Boas Práticas Para Tratamento de Dados (Analytics)

# Boas práticas de tratamento de Dados para Analytics - Papéis e Arquitetura básica de armazenamento

Quando começamos o estudo de Analytics Engineering (ou Engenharia de Analytics), é importante entendermos dois temas que balizam grande parte do estudo:

* Quem são os especialistas de dados?
* Como armazenamos Big Data?

## Quem são os especialistas de Dados?

Com o crescente volume de dados disponíveis, a demanda por especialistas em dados tem aumentado significativamente. Neste tópico, apresentaremos os principais papéis dos especialistas de dados bem como alguns conceitos utilizados ao longo da confecção dos pipeline de tratamento de dados. Lembrando que esta é uma relação proposta com base no que geralmente é visto no mercado, entretanto, existem customizações e variações destes termos de acordo com o projeto e/ou instituição.

1. Engenheiro de dados: a engenharia de dados é o processo de coletar, limpar, transformar e estruturar dados para análise. O especialista em engenharia de dados é responsável por criar pipelines de dados que conectem as fontes de dados aos sistemas de análise. Isso inclui trabalhar com bancos de dados, big data, sistemas de armazenamento em nuvem e outras tecnologias relacionadas.
2. Analytics engineer: o analytics engineer é um especialista em dados que projeta, implementa e mantém sistemas de análise de dados. Isso pode incluir a criação de painéis de visualização de dados, a construção de modelos de análise de dados e a realização de testes de desempenho em sistemas de análise de dados. O analytics engineer também é responsável por garantir a qualidade e precisão dos dados utilizados nos sistemas de análise.
3. Cientista de dados: o cientista de dados é um especialista em dados que usa métodos estatísticos, matemáticos e de machine learning para analisar e interpretar dados. O cientista de dados trabalha com os dados para encontrar padrões, identificar tendências e desenvolver modelos de previsão. Eles também são responsáveis por comunicar insights e resultados de análise de dados para a equipe de negócios.
4. Analista de dados: o analista de dados é responsável por coletar, analisar e interpretar dados para apoiar a tomada de decisões de negócios. Eles trabalham com equipes de negócios para identificar as necessidades de dados, coletar dados e gerar relatórios e visualizações que comuniquem os resultados da análise de dados.
5. Engenheiro de machine learning: o engenheiro de machine learning é responsável por desenvolver sistemas de machine learning que permitam que as organizações tomem decisões baseadas em dados. O engenheiro de machine learning trabalha com algoritmos de machine learning para criar modelos que possam aprender com dados e fazer previsões precisas. Eles também são responsáveis por integrar esses modelos em sistemas de análise de dados e garantir a escalabilidade e desempenho desses sistemas em ambientes produtivos.
6. Arquiteto de dados: o arquiteto de dados é responsável pelo projeto e implementação de soluções de armazenamento de dados em larga escala e pelo gerenciamento de todo o ciclo de vida de dados. Isso inclui a concepção da arquitetura de dados, o desenho de modelos de dados e esquemas de banco de dados, a definição de padrões e políticas de gerenciamento de dados, e a seleção de tecnologias de armazenamento e processamento de dados. O arquiteto de dados deve garantir que a arquitetura de dados seja escalável, segura, eficiente e possa lidar com grandes volumes de dados. Além disso, o arquiteto de dados também trabalha com outros especialistas de dados para garantir que os requisitos de dados sejam atendidos, que os dados sejam consistentes e precisos e que as soluções de dados sejam compatíveis com as necessidades de negócios da empresa.
7. Data Product Manager: o gerente de produtos de dados, também conhecido como data product manager, é responsável pela concepção, desenvolvimento e lançamento de produtos de dados. O data product manager trabalha em estreita colaboração com equipes de engenharia, cientistas de dados, analistas de dados, engenheiros de machine learning e outras partes interessadas para definir os requisitos de produtos de dados, estabelecer prioridades, coordenar esforços de desenvolvimento e garantir que os produtos de dados sejam lançados a tempo e dentro do orçamento. O data product manager deve estar familiarizado com as tecnologias de dados e possuir uma visão clara dos objetivos e metas do negócio para garantir que os produtos de dados atendam às necessidades dos usuários e do negócio. Além disso, o data product manager deve ter habilidades de liderança, comunicação e gerenciamento de projetos para coordenar os esforços de desenvolvimento de produtos de dados.

Os especialistas em dados são fundamentais para a gestão e análise de dados em todas as organizações. Eles desempenham papéis importantes em todas as etapas do ciclo de vida dos dados, desde a modelagem de dados até a análise e interpretação de resultados. Compreender esses papéis pode ajudar as organizações a construir equipes de dados eficazes e alcançar insights valiosos a partir de seus dados.

## Arquitetura de soluções em dados

Agora, vamos discutir um pouco sobre alguns conceitos da arquitetura de soluções de dados.

1. Modelagem de dados: A modelagem de dados é a arte e ciência de criar modelos que representem a estrutura, relacionamentos e regras de negócios subjacentes aos dados. O especialista em modelagem de dados trabalha com as equipes de negócios e de desenvolvimento de software para definir os requisitos de dados, identificar as fontes de dados e desenvolver modelos de dados que suportem as necessidades da organização.
2. Arquitetura de soluções de especialização de dados: A arquitetura de soluções de especialização de dados refere-se ao design de sistemas que suportam o processamento de dados em diferentes estágios de seu ciclo de vida. Isso inclui a coleta de dados brutos, a harmonização e estruturação de dados e a especialização de dados para fins específicos, como análise e relatórios. A arquitetura de soluções de especialização de dados pode incluir várias tecnologias, como bancos de dados, data warehousing, big data e nuvem.
3. Armazenamento de dados em camadas raw, trusted e specified: o armazenamento de dados em camadas é uma técnica de gerenciamento de dados que permite que as organizações armazenem e gerenciem dados em diferentes estágios do ciclo de vida dos dados. As camadas raw, trusted e specified são comumente usadas para categorizar dados em diferentes estágios. Também são usados outros nomes representando propósitos equivalentes, como raw, struct e refined ou bronze, silver e gold, por exemplo.

Dados brutos (raw): são dados coletados de fontes diversas, como sensores, aplicativos e sistemas. Esses dados geralmente não estão harmonizados ou estruturados e podem ser difíceis de analisar diretamente. Os benefícios da camada raw incluem a capacidade de coletar e armazenar grandes volumes de dados com baixo custo. O exemplo a seguir ilustra como podemos ler um arquivo CSV e representar a camada raw em Python:  
import pandas as pd

# Lendo um arquivo CSV

df\_raw = pd.read\_csv('dados\_brutos.csv')

# Exibindo as primeiras linhas do dataframe

print(df\_raw.head())

* + Note que no exemplo acima, não há tratamento algum nos dados. Eles apenas são carregados dentro do fluxo da pipeline a partir da sua fonte de origem.

Dados harmonizados e/ou estruturados (trusted): são dados que foram processados e transformados em um formato estruturado que permite a análise direta. Esses dados geralmente podem ser armazenados em um data lake, em um data warehouse ou banco de dados relacional (de acordo com volume, latência, entre outros). Os benefícios da camada trusted incluem a capacidade de analisar dados diretamente e com maior facilidade com base em dados que já foram estruturados e representam uma “fonte da verdade”. Continuando o exemplo anterior, podemos criar a camada trusted a partir da camada raw:  
# Carregando os dados brutos

df\_raw = pd.read\_csv('dados\_brutos.csv')

# Limpando os dados (removendo duplicatas, preenchendo valores ausentes, etc.)

df\_trusted = df\_raw.drop\_duplicates().fillna(0)

# Exibindo as primeiras linhas do dataframe

print(df\_trusted.head())

* + Agora, nossa única ação é aplicar algum tipo de tratamento para garantir a qualidade de consumo dos dados que estamos manipulando, conforme técnicas que são abordadas nos tópicos ao longo deste módulo.

Dados especializados (specified ou spec): são dados que foram otimizados para fins específicos, como análise de negócios ou relatórios. Isso pode incluir tabelas e visões especializadas que foram otimizadas para consultas específicas. Os benefícios da camada spec incluem consultas mais rápidas e menor custo de processamento. Por fim, podemos criar a camada spec a partir da camada trusted:  
# Carregando os dados confiáveis

df\_trusted = pd.read\_csv('dados\_confiaveis.csv')

# Criando uma nova coluna com o total de vendas por produto

df\_specified = df\_trusted.groupby('produto')['valor\_venda'].sum().reset\_index()

# Exibindo as primeiras linhas do dataframe

print(df\_specified.head())

* + Nesse exemplo, a camada specified é criada a partir da camada trusted ao agrupar os dados por produto e somar o valor de vendas para cada produto. Veja: o que fazemos nesta camada, nada mais é do que a geração de diferentes visões dos dados existentes na camada trusted.

1. Aplicação prática de tabelas e visões especializadas: tabelas e visões especializadas são exemplos de dados especializados. Essas tabelas e visões são projetadas para atender a propósitos específicos, como análise de negócios ou relatórios executivos. A criação de tabelas e visões especializadas pode melhorar o desempenho do sistema, reduzir o custo de processamento e melhorar a precisão dos resultados. Por exemplo, criar uma tabela resumida para um relatório pode reduzir significativamente o tempo necessário para executar consultas. E, normalmente, são armazenadas na camada spec.
2. Granularidade, sumarização, agrupamento e junção de dados: a granularidade de dados refere-se ao nível de detalhe dos dados armazenados. A sumarização é o processo de agregação de dados em um nível mais alto de granularidade, como somar as vendas por mês em vez de por dia. A sumarização pode ser útil para melhorar o desempenho do sistema e reduzir o custo de processamento, pois há menos dados a serem processados. O agrupamento é o processo de agrupar dados com base em uma ou mais colunas, como agrupar as vendas por região ou por categoria de produto. A junção é o processo de combinar dados de duas ou mais tabelas com base em uma coluna em comum, como combinar dados de vendas com dados de clientes com base no ID do cliente.

Essas técnicas podem ser usadas para atender a propósitos específicos, como relatórios e dashboards executivos. Por exemplo, um relatório executivo pode incluir informações agregadas, como vendas totais por região ou por categoria de produto, em vez de detalhes de vendas individuais. Isso pode ajudar os executivos a obter uma visão geral rápida do desempenho do negócio.

## Considerações finais

A arquitetura de soluções de especialização de dados ainda pode ajudar a melhorar a qualidade dos dados, permitindo que os dados sejam transformados e validados antes de serem carregados nas camadas subsequentes. Isso pode incluir a validação de dados, a eliminação de dados duplicados, a correção de erros de formatação e a normalização de dados. Esses processos podem melhorar a qualidade dos dados, tornando-os mais precisos e consistentes, o que, por sua vez, pode levar a decisões de negócios mais informadas e precisas.

Em termos de implementação, a arquitetura de soluções de dados pode ser implementada em uma variedade de plataformas, incluindo plataformas de nuvem, on-premises ou híbridas. Existem várias soluções de software disponíveis que podem ajudar a implementar a arquitetura de soluções de especialização de dados, incluindo soluções de data warehousing e soluções de big data.

Por fim, é importante destacar que a arquitetura de soluções de especialização de dados não é uma solução única para todos os problemas de dados e a arquitetura de soluções deve ser projetada com a escalabilidade em mente. À medida que a quantidade de dados aumenta, a arquitetura deve ser capaz de lidar com o aumento da carga de dados sem comprometer o desempenho. Cada organização tem suas próprias necessidades e requisitos de dados, e a arquitetura de soluções de especialização de dados deve ser adaptada para atender a essas necessidades.

# Referências

* Livro virtual Automate the Boring Stuff with Python: <https://automatetheboringstuff.com/>
* Playlist Python no canal da Ada no Youtube: <https://www.youtube.com/watch?v=cDqMbI02hRs&list=PLcmVV8telDGzH6wFY_9h_KZ3Sn7EfjT1D>
* Livro: REIS, Joe & HOUSLEY, Matt. Fundamentals of Data Engineering, O'Reilly, 2022.
* Livro: KLEPPMANN, Martin. Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, 2017
* Livro: DENSMORE, James. Data Pipelines Pocket Reference: Moving and Processing Data for Analytics, 2021
* Artigo \*\*\*\*Top 9 Job Roles in the World of Data Science for 2023: [https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/#:~:text=Developing and maintaining databases,data collection and quality processes](https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/#:~:text=Developing%20and%20maintaining%20databases,data%20collection%20and%20quality%20processes)
* Artigo The Types of Data Science Roles Explained: <https://365datascience.com/career-advice/types-of-data-science-roles-explained/>
* Artigo What is derived data, and how can you use it for your business?: <https://www.optimizely.com/insights/blog/what-is-derived-data-and-how-can-you-use-it-for-your-business/>
* Documentação Pandas: <https://pandas.pydata.org/docs/user_guide/index.html>
* Artigo raw data (source data or atomic data): <https://www.techtarget.com/searchdatamanagement/definition/raw-data>
* Artigo Different Types Of "Data Engineering" Teams: <https://seattledataguy.substack.com/p/different-types-of-data-engineering>
* Data-driven: Creating a Data Culture. Este livro de Hilary Mason e DJ Patil é um guia útil para ajudar a criar uma cultura de dados em uma organização, desde a coleta e gerenciamento de dados até a análise e visualização de dados. Ele também aborda como garantir que os dados sejam usados de maneira ética e responsável.
* Data-Driven: Creating a Data Culture, O'Reilly Media, Inc., 2020.
* Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. Este livro de Foster Provost e Tom Fawcett é um guia abrangente sobre como aplicar técnicas de ciência de dados para resolver problemas de negócios. Ele aborda desde a coleta e preparação de dados até a análise e modelagem, além de enfatizar a importância da comunicação e colaboração com as partes interessadas do negócio.
* Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, O'Reilly Media, Inc., 2013.
* Dataiku Blog. O blog da Dataiku fornece insights e conselhos sobre como aplicar ciência de dados em negócios. Ele inclui estudos de caso, dicas e truques, bem como discussões sobre as últimas tendências e ferramentas de ciência de dados: <https://blog.dataiku.com/>
* Harvard Business Review. A Harvard Business Review é uma publicação líder em negócios e gestão. Eles têm uma série de artigos e casos de estudo que abordam como usar dados de maneira eficaz para impulsionar os negócios. Eles também têm uma série de cursos online sobre análise de dados e tomada de decisão baseada em dados: <https://hbr.org/topic/data-and-analytics>