



Fundamentos de Engenharia de Dados

Fundamentos de Engenharia de Dados

Engenheiro de Dados x Cientista de Dados

Antes de mais nada, compreenda que Cientistas de Dados e Engenheiros de Dados são profissionais diferentes, com perfis diferentes e que usam essencialmente ferramentas diferentes. É claro que nada impede que um único profissional tenha habilidades para exercer as duas funções e em empresas menores ou Startups isso será uma realidade.

Mas compreender as diferenças ajuda a definir o perfil adequado de cada profissional.

O Cientista de Dados desenvolve modelos e realiza análises usando Matemática, Estatística, Programação e Machine Learning para explicar e prever comportamentos complexos, resolvendo problemas das áreas de negócio, no mundo real.

Um Engenheiro de Dados projeta e constrói arquiteturas de dados e pipelines para ingestão, armazenamento, processamento e execução de aplicações de grande escala com Big Data.

“Ciência de Dados” e “Machine Learning” (ML) são disciplinas relacionadas a projetos que tendem a ser concluídos por indivíduos com títulos como “Cientista de Dados”.

Os Cientistas de Dados geralmente estão acostumados a trabalhar com todos os tipos de dados e podem usar os mesmos Data Lakes e várias ferramentas de preparação de dados que os Engenheiros de Dados usam.

No entanto, os Cientistas de Dados geralmente transformam seus dados com o objetivo final de lidar com a Ciência de Dados ou problemas de ML, enquanto os Engenheiros de Dados estão mais interessados em criar processos de engenharia para dar suporte a outras partes da empresa.

Embora possam usar as mesmas ferramentas os propósitos são diferentes. Engenheiros de Dados estão preocupados com o fluxo de dados. Cientistas de Dados estão preocupados com o processo científico de análise dos dados e Machine Learning.

Os Engenheiros de Dados geralmente coletam dados de diferentes fontes, transformam os dados em diferentes formatos e, em seguida, entregam os dados a Analistas de Dados, Cientistas de Dados ou Engenheiros de IA. Essa “entrega” pode se dar por meio de repositórios de dados como Data Warehouses e Data Lakes, por meio de APIs de acesso, por meio de containers, por meio de pipelines ou outras opções.