

Fundamentos de Engenharia de Dados

Projete e Construa Sistemas de Dados Robustos

Joe Reis • Matt Housley

O'REILLY®
Novatec

Authorized Portuguese translation of the English edition of Fundamentals of Data Engineering
ISBN 9781098108304 © 2022 Joseph Reis and Matthew Housley. This translation is published
and sold by permission of O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

Tradução em português autorizada da edição em inglês da obra Fundamentals of Data Engineering
ISBN 9781098108304 © 2022 Joseph Reis and Matthew Housley. Esta tradução é publicada e
vendida com a permissão da O'Reilly Media, Inc., detentora de todos os direitos para publicação
e venda desta obra.

© Novatec Editora Ltda. [2023].

Todos os direitos reservados e protegidos pela Lei 9.610 de 19/02/1998. É proibida a reprodução
desta obra, mesmo parcial, por qualquer processo, sem prévia autorização, por escrito, do autor e
da Editora.

Editor: Rubens Prates

Assistente editorial: Karina Tsuchiya

Tradução: Márcio Martins

Revisão gramatical: Tássia Carvalho

ISBN do impresso: 978-85-7522-876-0

ISBN do ebook: 978-85-7522-877-7

Histórico de impressões:

Novembro/2023 Primeira edição

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110

02460-000 — São Paulo, SP — Brasil

Tel.: +55 11 2959-6529

Email: novatec@novatec.com.br

Site: <https://novatec.com.br>

Twitter: twitter.com/novateceditora

Facebook: facebook.com/novatec

LinkedIn: linkedin.com/company/novatec-editora

GRA20231031

Sumário

Prefácio	17
Parte I ■ Fundamentos e componentes essenciais da engenharia de dados	25
Capítulo 1 ■ Descrição do que é engenharia de dados	26
O que é engenharia de dados?	26
Definição de engenharia de dados	28
O ciclo de vida da engenharia de dados	28
Evolução do engenheiro de dados	29
Engenharia de dados e ciência de dados	36
Habilidades e funções da engenharia de dados	38
Maturidade de dados e o engenheiro de dados	39
Formação e habilidades de um engenheiro de dados	43
Responsabilidades comerciais	44
Responsabilidades técnicas	45
As diversas funções na engenharia de dados, do tipo A ao tipo B	49
Engenheiros de dados dentro de uma organização	50
Diferença entre engenheiros de dados voltados para o público interno e aqueles voltados para o público externo	50
Engenheiros de dados e outras funções técnicas	52
Engenheiros de dados e liderança corporativa	57
Resumo	61
Recursos adicionais	62
Capítulo 2 ■ O ciclo de vida da engenharia de dados	64
O que é o ciclo de vida da engenharia de dados?	64
O ciclo de vida dos dados versus o ciclo de vida da engenharia de dados	66
Geração: Sistemas de origem	66
Armazenamento	70
Ingestão	72
Transformação	77
Disponibilização de dados	78
Principais elementos subjacentes em todo o ciclo de vida da engenharia de dados	84
Segurança	85
Gerenciamento de dados	86

DataOps.....	98
Arquitetura de dados	103
Orquestração.....	104
Engenharia de software.....	106
Resumo	109
Recursos adicionais.....	110
Capítulo 3 ■ Desenvolvendo uma boa arquitetura de dados.....	111
O que é arquitetura de dados?	111
Definição de arquitetura corporativa	112
Definição de arquitetura de dados	115
A “boa” arquitetura de dados.....	117
Princípios de uma boa arquitetura de dados.....	118
Princípio 1: Escolha componentes comuns com sabedoria	119
Princípio 2: Planeje para falhas.....	120
Princípio 3: Projete para escalabilidade.....	121
Princípio 4: Arquitetura é o cerne da liderança	122
Princípio 5: Esteja sempre arquitetando	123
Princípio 6: Desenvolva sistemas fracamente acoplados.....	123
Princípio 7: Tome decisões que possam ser revertidas	126
Princípio 8: Priorize a segurança	126
Princípio 9: Adote o FinOps.....	128
Principais conceitos de arquitetura	130
Domínios e serviços.....	130
Sistemas distribuídos, escalabilidade e planejando para falhas	132
Acoplamento forte versus acoplamento fraco: Camadas, monolitos e microserviços	134
Acesso do usuário: locatário único versus multilocatários	140
Arquitetura orientada a eventos	140
Projetos brownfield versus projetos greenfield.....	141
Exemplos e tipos de arquitetura de dados	143
Data Warehouse	144
Data Lake.....	148
Convergência, a nova geração de data lakes e a plataforma de dados	149
Pilha de dados moderna.....	150
Arquitetura Lambda	151
Arquitetura Kappa.....	152
O modelo Dataflow e o processamento em lote e em fluxo (streaming) unificado	153
Arquitetura para IoT	154
Data Mesh.....	157
Outros exemplos de arquitetura de dados.....	158
Quem faz parte do desenvolvimento de uma arquitetura de dados?	159
Resumo	160
Recursos adicionais.....	160

Capítulo 4 ■ Escolhendo tecnologias ao longo do ciclo de vida da engenharia de dados 164

Tamanho e capacidades da equipe.....	165
Velocidade de entrada no mercado	166
Interoperabilidade.....	167
Otimização de custos e valor comercial	168
Custo total de propriedade.....	168
Custo de oportunidade total de propriedade.....	169
FinOps	170
Hoje versus o futuro: tecnologias imutáveis versus tecnologias transitórias.	171
Nossa recomendação	173
Local	174
On-premises.....	174
Nuvem	175
Nuvem híbrida	179
Multicloud	180
Computação descentralizada: o blockchain e a computação de borda	181
Nossa recomendação	182
Argumentos sobre a repatriação dos dados da nuvem	183
Construir versus comprar.....	186
Software de código aberto.....	187
Jardins murados proprietários.....	191
Nossa recomendação	194
Sistemas monolíticos versus sistemas modulares	194
Monolito	195
Modularidade.....	196
O padrão de monolito distribuído	197
Nossa recomendação	198
Arquitetura sem servidor (serverless) versus arquitetura com servidor	199
Arquitetura sem servidor	199
Contêineres	200
Avaliando arquitetura sem servidor versus arquitetura com servidor	202
Nossa recomendação	203
Otimização, desempenho e a guerra de benchmarks	204
Big Data...no contexto dos anos 1990s	205
Comparações de custos sem sentido	205
Otimização assimétrica.....	206
O “caveat emptor” — comprador, tenha cautela	206
Elementos subjacentes e seus impactos na escolha de tecnologias	206
Gerenciamento de dados.....	206
DataOps.....	207
Arquitetura de dados	208
Exemplo de orquestração: Airflow.....	208

Engenharia de software.....	209
Resumo	210
Recursos adicionais.....	210

Parte II = O ciclo de vida da engenharia de dados em detalhes.....211

Capítulo 5 = Geração de dados em sistemas de origem 212

Fontes de dados: como os dados são criados?	213
Sistemas de origem: principais conceitos	214
Arquivos e dados não estruturados.....	214
APIs	214
Bancos de dados de aplicação (sistemas OLTP)	215
Sistema de processamento analítico online	217
Captura de dados de alteração	218
Logs	218
Logs de banco de dados	220
CRUD	221
Padrão insert-only (apenas inserir)	221
Mensagens e fluxos	222
Tipos de registro de tempo.....	224
Detalhes práticos dos sistemas de origem.....	225
Banco de dados.....	225
APIs	236
Compartilhamento de dados.....	239
Fonte de dados de terceiros	239
Fila de mensagens e plataformas de streaming de eventos	240
Com quem você trabalhará	245
Elementos subjacentes e seus impactos em sistemas de origem	247
Segurança.....	247
Gerenciamento de dados.....	247
DataOps.....	248
Arquitetura de dados	250
Orquestração.....	251
Engenharia de software.....	251
Resumo	252
Recursos adicionais.....	253

Capítulo 6 = Armazenamento254

Componentes básicos do armazenamento de dados	256
Unidade de disco magnético	256
Unidades de estado sólido.....	259
Memória de acesso aleatório	260
Redes e CPU.....	261

Serialização	262
Compressão	263
Cache	264
Sistemas de armazenamento de dados	265
Armazenamento em uma única máquina versus armazenamento distribuído	265
Consistência eventual versus consistência forte	266
Armazenamento de arquivos	267
Armazenamento em blocos	270
Armazenamento de objetos	275
Sistemas de armazenamento baseado em memória e cache	282
O Hadoop Distributed File System	283
Armazenamento de streaming	284
Índices, particionamento e clustering	285
Abstrações de armazenamento na engenharia de dados	287
Data warehouse	288
Data lake	289
Data lakehouse	289
Plataformas de dados	290
Arquitetura de armazenamento de fluxo para lote	291
Principais conceitos e tendências em armazenamento	291
Catálogo de dados	292
Compartilhamento de dados	293
Esquema	294
Separação entre computação e armazenamento	294
Ciclo de vida do armazenamento de dados e retenção de dados	299
Armazenamento de único locatário versus armazenamento multilocatário	303
Com quem você trabalhará	305
Elementos subjacentes	305
Segurança	305
Gerenciamento de dados	306
DataOps	307
Arquitetura de dados	307
Orquestração	308
Engenharia de software	308
Resumo	309
Recursos adicionais	309
Capítulo 7 ■ Ingestão	310
O que é ingestão de dados?	310
Principais considerações de engenharia para a etapa de ingestão	312
Dados limitados versus ilimitados	313
Frequência	314
Ingestão síncrona versus ingestão assíncrona	315

Serialização e desserialização	317
Taxa de transferência e escalabilidade	317
Confiabilidade e durabilidade	318
Carga útil	319
Diferença entre os padrões push, pull e poll	322
Considerações sobre ingestão em lote	323
Snapshot ou extração diferencial	325
Exportação e ingestão baseada em arquivos	325
ETL versus ELT	325
Inserções, atualizações e tamanho de lote	326
Migração de dados	327
Considerações sobre ingestão de streaming e mensagens	327
Evolução do esquema	327
Dados que chegam atrasados	328
Ordenação e entrega múltipla	329
Replay	329
Tempo de vida	329
Tamanho das mensagens	330
Tratamento de erros e filas de mensagens mortas	330
Operações de pull e push realizadas pelo consumidor	331
Localização	331
Formas de ingestão de dados	331
Conexão direta com o banco de dados	331
Captura de dados de alteração	333
APIs	336
Filas de mensagens e plataformas de streaming de eventos	337
Conectores de dados gerenciados	339
Movimentação de dados com armazenamento de objetos	339
EDI	340
Bancos de dados e exportação de arquivos	340
Questões práticas com formatos de arquivo comuns	341
Shell	342
SSH	342
SFTP e SCP	342
Webhooks	343
Interface web	344
Web scraping	344
Appliances de transferência para migração de dados	345
Compartilhamento de dados	346
Com quem você trabalhará	346
Stakeholders do upstream	347
Stakeholders do downstream	347

Elementos subjacentes.....	348
Segurança.....	348
Gerenciamento de dados.....	349
DataOps.....	351
Orquestração.....	354
Engenharia de software.....	354
Resumo	355
Recursos adicionais.....	355

Capítulo 8 ■ Consultas, modelagem e transformação 356

Consultas	357
O que é uma consulta?.....	358
A vida útil de uma consulta.....	360
O otimizador de consultas	360
Aprimorando o desempenho da consulta.....	361
Consultas em dados de streaming.....	368
Modelagem de dados	376
O que é um modelo de dados?.....	377
Modelos conceitual, lógico e físico de dados	378
Normalização.....	379
Técnicas para modelagem de dados analíticos em lote	384
Modelando dados de streaming	399
Transformações	400
Transformações em lote	401
Visualizações materializadas, federação e virtualização de consultas.....	419
Transformações e processamento de streaming.....	422
Com quem você trabalhará	426
Stakeholders do upstream	426
Stakeholders do downstream	427
Elementos subjacentes.....	427
Segurança.....	427
Gerenciamento de dados.....	428
DataOps.....	429
Arquitetura de dados	430
Orquestração.....	430
Engenharia de software.....	431
Resumo	432
Recursos adicionais.....	432

Capítulo 9 ■ Disponibilizando dados para análise, machine learning e ETL reverso..... 435

Considerações gerais para a disponibilização de dados.....	436
Confiança.....	436
Qual é o caso de uso e quem é o usuário?	438

Produtos de dados	439
Autoatendimento ou não?	440
Definições e lógica de dados.....	441
Data Mesh.....	443
Análise	443
Análise de negócios	444
Análise operacional	446
Análise incorporada.....	448
Machine learning.....	450
O que um engenheiro de dados deve saber sobre ML	450
Maneiras de disponibilizar dados para análise e ML	452
Troca de arquivos	453
Bancos de dados	454
Sistemas de streaming.....	456
Federação de consultas	456
Compartilhamento de dados.....	457
Camadas semânticas e camadas de métricas.....	458
Disponibilizando dados em notebooks	459
ETL reverso	461
Com quem você trabalhará	463
Elementos subjacentes.....	464
Segurança.....	464
Gerenciamento de dados.....	466
DataOps.....	467
Arquitetura de dados	467
Orquestração.....	468
Engenharia de software.....	469
Resumo	470
Recursos adicionais.....	470

Parte III = Segurança, privacidade e o futuro da engenharia de dados 472

Capítulo 10 = Segurança e privacidade 473

Pessoas	474
O poder do pensamento negativo	474
Mantenha-se sempre paranoico com a segurança.....	475
Processos.....	475
Falsa sensação de segurança versus hábito de segurança	476
Segurança ativa.....	476
O princípio do privilégio mínimo	477
Responsabilidade compartilhada na nuvem	477
Sempre faça o backup dos seus dados.....	477
Um exemplo de política de segurança.....	478

Tecnologia	480
Aplique patches e atualize os sistemas	480
Criptografia.....	480
Registro de logs, monitoramento e emissão de alertas	481
Acesso à rede.....	483
Segurança para engenharia de dados de baixo nível	484
Resumo	485
Recursos adicionais.....	485

Capítulo 11 ■ O futuro da engenharia de dados486

O ciclo de vida da engenharia de dados veio para ficar	487
A diminuição da complexidade e a ascensão de ferramentas de dados fáceis de usar	487
O sistema operacional de dados em escala de nuvem e a melhoria na interoperabilidade	489
Engenharia de dados “empresarial”	491
Funções e responsabilidades se transformarão... ..	492
Indo além da pilha de dados moderna, rumo à pilha de dados em tempo real	493
A pilha de dados em tempo real	494
Pipelines de streaming e bancos de dados analíticos em tempo real	495
A fusão de dados com aplicações.....	496
A estreita relação entre aplicações e ML.....	497
Dados de matéria escura e o surgimento de... planilhas ?!	497
Conclusão.....	498

Apêndice A ■ Detalhes técnicos de serialização e compressão500

Formatos de serialização	500
Serialização baseada em linhas.....	500
Serialização colunar.....	502
Serialização híbrida	505
Mecanismos de armazenamento de banco de dados.....	506
Compressão: gzip, bzip2, Snappy e outros.	506

Apêndice B ■ Rede em nuvem.....508

Topologia de rede em nuvem	508
Custos de saída de dados	508
Zonas de disponibilidade	509
Regiões	510
Redundância de multirregião e rede específicos do GCP	510
Conexões de rede diretas para as nuvens	511
CDNs.....	511
O futuro das taxas de saída de dados.....	512

Índice remissivo515