário. Esse fluxo é demonstrado na imagem abaixo:



Esse cenário típico que acabamos de descrever é um exemplo de tarefa realizada por Engenheiros de Dados, profissionais responsáveis por construir e manter a infraestrutura para que os dados possam seguir da origem para o destino e então possam ser usados para análises de todos os tipos. Se a possibilidade de construir esse tipo de infraestrutura é algo que desperta seu interesse, a função de Engenheiro de Dados pode ser a ideal para você. A boa notícia é que os salários são altos, a demanda é alta e há poucos profissionais capacitados disponíveis no mercado de trabalho.

Cientista de Dados x Engenheiro de Dados

Cientistas de Dados e Engenheiros de Dados trabalharão juntos em uma equipe de Data Science ou Analytics para criar uma solução de ponta a ponta nas empresas que necessitem de modelos analíticos avançados e que sejam operacionalizados em escala.



O Cientista de Dados está interessado em analisar conjuntos de dados, limpá-los, transformá-los, pré-processá-los, executar seus modelos preditivos de Machine Learning, buscar padrões e gerar insights para os tomadores de decisão.

Já o Engenheiro de Dados é o profissional que vai projetar, implementar e manter a infraestrutura necessária para assegurar que todo o processo analítico funcione corretamente, de forma segura e com boa performance. Cabe ao Engenheiro de Dados definir os processos para coletar, armazenar e processar os dados usados nos modelos analíticos criados pelo Cientista de Dados.

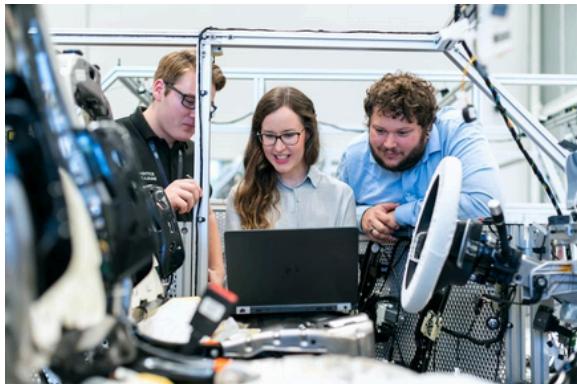
Os requisitos para um Cientista de Dados são tipicamente mais “científicos”, pois é necessário realizar pesquisas e saber como construir e testar modelos avançados. Para os Cientistas de Dados é necessário ênfase em Matemática e Estatística, os pilares da Ciência de Dados.

Os Engenheiros de Dados geralmente têm um perfil mais voltado para infraestrutura de tecnologia. Os profissionais interessados em projetar e construir arquiteturas em larga escala para aplicações intensivas em dados podem avançar para este campo, pois há muito menos ênfase em ciência e matemática e mais em engenharia de software, infraestrutura e arquitetura de soluções.

Engenheiro de Dados é uma função e não uma profissão regulamentada (pelo menos ainda). Logo, pode ser exercida por pessoas com qualquer formação acadêmica, desde que tenham o conhecimento necessário, obviamente.

Perfil Profissional do Engenheiro de Dados

Os Engenheiros de Dados devem entender os conceitos fundamentais em ciência da computação e devem ser bem versados na construção e concepção de aplicações em grande escala; de ponta a ponta.



Eles devem entender os prós e os contras da utilização de bancos de dados relacionais e NoSQL, eles devem saber como projetar soluções para casos de uso com dados em lote (Batch) e fluxo contínuo (Streaming de Dados), eles devem saber o que é preciso para operacionalizar um modelo preditivo e como ajudar a publicar esses modelos criados no “laboratório” (treinamento e validação) para aplicações em tempo real.

Eles devem entender a computação distribuída e ambientes em cluster e devem ser capazes de trabalhar com o Cientista de Dados para ajudar a construir a infraestrutura de armazenamento e processamento de dados.

Eles devem saber quando definir esquemas a fim de implementar projetos de “Data Lakes” que ajudem na análise em grande escala, mas ainda atendam aplicativos específicos do domínio.

Eles devem ainda conhecer sobre governança e segurança de dados e eles devem estar muito familiarizados com as principais tecnologias que são usadas para construir esses sistemas.

Inevitavelmente surge a pergunta: e por onde eu começo? A seguir os 7 passos que nós sugerimos, fruto da experiência capacitando milhares de profissionais aqui na Data Science Academy.

Por Onde Começar em 7 Passos

Agora que você tem uma boa visão sobre o que é e o que faz um Engenheiro de Dados e decidiu que deseja trabalhar nessa função, aqui estão os 7 passos por onde começar.



1- Faça Uma Auto-Análise



Antes de pensar na carreira de Engenheiro de Dados, faça uma auto-análise sobre seu nível atual de conhecimento. Use papel e caneta se necessário e liste as tecnologias que você conhece atribuindo notas de 1 a 10 para o nível de conhecimento que você acredita ter.

Pode ser meio complicado no começo, mas isso vai dar uma boa ideia de onde você está e o que precisa fazer para adquirir o conhecimento necessário para trabalhar como Engenheiro de Dados. Essas perguntas podem ajudar:

- Qual o seu nível atual de conhecimento sobre tecnologias de armazenamento e processamento de dados?
- Você se sente confortável em trabalhar com sistema operacional Linux?
- Você comprehende por que computação distribuída é fundamental ao armazenar e processar grandes volumes de dados?
- Você sabe trabalhar com Cloud Computing? Domina algum provedor em nuvem?
- Qual seu nível de proficiência em programação de computadores?
- Domina ou tem alguma proficiência em ferramentas ETL?
- Saberia extrair dados em tempo real, usando tecnologia local ou em nuvem?

Respondendo essas e outras perguntas você terá uma visão mais clara das áreas onde precisa desenvolver seus skills.

2- Aprenda a Trabalhar com Bancos de Dados Relacionais e Linguagem SQL

Bancos de dados relacionais farão parte de projetos de Engenharia de Dados, seja como fonte, como destino ou durante a execução de um pipeline de dados.

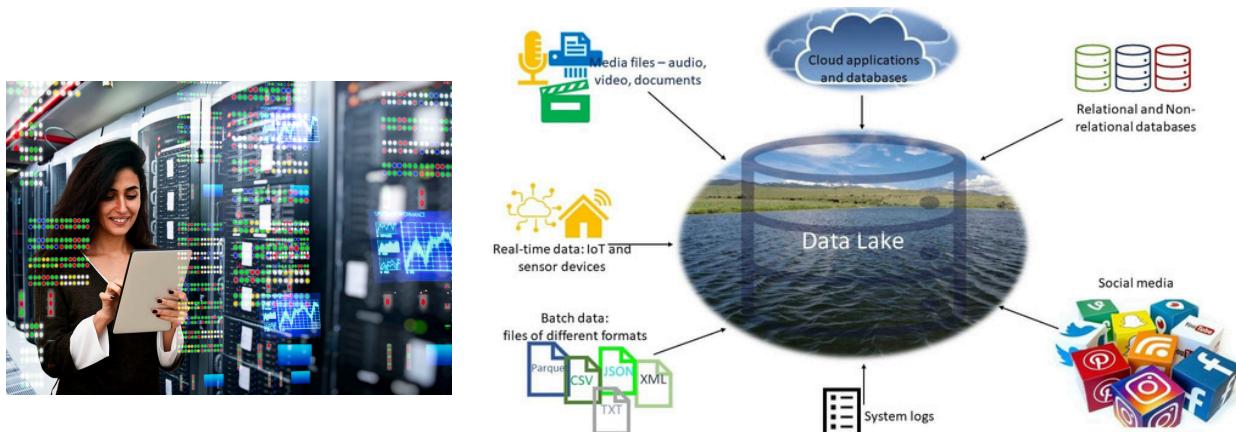


Saber trabalhar com bancos de dados relacionais é, portanto, mandatório para Engenheiros de Dados. E o que significa “saber trabalhar”?

Significa conhecer o funcionamento do SGBD (Sistema Gerenciador de Banco de Dados), conhecer as alternativas de otimização e tuning do banco de dados gerenciado pelo SGBD, conhecer os modelos de dados (como modelo transacional para bancos transacionais e modelo dimensional para Data Warehouses), conhecer como os dados podem ser carregados e extraídos de um banco de dados.

SQL é a principal linguagem usada para manipular dados em bancos de dados e amplamente usada em processos ETL (Extração, Transformação e Carga de Dados). SQL fará parte do seu dia a dia como Engenheiro de Dados e conhecer algumas ferramentas usadas em ETL como Airbyte, Apache Spark e Apache Airflow pode ser necessário para entrar no mercado de trabalho.

3- Aprenda a Projetar, Implementar e Manter Data Lakes



Existem muitas formas e tecnologias para projetar, implementar e manter Data Lakes, sendo Apache Hadoop HDFS, Bancos de Dados NoSQL e Data Stores como o Amazon S3 na nuvem, as tecnologias mais comuns para esse fim.

O Engenheiro de Dados deve dominar o máximo possível sobre Data Lakes e seu tipo de armazenamento. Cada vez mais e mais empresas implementam Data Lakes à medida que avançam em suas estratégias de dados. O Data Lake está se tornando o coração de muitas estratégias de dados.

Portanto, saber projetar, implementar e manter um Data Lake é requisito básico para Engenheiros de Dados. Dominar ferramentas ETL e ELT que carregam e extraem dados de Data Lakes é igualmente importante.

E aqui podemos incluir também o conhecimento em extração de dados em tempo real, ou seja, saber trabalhar com Streaming de Dados, seja para análise em tempo real ou para armazenamento. O Apache Kafka é um exemplo de ferramenta que pode ser usada nesse tipo de tarefa.

Além de saber implementar Data Lakes, é importante saber como implementar uma das suas variantes, o Data Lakehouse (Data Lake + Data Warehouse), especialmente nos ambientes de Cloud Computing.

4- Desenvolva Suas Habilidades em Cloud Computing e Computação Distribuída

Todos estão indo para a nuvem.

Dominar Cloud Computing agora é um requerimento básico para profissionais de tecnologia em geral e no caso do Engenheiro de Dados isso é ainda mais relevante.



Dificilmente um Engenheiro de Dados participará de um projeto onde alguma solução em nuvem não seja utilizada e dominar pelo menos um provedor de Cloud Computing é fundamental. Os principais provedores em nuvem são AWS, Microsoft Azure e Google Cloud Platform.

A computação distribuída em clusters de computadores é a base para o armazenamento e processamento de grandes volumes de dados e quase todos os frameworks usados para esse fim são baseados no conceito de computação distribuída. Esse conhecimento é requisito básico para Engenheiros de Dados.

Você pode adquirir o conhecimento sobre computação distribuída construindo um ambiente de cluster no seu computador através de máquinas virtuais, exatamente como ensinamos na Formação Engenheiro de Dados aqui na DSA. Uma forma prática de adquirir o conhecimento que será usado no dia a dia, montando um laboratório no seu próprio computador.

5- Desenvolva Suas Habilidades Sobre Governança e Segurança de Dados

Ao implementar uma arquitetura de dados e construir um pipeline de extração, transformação e carga de dados, duas questões farão inevitavelmente parte do trabalho do Engenheiro de Dados.

Primeiro a Governança de Dados e Metadados, o que será usado para organizar, pesquisar, selecionar e descartar dados de forma criteriosa.

Segundo, algo bastante óbvio: segurança. Se os dados estão se tornando um dos ativos mais importantes da empresa, a segurança de armazenamento e acesso se torna algo crítico.

- **Quem pode acessar os dados no Data Lake?**
- **Um departamento pode ter acesso de leitura a dados de outro departamento?**
- **Os dados devem ser criptografados?**
- **Qual a política de backup dos dados?**
- **Qual a política de anonimização dos dados?**

Todas essas perguntas devem ser respondidas e uma política de segurança de acesso aos dados deve ser criada, implementada e mantida.

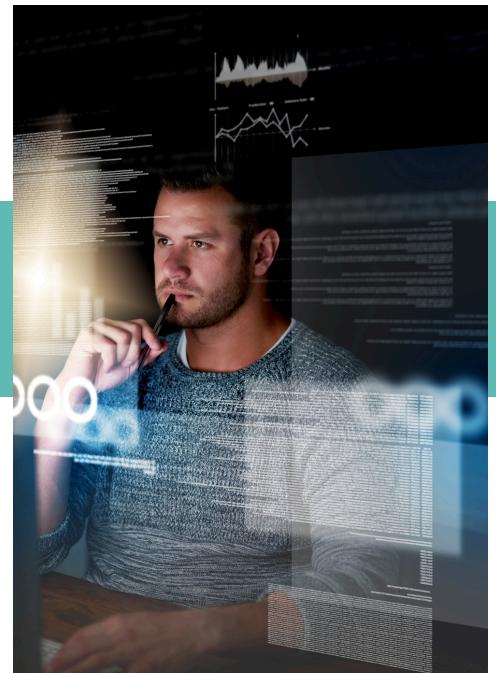
É importante conhecer as regras estabelecidas na LGPD (Lei Geral de Proteção aos Dados) e o impacto na construção dos pipelines de dados.

6 - Domine Pelo Menos Uma Linguagem de Programação

O Engenheiro de Dados precisa saber programar?

Não e Sim!

Vamos explicar!



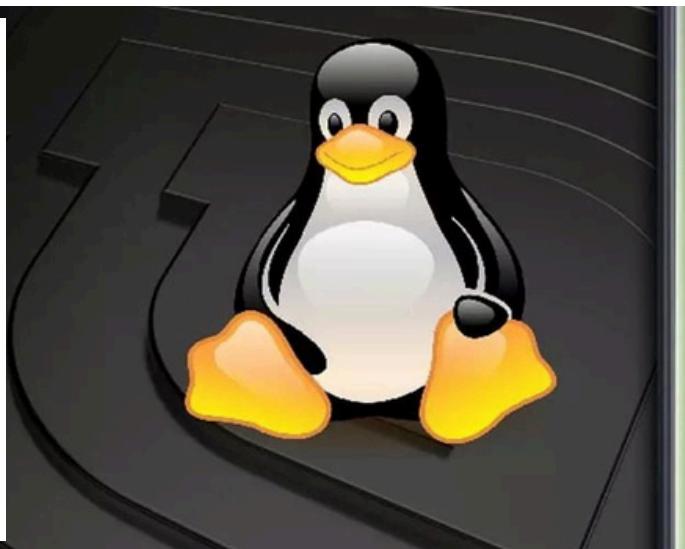
O conhecimento em programação de um Engenheiro de Dados não precisa ser do nível de um Engenheiro de Software, mas é importante dominar pelo menos uma linguagem, pois isso vai permitir construir pipelines de dados personalizados e customizar as soluções de dados para as necessidades da empresa.

Além disso, todo e qualquer software é construído com uma linguagem de programação, certo? Dominar uma linguagem, portanto, vai permitir que você leia e compreenda mensagens de erro, faça debug no caso de eventuais problemas e até mesmo corrija pequenos bugs ou faça customização nas ferramentas quando forem open-source, o que predomina em projetos de Ciência de Dados.

Python, Java e Scala são as linguagens mais comumente usadas por Engenheiros de Dados no dia a dia.

7 - Domine o Sistema Operacional Linux

Raramente (muito raramente),
você encontrará um
Engenheiro de Dados que não
saiba trabalhar com Linux, seja
localmente ou em nuvem.

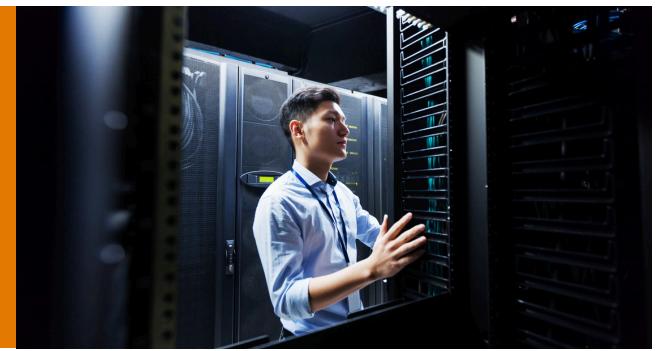


Todas as principais tecnologias de armazenamento e processamento de dados foram criadas para o universo Unix e Linux é um tipo de Unix. Não tente lutar contra isso. Use seu Windows para assistir vídeos no YouTube ou navegar pela internet, mas na hora de trabalhar de forma profissional busque sua proficiência em Linux.

Em diversas situações, ao acessar remotamente um servidor na nuvem por exemplo, tudo que estará à disposição é um terminal para executar as ações necessárias de configuração. Saber trabalhar com Linux é fundamental para quem pretende seguir a carreira de Engenheiro de Dados. Exatamente por isso oferecemos o curso gratuito ***Sistema Operacional Linux, Docker e Kubernetes***, para quem adquire a [Formação Engenheiro de Dados](#).

Engenheiro de Dados é uma das funções com maior remuneração no universo da Ciência de Dados. Prepare-se e busque sua vaga!

**Quer iniciar sua carreira como
Engenheiro de Dados e se tornar um
profissional altamente requisitado no
mercado de trabalho?**



Que tal começar com uma Formação completa, com cursos 100% online e 100% em português? E se os cursos forem totalmente práticos, com exemplos e casos de uso do mundo real? E se você tivesse suporte em até 24 horas, até mesmo nos finais de semana e feriados? E se você tivesse a chance de fazer networking de qualidade com profissionais de alto nível em todo Brasil e no exterior? E se ao final você tivesse um certificado de conclusão para compartilhar nas suas redes sociais e demonstrar suas habilidades aos empregadores? E se você ainda pudesse estudar do smartphone ou tablet?

Tudo isso e muito mais já é possível.

Clique no link abaixo e comece agora mesmo:

Formação Engenheiro de Dados

Este guia tem como objetivo ajudar você a entender um pouco melhor como se preparar e se tornar um profissional de Engenharia de Dados. Faça da sua jornada de aprendizagem uma experiência prazerosa e divertida! De qualquer forma o resultado será recompensador.

Como se Preparar?

Formação Engenheiro de Dados

Sua Carreira Elevada a Outro Nível.

Formação Engenheiro de Dados 4.0



Transformação Digital na Era da Inteligência Artificial

64h

A qualquer hora, em qualquer lugar. 100% Online 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 de Dados.
- Versão 4.0
- Site: www.datascienceacademy.com.br