A Importância da Engenharia de Dados para a Ciência de Dados

Olá! Seja bem-vindo(a) à nossa newsletter semanal sobre Ciência de Dados!
Nesta edição, vamos explorar a importância da engenharia de dados e sua correlação com a ciência de dados. A engenharia de dados é uma disciplina fundamental que viabiliza a coleta, armazenamento e processamento de grandes volumes de dados, tornando-os acessíveis e utilizáveis para análise e modelagem. De forma mais detalhada... a engenharia de dados é uma área da tecnologia da informação que se concentra na criação e gestão de infraestruturas de dados. Envolve o design, construção e manutenção de sistemas que coletam, armazenam, processam e transformam dados brutos em formatos utilizáveis e acessíveis para análise. A engenharia de dados é essencial para suportar as necessidades de análise de dados, aprendizado de máquina e tomada de decisão baseada em dados.

A Relação Entre Engenharia de Dados e Ciência de Dados

Enquanto os cientistas de dados se concentram em extrair insights e criar modelos preditivos a partir dos dados, os engenheiros de dados garantem que esses dados estejam disponíveis, limpos e organizados. Sem uma infraestrutura robusta de dados, o trabalho dos cientistas de dados seria significativamente limitado. Portanto, a engenharia de dados é uma base crítica para qualquer projeto de ciência de dados bem-sucedido.
A seguir, apresentamos uma tabela com as dez principais tarefas realizadas por um engenheiro de dados:

Tarefa	Detalhamento dessas principais Tarefas	Exemplo Real
Aquisição de Dados	• Coleta de dados de diversas fontes, como bancos de dados, APIs, sensores, logs de servidores e arquivos CSV. • Garantir a integridade e a qualidade dos dados adquiridos.	Uma empresa de e-commerce coleta dados de transações de clientes de seu banco de dados e de APIs de parceiros.
Armazenamento de Dados	• Design e implementação de bancos de dados, data Warehouse e data lakes para armazenar grandes volumes de dados. • Gerenciamento de esquemas de banco de dados e otimização de consultas para melhorar o desempenho.	Um banco utiliza um data lake para armazenar dados históricos de transações financeiras.
Processamento de Dados	• Criação de pipelines de dados para processar e transformar dados brutos em informações úteis. • Utilização de frameworks como Apache Hadoop, Apache Spark e Apache Kafka para processamento em larga escala.	Uma startup de saúde utiliza Apache Spark para processar grandes volumes de dados de dispositivos médicos.
Integração de Dados	• Integração de dados provenientes de múltiplas fontes para criar uma visão unificada. • Utilização de ferramentas de ETL (Extract, Transform, Load) para mover dados entre sistemas.	Uma empresa de logística integra dados de GPS de caminhões e sensores de temperatura para monitoramento.
Gerenciamento da Qualidade	• Implementação de práticas de limpeza de dados para remover inconsistências, duplicidades e erros. • Monitoramento contínuo da qualidade dos dados e implementação de controles para garantir a precisão e a confiabilidade dos dados.	Uma instituição financeira aplica limpeza de dados para remover duplicidades em registros de clientes.
Segurança e Governança	• Garantia da segurança dos dados, implementando controles de acesso e políticas de segurança. • Manutenção da conformidade com regulamentações de proteção de dados, como GDPR e LGPD.	Uma empresa de telecomunicações implementa políticas de acesso para proteger dados de clientes.
Modelagem de Dados	• Design de modelos de dados que representem a estrutura dos dados e suas relações. • Criação de diagramas ER (Entidade-Relacionamento) e mapeamento de dados.	Uma rede de supermercados cria um modelo de dados para analisar padrões de compras dos consumidores.
Otimização de Desempenho	• Monitoramento e otimização do desempenho de sistemas de dados para garantir tempos de resposta rápidos. • Implementação de técnicas de indexação, particionamento e caching para melhorar a eficiência.	Uma empresa de fintech otimiza consultas em seu banco de dados para acelerar processos de análise financeira.
Colaboração com Cientistas	• Trabalhar em estreita colaboração com cientistas de dados para entender as necessidades de dados e fornecer suporte técnico. • Preparação de conjuntos de dados para análise e desenvolvimento de modelos de aprendizado de máquina.	Uma equipe de engenharia de dados prepara conjuntos de dados para um projeto de aprendizado de máquina.
Documentação e Manutenção	• Documentação de processos, arquiteturas de dados e fluxos de trabalho. • Manutenção contínua das infraestruturas de dados para garantir sua disponibilidade e escalabilidade.	Uma corporação global mantém documentação detalhada de seus fluxos de dados e processos de ETL.

A engenharia de dados desempenha um papel essencial na construção de um ecossistema de dados robusto e eficiente, permitindo que os cientistas de dados concentrem seus esforços na análise e modelagem para gerar insights valiosos. A sinergia entre essas duas áreas é fundamental para o sucesso das iniciativas baseadas em dados.
Esperamos que você tenha achado estas informações úteis. Seja você um iniciante no mundo da Ciência de Dados ou um profissional experiente, entender a importância da engenharia de dados pode aumentar suas chances de sucesso em projetos e na carreira.
Se você tiver alguma dúvida ou sugestão, não hesite em entrar em contato. Queremos garantir que nossos conteúdos sejam sempre relevantes e úteis para você.
Fique atento à nossa próxima edição, onde continuaremos a explorar ferramentas técnicas e dicas valiosas para aprimorar suas habilidades em Ciência de Dados. Até lá, continue explorando e aprendendo!

Saudações,

Prof. Dr. Dilermando Piva Jr
Coordenador de Ciência de Dados para Negócios / Fatec Votorantim
E-mail: f301.cdn@fatec.sp.gov.br