

Engenharia de dados com DBT

...

Bem-Vindos

Referências

Cursos oficiais

...

<https://courses.getdbt.com/collections>

Documentação oficial

...

<https://docs.getdbt.com/>

Blog Oficial

...

<https://www.getdbt.com/blog/>

Certificação Oficial

...

<https://www.getdbt.com/blog/dbt-certification-program/>

Engenharia de dados com DBT

<https://www.amazon.com/Data-Engineering-dbt-cloud-based-dependable-ebook/dp/B0C4LL19G7>

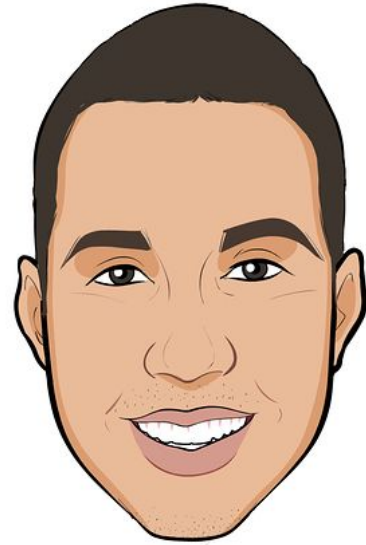


O que você está comprando?

- Tempo
- Didática
- Organização
- Material em português
- Demos
- Canal de dúvidas
- Suporte
- Disciplina
- Bônus
- Experiência do instrutor
- Base e contexto
- Aprendizado semi customizado (Pesquisa)

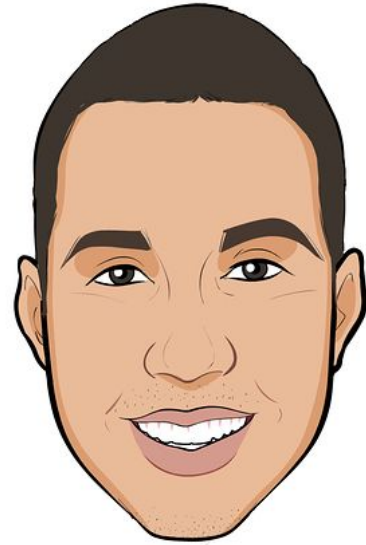
Instrutor

Matheus Willian



Contatos

[linkedin.com/in/matheuswillian](https://www.linkedin.com/in/matheuswillian)
matheuswilliandf@gmail.com



Roadmap

Roadmap

- 5 dias de treinamento das 19:00 às 22:00.
- Dia 01 - Base
- Dia 02 - Fundamentos
- Dia 03 - Cloud
- Dia 04 - Intermediário
- Dia 05 - Avançado

Dia 01 - Base

- Business Intelligence
 - ETL
 - Traditional DW
 - Data Viz
 - Q&A: Business intelligence
- DW
 - Snowflake, PostgreSQL, Big Query
 - Modelagem dimensional
 - Star Schema, Snowflake, Data Vault
 - Slowly Changing Dimensions (SCD)
 - Demo: PostgreSQL and Dbeaver Instalation
- DBT Core
 - Contexto
 - Timeline
 - Concepts
 - Project Structure
 - Demo: DBT Core Installation
- Just Enough Python for BI
 - Virtual env
 - Pandas
 - SQL Alchemy
 - Data Viz
 - Demo: BI with Python, DBT Core, and PostgreSQL

Dia 02 - Fundamentos

- Big Data and Cloud
 - Big Data and Cloud
 - Cloud Data Warehouse (CDW)
 - Data lake vs Data Warehouse vs lakehouse
 - External tables
- Modern Data Stack
 - Modern Data Stack
 - Analytics Engineer
 - BI vs MDS
 - ELT vs ETL
 - CDW vs TDW
 - Q&A: BI vs MDS
- Git Overview
 - DevOps
 - Branchs (development, main, etc.)
 - Commands
 - Gitflow
 - Pull Request
 - Demo: Git installation and workflow
 - Demo: Github setup
- DBT Cloud
 - DBT Cloud
 - DBT Core vs Cloud
 - Demo: Airbyte Account Creation
 - Demo: Big Query Account Creation
 - Demo: DBT Cloud Account Creation
 - Demo: MDS with Airbyte, DBT Cloud and Big Query

Dia 03 - Cloud

- Engineering Best Practices
 - DevOps/DataOps
 - Tests and Documentation
 - Data Quality
 - Q&A: Best Practices
- SQL Advanced
 - Joins and unions
 - Aggregations and functions
 - CTEs and subselects
 - Window Functions
 - Query Execution Plan
 - Explain and Optimization
 - Demo: Reviewing SQL Queries
 - Demo: Query optimization
- DBT Features
 - Sources
 - Tags
 - Tests and Advanced Tests
 - Documentation
 - CI/CD
 - Logging and Alerting
 - Demo: Snowflake Account Creation
 - Demo: DBT best practices

Dia 04 - Intermediário

- Analyses
- Seeds
- Advanced Materialization (Incremental & Ephemeral)
- Snapshots
- Singular and Custom Tests
- Exposures

Dia 05 - Avançado

- Refactoring SQL
- Jinja
- Macros
- Packages
- Hooks
- Certification Tips

Cronograma

Cronograma

- 19:00-19:10 (10 min) espera/review
- 19:10-19:40 (30 min) conteúdo I
- 19:40-20:20 (40 min) demo I
- 20:20-20:40 (20 min) pausa
- 20:40-21:10 (30 min) conteúdo II
- 21:10-21:50 (40 min) demo II
- 21:50-22:00 (10 min) encerramento/duvidas

Dicas

Dicas - Como fazer esse curso?

- Prestar mais atenção do que tentar fazer junto
- Vocês tem acesso ao curso durante 1 ano
- Como fazer um curso online
- Qual a diferença para um fast-track

Dicas - O que está incluso?

- Slides
- Repositório
- Bônus

Bônus

- Série - Os fundamentos
- Fast Track - SQL para times de dados
- Workshop - Construindo, na prática, uma Stack de Dados Moderna com airbyte, airflow, dbt & snowflake

Dúvidas

Pesquisa

Dia 01

Contexto

Analytics

Analytics - Contexto

Info

Data is generated all the time, both inside and outside the company.

Data Producer



Info

Data professionals need to access multiple data sources and combine them to give a meaning for the data.

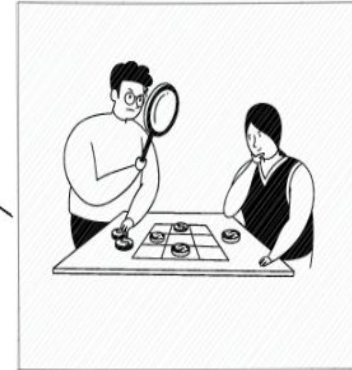
Analytics



Info

Based on meaningful data, the decision level can now use facts to decide which course of action we can take.

Decision Makers



Analytics - Pirâmide das necessidades de AI

THE DATA SCIENCE **HIERARCHY OF NEEDS**

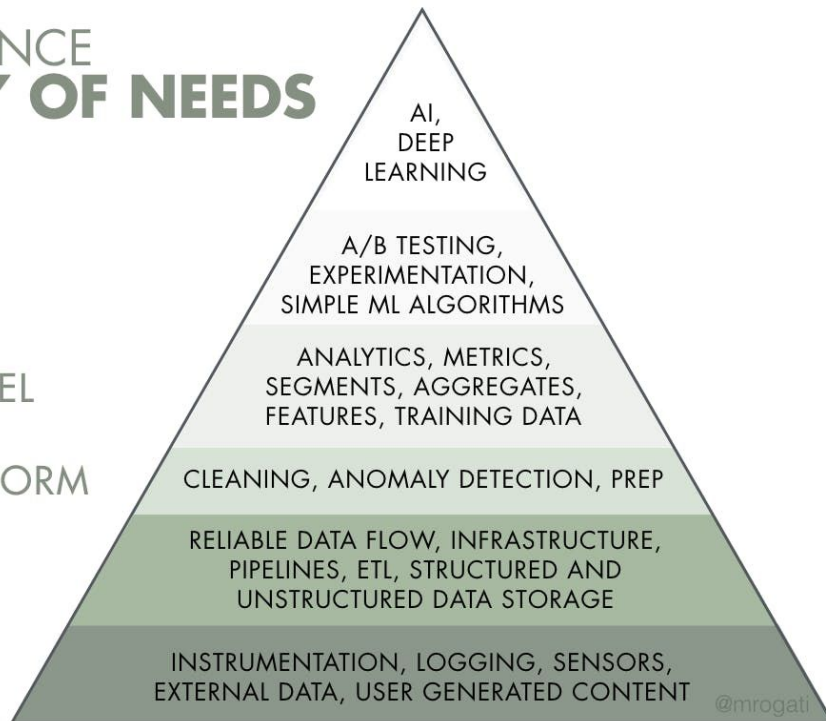
LEARN/OPTIMIZE

AGGREGATE/LABEL

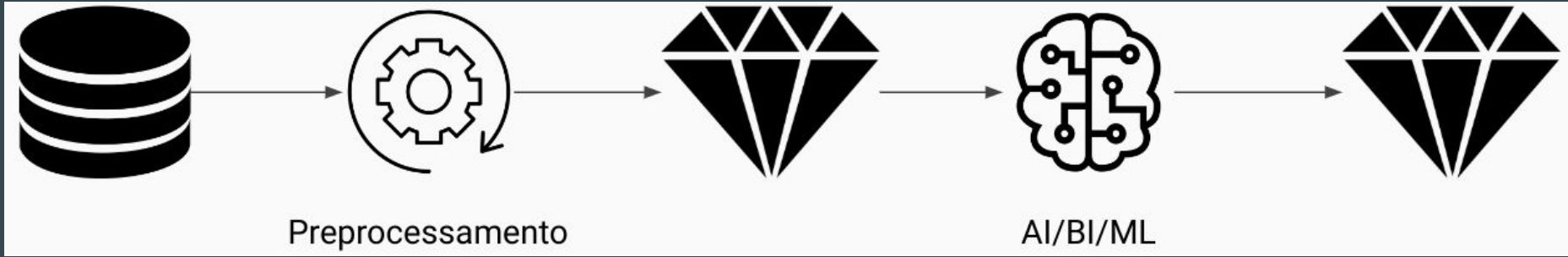
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

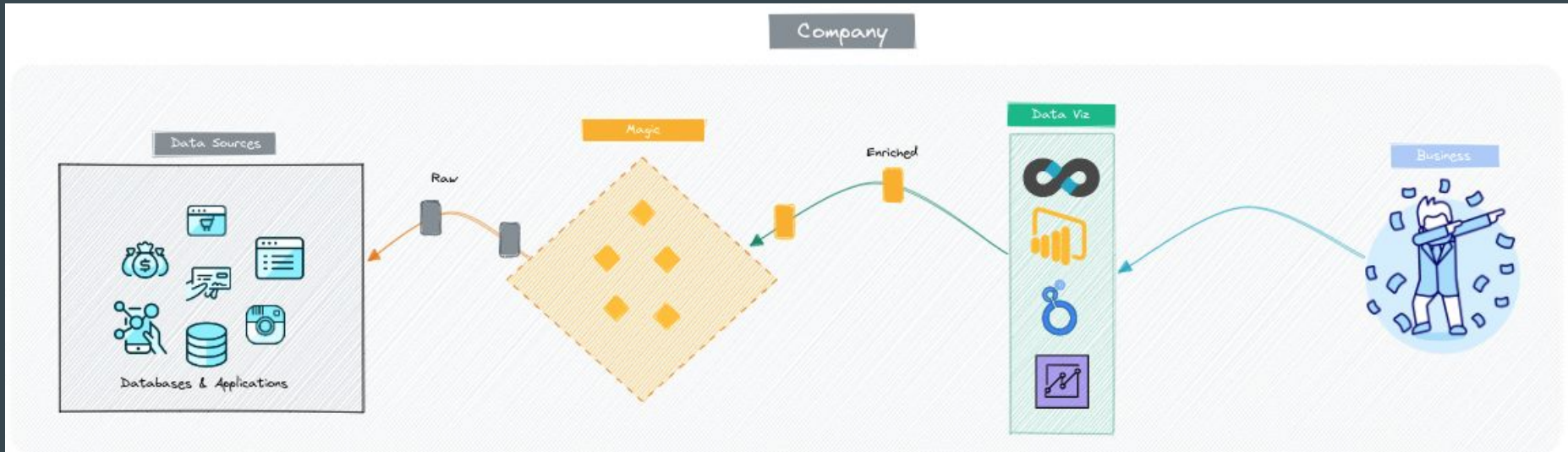


Analytics - Pirâmide das necessidades de AI



BI

BI - Contexto



BI - Contexto

- ETL
- DW
- Data Viz

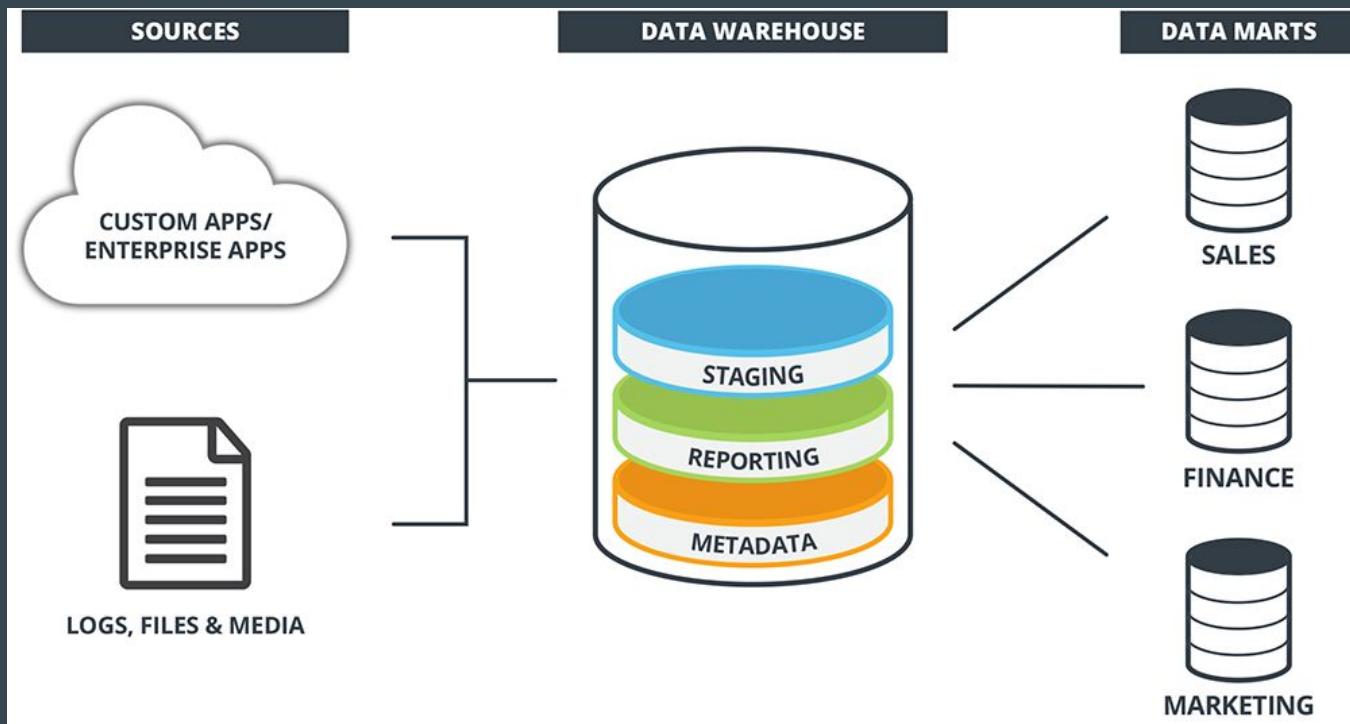
ETL



BI - ETL

- Extração, transformação e carga;
- Processo de alimentação de uma fonte de dados;
- Diretamente associado à alimentação de um DW;
- Ferramentas;

TDW



BI - Data Warehouse

- Unificação de fontes de dados;
- Não substitui banco operacional da empresa;
- Modelo de dados: cubo multidimensional;

Data Visualization

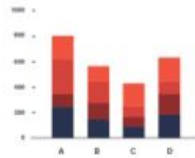
KPIs



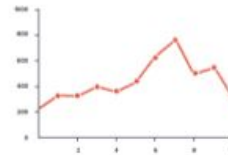
Tables

| | A | B | C |
|---|------|-----|----|
| X | \$40 | 240 | 48 |
| Y | \$50 | 200 | 59 |
| Z | \$60 | 310 | 79 |

Bar charts



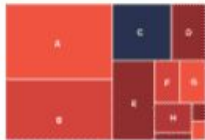
Line charts



Donut charts



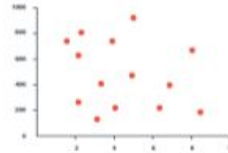
Tree Maps



Bullet Charts



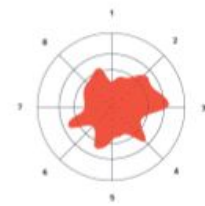
Scatter plots



Geo Maps



Radial charts



BI - Visualização de dados

- Visualização dos dados finais;
- Impacto na tomada de decisão;
- Clareza e simplicidade;

Dúvidas

Data Warehouse

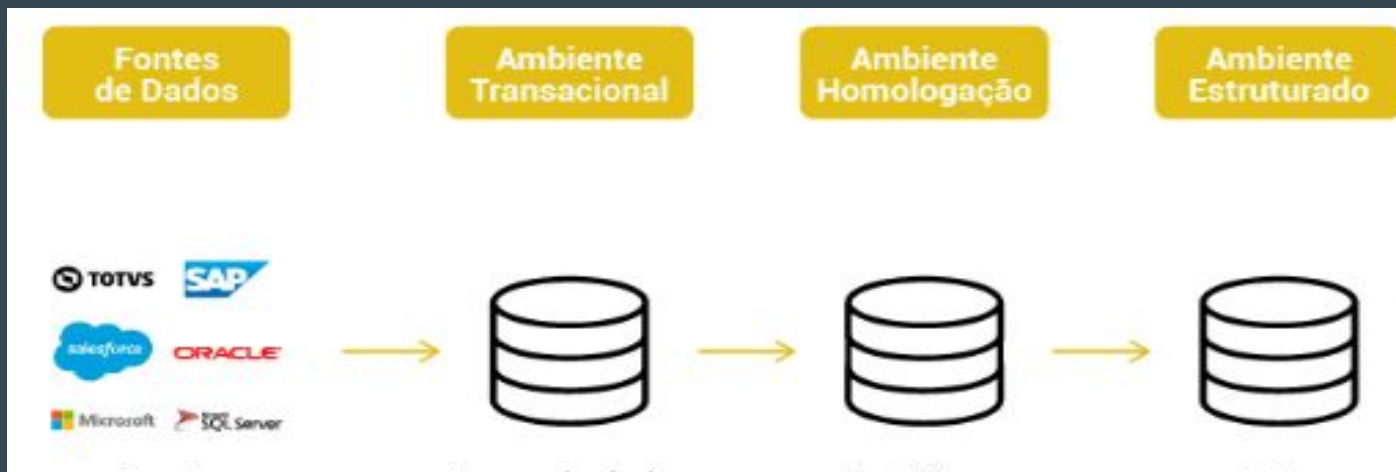
DW - Problemática Banco de dados Relacional

- Dificuldade de acesso aos dados
- Fragmentação da infraestrutura
- Baixo desempenho
- Dados despadronizados

DW - Definição

- Unir e harmonizar distintas e diversas fontes de dados desejadas para um contexto BI em suporte para tomada de decisão

DW - Definição



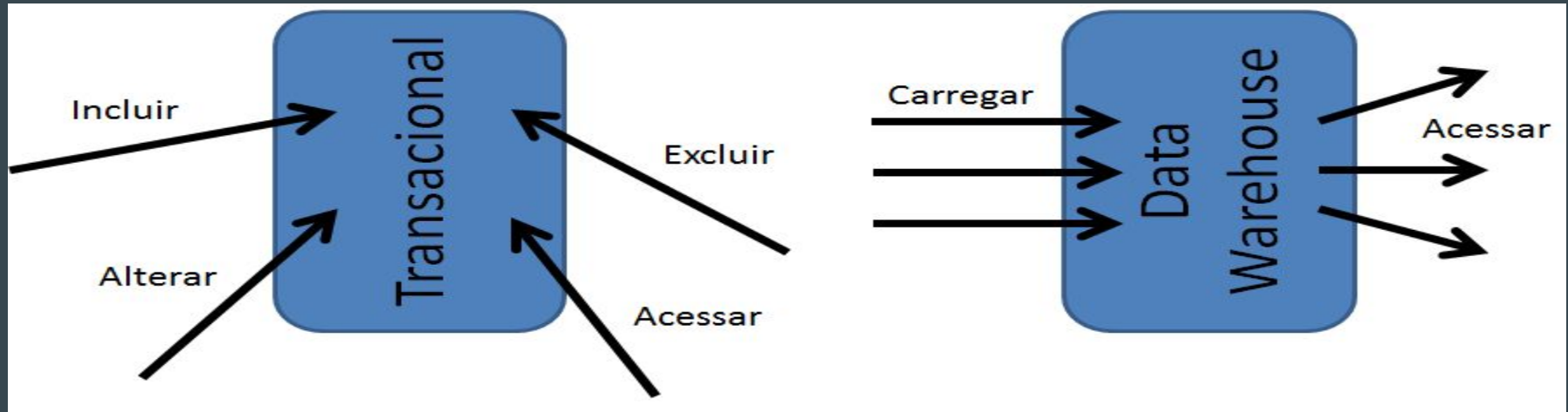
DW - ETL

Para isso é preciso um processo de inserção e transformação dos dados.

1. Os dados são extraídos da fonte
2. Os dados são transformados/manipulados
3. Os dados são inseridos no Data Warehouse

Os dados podem ser inseridos primeiramente no DW (Data warehouse) e transformados em outro momento (ELT).

DW - DW vs Database

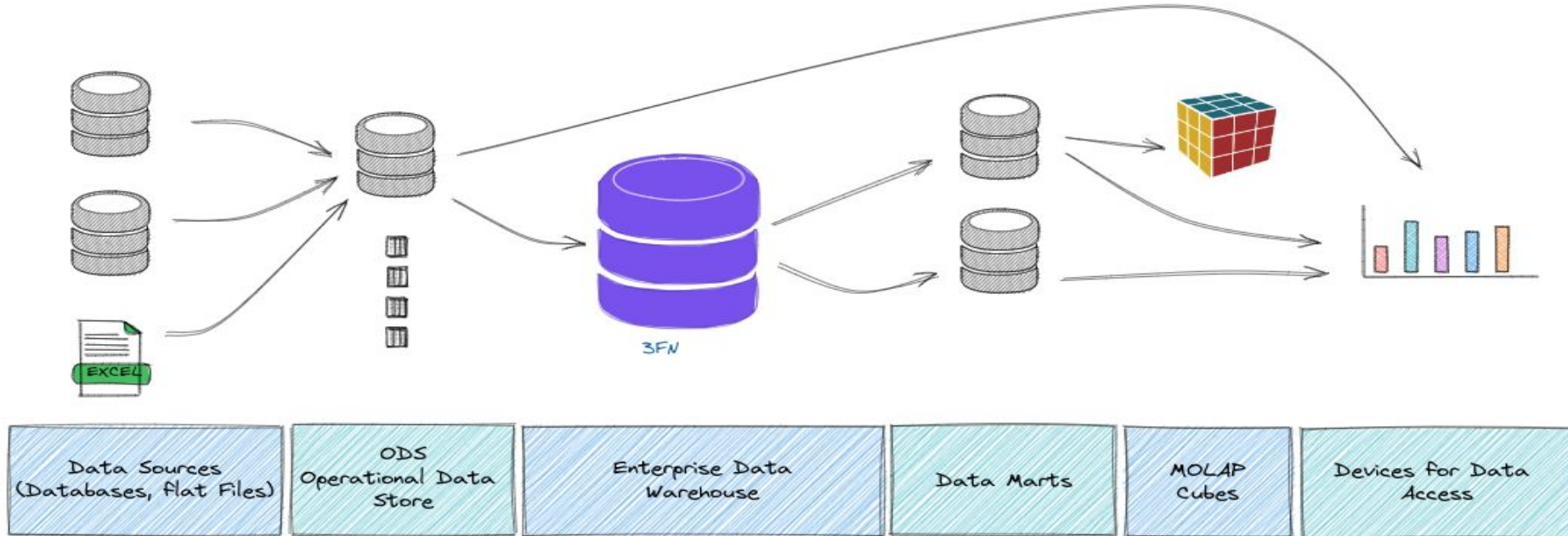


DW - Bill Inmon

Info

Bill Inmon
1992
Book: Building the Data Warehouse
Top Down, 3Fn

To move canvas, hold mouse wheel or spacebar while dragging, or use the hand tool



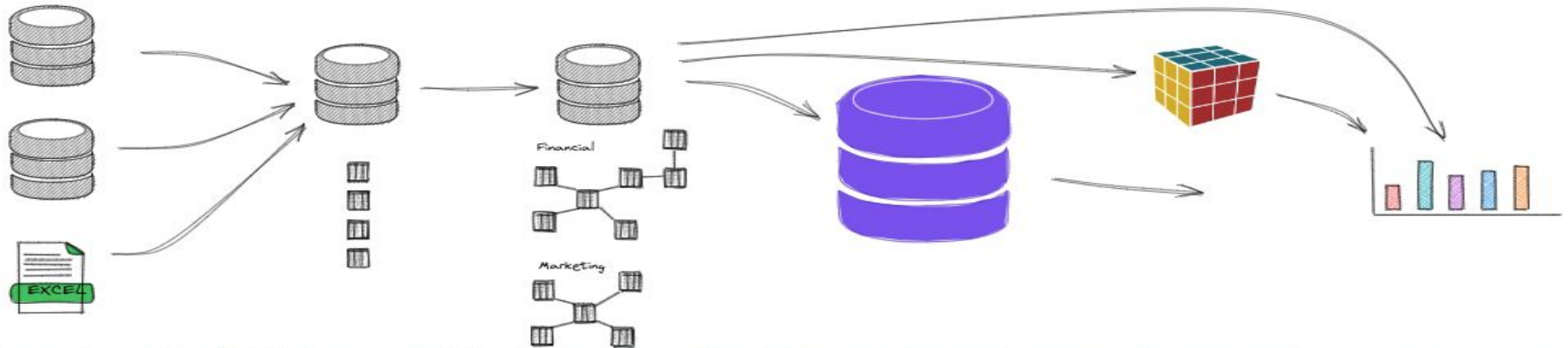
DW - Inmon

- Visão da alta gestão
- Top-Down
- Comumente formado o data warehouse e depois dividido de forma departamental (data mart)

BI Tradicional - Ralph Kimball

Info

Ralph Kimball
1996
Book: The Data Warehouse Toolkit
Bottom Up, Star Schema and Snowflake



Data Sources
(Databases, flat Files)

Staging Area

Data Marts

Enterprise Data
Warehouse

MOLAP
Cubes

Devices for Data
Access

DW - Kimbal

- Visão departamental
- Bottom-up
- Alta probabilidade de ilhagem dos dados
- Sem integração entre os data mart

DW - Kimball x Inmon

- DW dissolvendo em data marts (Inmon);
- Data marts compondo DW (Kimball);
- Vantagens e Desvantagens;
- O que é mais usado?;

BI - Data Team



Analista de
ETL

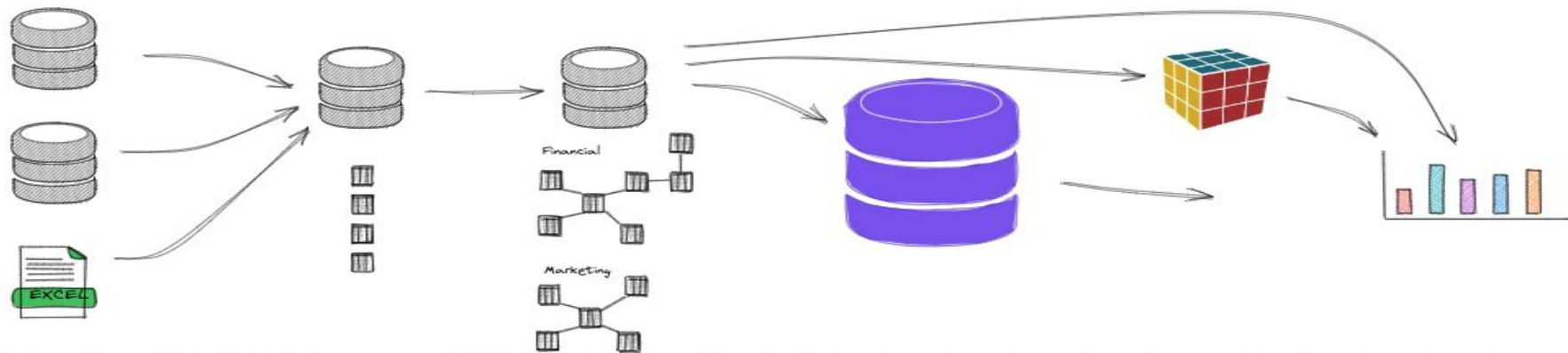


DBA DW



Analista de BI

BI - Data Team



Data Sources
(Databases, Flat Files)

Stagin Area

Data Marts

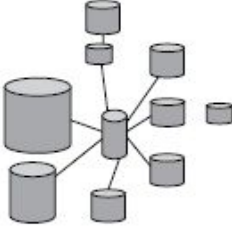
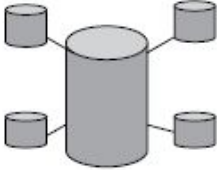

Enterprise Data
Warehouse

MOLAP
Cubes

Devices for Data
Access

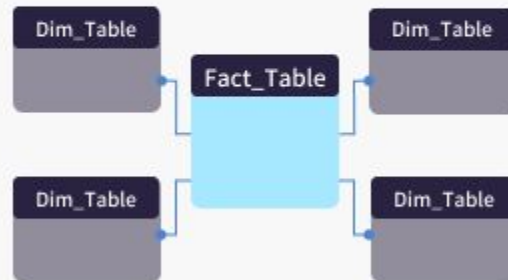
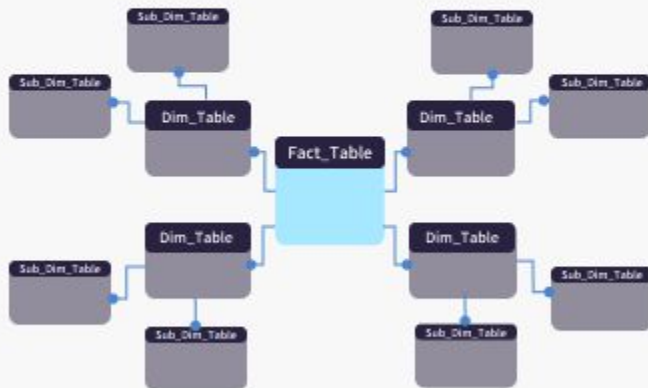
DW - Modelagem

Modelagem - Snowflake, Star Schema, Single Table

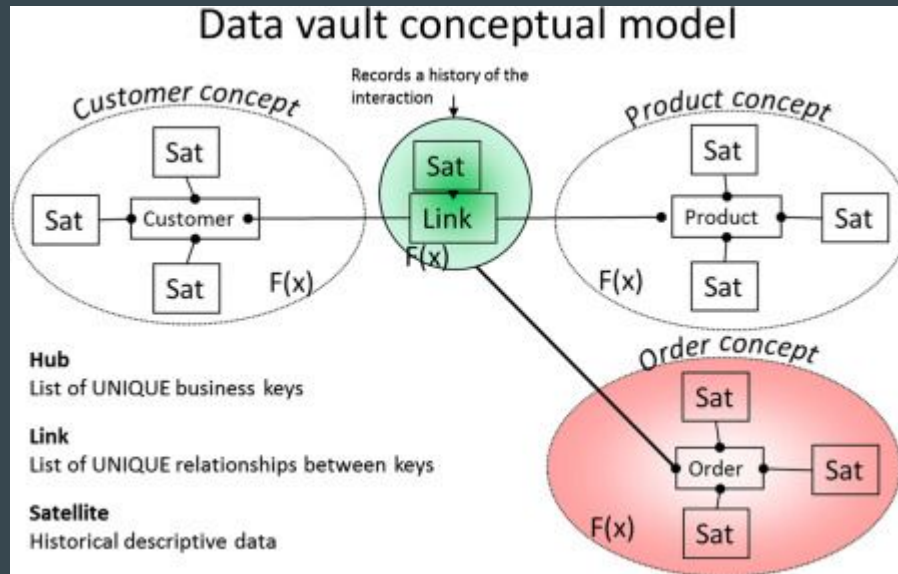
| | Option 1 Snowflake | Option 2 Star Schema | Option 3 Single Table |
|-------------------|---|--|---|
| |  |  |  |
| Response Time | Satisfactory | Good | Excellent |
| RAM consumption | Good | Good | Bad |
| Script run time | Good | Excellent | Bad |
| Flexibility Model | Poor | Excellent | Excellent |
| Complexity Script | Poor | Excellent | Excellent |

Modelagem - Star x snowflake

Snowflake Schema vs Star Schema



Modelagem dimensional: Data Vault



Modelagem dimensional: Data Vault

- O que é a metodologia Data Vault?;
- Escalabilidade;
- Hubs, Links e Satélites;
- Processo de construção;

DW - SCD

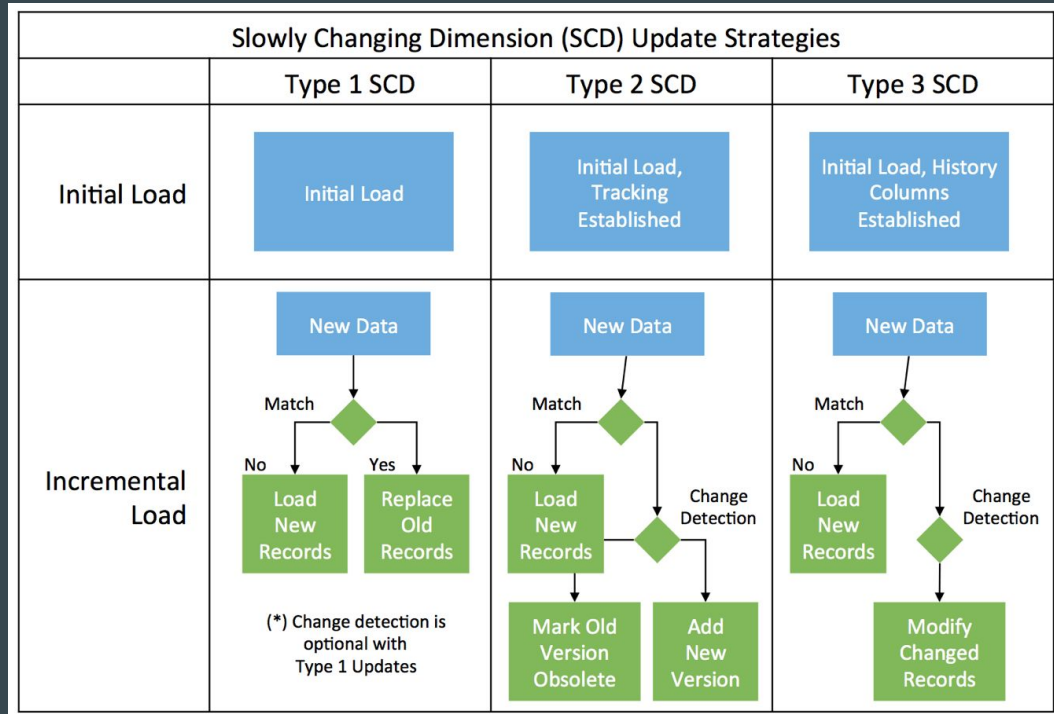
SCD

- SCD, ou "Slowly Changing Dimension", refere-se às técnicas utilizadas para gerenciar mudanças históricas em dimensões em um armazém de dados. Aqui estão os tópicos básicos sobre SCD
- Refere-se à forma como os armazéns de dados lidam com mudanças nas dimensões ao longo do tempo.
- Utilizado em armazéns de dados para refletir a evolução de informações ao longo do tempo.
- Importante para análises históricas e tomada de decisão.
- Garante a integridade e a acurácia dos dados históricos.
- Fornece uma visão clara da evolução dos dados ao longo do tempo.

SCD - Tipo 0 (Histórico não rastreado)

- Dados não são alterados após a inserção inicial;
- Não há rastreamento de mudanças ou histórico de atualizações;
- Útil em casos onde a consistência histórica é crucial e as alterações subsequentes não são relevantes ou desejadas.

SCD - Tipos



SCD - Tipo 1 (Sobrescrição)

Type 1 Slowly Changing Dimension

| Product Dim (Source) | | | Product Dim (Target) | | | |
|----------------------|------------|--|----------------------|------|-------------------|---------------------------|
| Product Name | Product ID | Product Descr | Product Name | SID | Source Product ID | Product Descr |
| 10 inch box | 010 | 10 inch glued pasted box | 10 inch box | 0001 | 010 | 10 inch pasted box |
| 12 inch box | 012 | 12 inch glued box | 12 inch box | 0002 | 012 | 12 inch glued box |

SCD - Tipo 1 (Sobrescrição)

- Substitui os dados antigos pelos atuais;
- Apenas a informação mais recente é retida;
- Útil em casos onde não é necessário manter dados históricos;

SCD - Tipo 2 (Nova instância)

Type 2 Slowly Changing Dimension

| Product Dim (Source) | | | Product Dim (Target) | | | | | | |
|----------------------|------------|--------------------|----------------------|------|-------------------|--------------|--------------------|--------------|-------------|
| Product Name | Product ID | Product Descr | | SID | Source Product ID | Product Name | Product Descr | EFF_START_DT | EFF_END_DT |
| 12 inch box | 012 | 12 inch glued box | → | 0001 | 012 | 12 inch box | 12 inch glued box | Jan-01-1753 | Dec-31-9999 |
| 10 inch box | 010 | 10 inch glued box | ↙ | 0002 | 010 | 10 inch box | 10 inch glued box | Jan-01-1753 | May-12-06 |
| | | 10 inch pasted box | | 0003 | 010 | 10 inch box | 10 inch pasted box | May-12-06 | Dec-31-9999 |

SCD - Tipo 2 (Nova instância)

- Cria um novo registro a cada mudança;
- Útil em casos onde dados históricos são necessários;
- Pode exigir uma alta capacidade de armazenamento;

SCD - Tipo 3 (Mantém ambos os valores)

Type 3 Model – Slowly Changing Dimension

Before The Change

Vendor Dimension Table

| Vendor Key | Vendor Name | Old Phone | Current Phone | Effective Date |
|------------|--------------------------|-----------|---------------|----------------|
| 001 | At-Your-Service Plumbing | | 202-555-8639 | 2/1/2001 |



After The Change

Vendor Dimension Table

| Vendor Key | Vendor Name | Old Phone | Current Phone | Effective Date |
|------------|--------------------------|--------------|---------------|----------------|
| 001 | At-Your-Service Plumbing | 202-555-8639 | 858-555-6555 | 12/15/2008 |

SCD - Tipo 3 (Mantém ambos os valores)

- Mantém os registros novos e antigos;
- Possibilita manter dados históricos sem ocupar muito armazenamento;
- Útil em cenários em que poucos dados históricos precisam ser armazenados;

Dúvidas

Modern Data Stack

Contexto

Contexto - AS-IS

- CSV, Excel, Google Sheets
- Transformações manuais
- Dashboards de BI
- Transformações recorrentes
- Times de dados tradicionais
- Ferramentas tradicionais

MDS - To-BE

- Como os avanços da tecnologia permitiram a transição de ETL para ELT
- Times de dados modernos
- O papel do Analytics Engineer
- Ferramentas de dados modernas (Modern Data Stack)
- Como o DBT se encaixa no Modern Data Stack

MDS - Modern Data Stack

MDS - Modern Data Stack

Info

The main core of the modern data stack is to simplify the data pipeline and make easy for the business to consume, process and deliver the data.

Data Sources



Info

Ingestion of data using a web UI that connects multiple data sources, batch 5 in 5 minutes minimum, simple and objective.

Ingestion



Orchestration & Processing



Info

Orchestrate the processing jobs with Airflow running DBT scripts.

raw data

Data Store



Consumers



Info

Consume processed data using visualization tools.

Info

Store the raw data inside Snowflake, the data platform. Optimize for data warehousing.

MDS - Modern Data Stack

- Data Sources (Sources)
- Loaders (EL)
- Cloud Data Warehouse (CDW)
- Transformation tools (T/ETL)
- Data Sinks (Exposures)
- Orchestration Tools

MDS - Modern Data Stack

- Data Sources (Sources)
 - Eventos
 - Bancos de dados
 - Arquivos
 - APIs
- Loaders (EL)
- Cloud Data Warehouse (CDW)
- Transformation tools (T/ETL)
- Data Sinks (Exposures)
- Orchestration Tools

MDS - Modern Data Stack

- Data Sources (Sources)
- Loaders (EL)
 - Airbyte
 - Kafka
 - AWS Glue
 - Azure Data Factory
 - GCP Data Flow
 - Script (python, go, shell, java, etc.)
- Cloud Data Warehouse (CDW)
- Transformation tools (T/ETL)
- Data Sinks (Exposures)
- Orchestration Tools

MDS - Cloud Data Warehouses (CDW)

- Escalabilidade
- Custo
 - Computação
 - Storage
 - Tempo de transferência

MDS - Modern Data Stack

- Data Sources (Sources)
- Loaders (EL)
- Cloud Data Warehouse (CDW)
 - Databricks
 - Snowflake
 - AWS Redshift
 - Azure SQL Pools (Synapse/Fabric)
 - GCP Bigquery
- Transformation tools (T/ETL)
- Data Sinks (Exposures)
- Orchestration Tools

MDS - EL+T/ETL

- Extrair e carregar dados sem transformação (dados cru)
- Manipular/Transformar dados no destino (Data Warehouse)
- Pipelines automatizados (orquestração)
- Geralmente manipulados por Engenheiros de dados
- Benefícios:
 - Evita repetição do processo de EL em caso de erro na transformação
 - Focado em levar dados para o DW
 - Transformar dados diretamente no DW
 - Possibilidade de escalar apenas a extração (EL) ou a transformação (T)
 - Facilidade de rodar as transformações novamente sem a necessidade de uma nova extração
 - Facilidade de criar/alterar as transformações sem alterar a extração (e vice-versa)
 - Facilidade de adicionar novas regras de negócio sem alterar todo o pipeline

MDS - Modern Data Stack

- Data Sources (Sources)
- Loaders (EL)
- Cloud Data Warehouse (CDW)
- Transformation tools (T/ETL)
 - DBT (T)
 - Spark (ETL)
- Data Sinks (Exposures)
- Orchestration Tools

MDS - Modern Data Stack

- Data Sources (Sources)
- Loaders (EL)
- Cloud Data Warehouse (CDW)
- Transformation tools (T/ETL)
- Data Sinks (Exposures)
 - Ferramentas de BI
 - Modelos de Machine Learning
 - Relatórios operacionais
- Orchestration Tools

MDS - Modern Data Stack

- Data Sources (Sources)
- Loaders (EL)
- Cloud Data Warehouse (CDW)
- Transformation tools (T/ETL)
- Data Sinks (Exposures)
- Orchestration Tools
 - Airflow
 - Dagster
 - Kestra

MDS - Modern Data Team

MDS - Times de Dados Modernos - Antes

- Engenheiro de dados (mais perto da TI)
- Analista de dados (mais perto do negócio)

MDS - Modern Data Stack



MDS - Times de Dados Modernos - Antes

- Engenheiro de dados (mais perto da TI):
 - Construir infraestrutura para armazenamento de dados (data lake, database, data warehouse)
 - Gerenciar processos de ETL/ELT
 - SQL, Python, ETL, orquestração
 - Sabe como construir
- Analista de dados (mais perto do negócio)

MDS - Times de Dados Modernos - Antes

- Engenheiro de dados (mais perto da TI)
- Analista de dados (mais perto do negócio):
 - Consultar dados
 - Construir dashboards e relatórios
 - Conhecimento das métricas e processos do negócio
 - SQL, Excel, ferramentas de visualização de dados
 - Sabe o que construir

MDS - Modern Data Stack



Data Engineer



Analytics Engineer

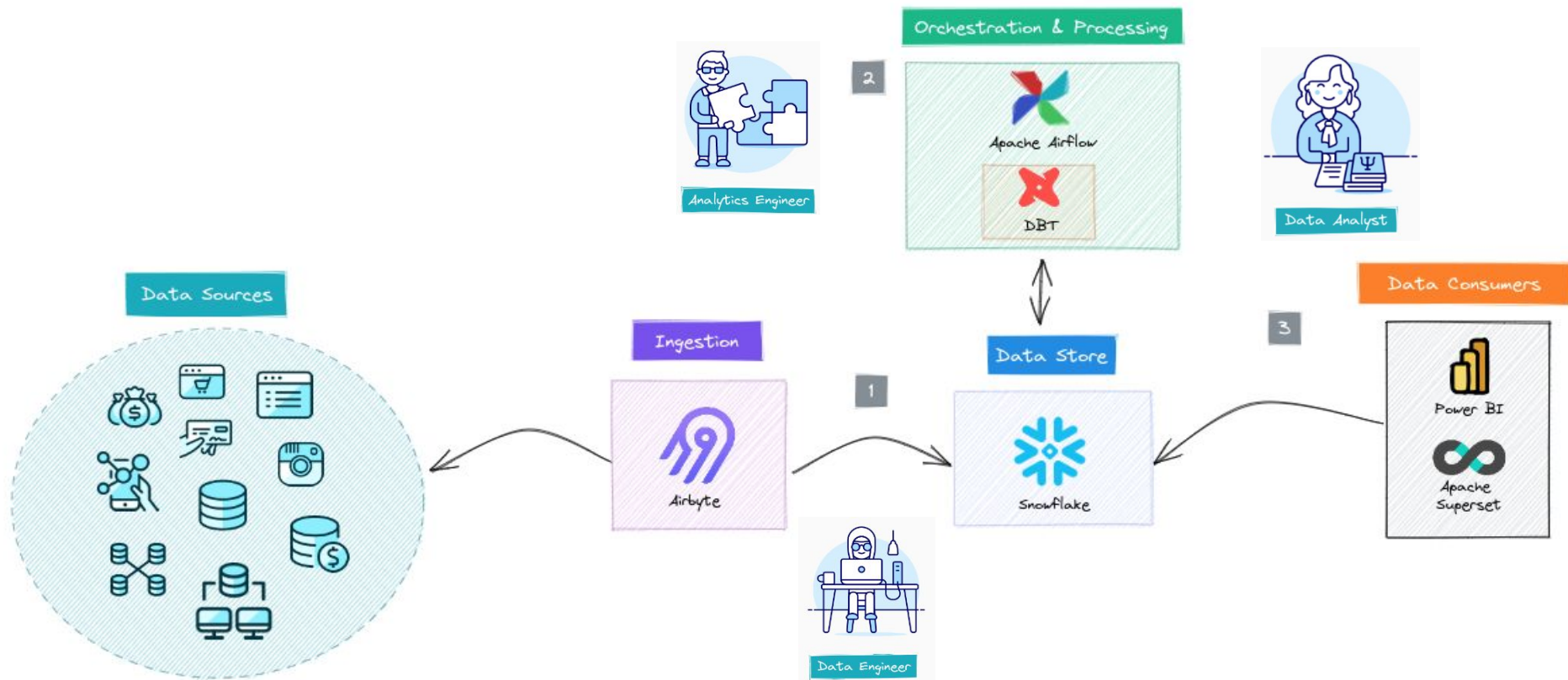


Data Analyst

MDS - Times de Dados Modernos - Analytics Engineer

- Engenheiro de dados (mais perto da TI)
- Analytics Engineer (meio de campo)
- Analista de dados (mais perto do negócio)

MDS - Modern Data Stack



MDS - Times de Dados Modernos - Engenheiro de dados

- Engenheiro de dados (mais perto da TI):
 - Construir infraestrutura para armazenamento de dados (data lake, database, data warehouse)
 - Gerenciar processos de streaming e batch (EL)
 - Carregar dados no Data Warehouse
 - Orquestração de pipelines de dados
 - Construir e manter plataformas de dados
 - SQL, Python, ETL, Streaming
 - Cuida da infraestrutura
- Analytics Engineer (meio de campo)
- Analista de dados (mais perto do negócio)

MDS - Times de Dados Modernos - Analytics Engineer

- Engenheiro de dados (mais perto da TI)
- Analytics Engineer (meio de campo)
 - Realiza limpezas e transformações de dados (T)
 - Cuida da qualidade dos dados (Data Quality)
 - Aplica boas práticas de desenvolvimento de software (DevOps)
 - Otimização de performance do Data Warehouse
 - Mantém documentação dos dados atualizada
 - Auxilia usuários de negócio a utilizar a plataforma de dados
 - Cuida dos dados
- Analista de dados (mais perto do negócio)

MDS - Times de Dados Modernos - Analista de dados

- Engenheiro de dados (mais perto da TI)
- Analytics Engineer (meio de campo)
- Cientista/Analista de dados (mais perto do negócio):
 - Consultar dados
 - Construir dashboards e relatórios
 - Conhecimento das métricas e processos do negócio
 - Trabalha com o negócio para entender requisitos de dados
 - SQL, Excel, ferramentas de visualização de dados
 - Cuida da informação

MDS - Comparativo

MDS - Comparativo - TBI vs MDS

| TÓPICOS | TDW | <u>MDW</u> |
|---------------------|--|--|
| Arquitetura | Centralizada | Distribuída |
| Armazenamento | Estrutura de dados rígida (<u>schema-on-write</u>) | Armazenamento de dados flexível (<u>schema-on-read</u>) |
| Escalabilidade | Escalabilidade vertical (adicionar recursos ao hardware existente) | Escalabilidade horizontal (adicionar nós de processamento) |
| Processamento | Processamento em lotes (<u>ETL</u>) | Processamento em tempo real (streaming) e em lote (<u>batch</u>) |
| Integração de dados | Integração de dados antes do armazenamento | Integração de dados após o armazenamento |

MDS - Comparativo - TDW vs MDW

| TÓPICOS | ANTIGAMENTE | ATUALMENTE |
|--------------------|------------------------|---|
| Armazenamento | Custoso e lento | Barato e rápido |
| Escalabilidade | Vertical | Horizontal |
| Memória | Processamento em disco | Processamento em memória |
| Tipos de dados | Primitivos | Complexos (<u>nested</u> , <u>array</u> , <u>json</u> , <u>xml</u> , etc.) |
| Modelagem de dados | Modelagem dimensional | Grandes tabelas |

MDS - Comparativo - ETL vs ELT

| TÓPICOS | <u>ETL</u> | ELT |
|------------------------------------|---|---|
| Definição | Processo de Extração, Transformação e Carregamento de dados. | Processo de Extração, Carregamento e Transformação de dados. |
| Transformação de dados | A transformação ocorre durante a fase de <u>ETL</u> , antes do carregamento no <u>destino final</u> . | A transformação ocorre após o carregamento dos dados no <u>destino final</u> . |
| Armazenamento temporário | Os dados podem ser armazenados temporariamente em um local intermediário. | Os dados são carregados diretamente no <u>destino final</u> sem armazenamento temporário adicional. |
| Uso de ferramentas <u>ETL</u> /ELT | Ferramentas especializadas de <u>ETL</u> são utilizadas para o processo. | Pode-se usar ferramentas de EL e aproveitar as funcionalidades do banco de dados. |

MDS - Comparativo - Data Team

| | <u>Tradicional</u> | <u>Moderno</u> |
|--|------------------------|--------------------|
| | <u>Analista</u> de ETL | Data Engineer |
| | DBA de banco MPP | Analytics Engineer |
| | <u>Analista</u> de BI | Data Analyst |

Dúvidas

Revisão de Python

Demo



<https://colab.research.google.com/>

Fast-Track Python para Engenharia de dados

<https://theplumbers.com.br/ft-python/>



Dúvidas