Dia 03

Revisão - Dia 02

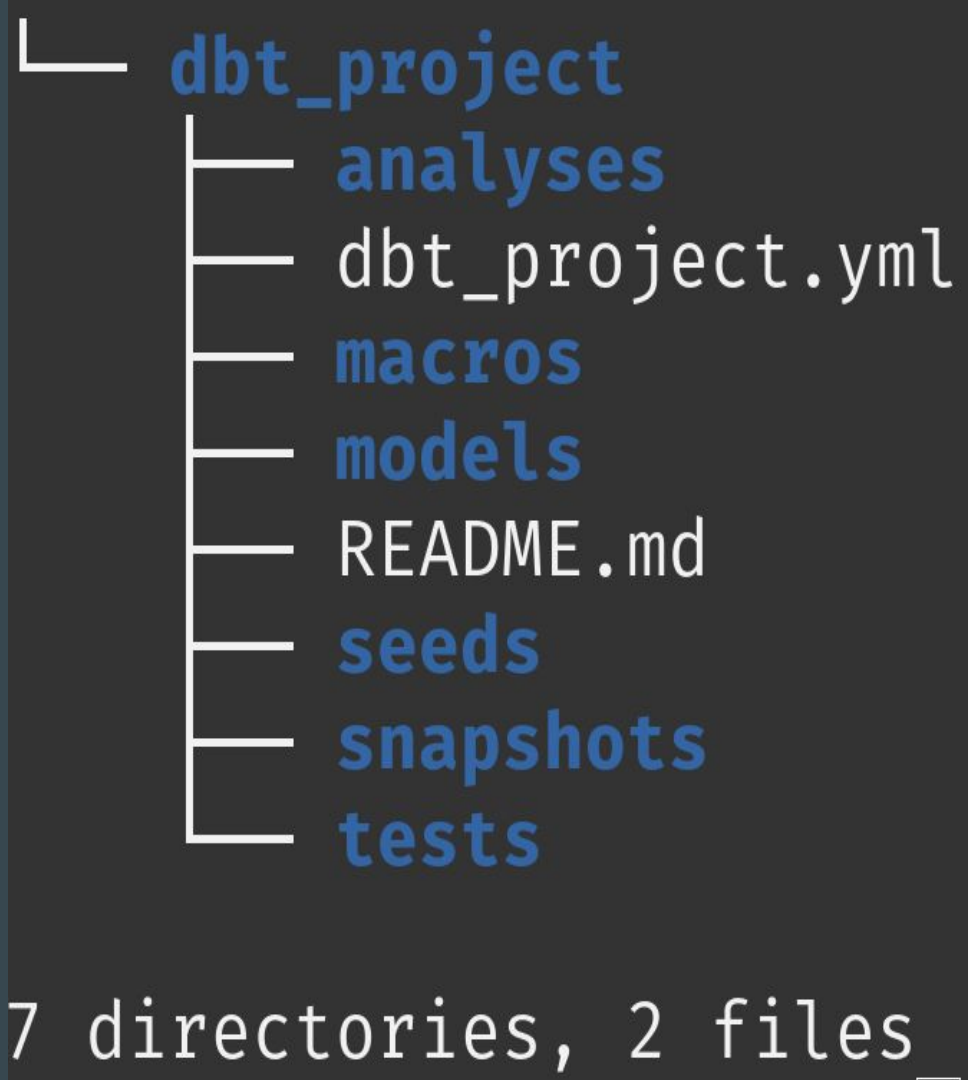
DBT - Fundamentos

Implementando Pipelines SQL com DBT

https://theplumbers.com.br/workshop-08/

DBT

Estrutura de Projeto



DBT - Estrutura de projeto

- Nome do projeto
  - Analyses
  - Macros
  - Models
  - Seeds
  - Snapshots
  - Tests
  - dbt_project.yml
  - readme.md

DBT - Nome do projeto

- é a pasta raiz do seu projeto DBT. Ele atua como um container para todos os componentes do seu projeto.

DBT - Analyses

- Arquivos de análise permitem que você execute consultas exploratórias e análises ad hoc. Embora sejam semelhantes aos modelos, eles não são materializados no banco de dados. Eles são ótimos para experimentar e testar lógica antes de integrá-la em modelos ou para análises que não precisam ser executadas regularmente.

DBT - Macros

- Macros são como funções que permitem reutilizar código SQL em seus modelos, testes, etc. Eles são úteis para encapsular lógica que é usada repetidamente em todo o projeto.

DBT - Models

- Estes são os scripts de transformação de dados escritos em SQL. Quando você "executa" o DBT, ele materializa esses modelos como view ou tabelas no seu data warehouse.

DBT - Seeds

- Seeds são arquivos CSV que você pode usar para carregar dados estáticos em seu data warehouse, como tabelas de dimensão com poucas linhas. O DBT carrega esses dados para você.

DBT - Snapshots

- Snapshots capturam a mudança de dados ao longo do tempo. Eles são usados para criar tabelas que armazenam versões históricas de linhas em um dado modelo.

DBT - Tests

- Permite escrever testes para seus modelos, como testar se uma coluna é única, ou se os valores em uma coluna estão dentro de um certo conjunto.

DBT - readme.md

- Documentação principal do seu projeto. O DBT pode gerar uma documentação baseada na web a partir desse arquivo e de descrições de modelo adicionais. Ajuda novos membros da equipe a entender o propósito e a estrutura do projeto e é uma boa prática manter atualizado.

DBT - dbt_project.yml

- Este é o arquivo de configuração principal para o seu projeto DBT. Ele define coisas como quais modelos devem ser executados e em que ordem, e outras configurações globais para o projeto.

DBT - Threads

- São o número de conexões simultâneas que o dbt tentará usar ao executar comandos contra o armazém de dados, sem violar as dependências entres os modelos.
- Quando você aumenta o número de threads, o dbt pode executar múltiplas operações em paralelo, acelerando o processo de transformação de dados ou qualquer outro comando executado.
- Porém, aumentar o número de threads aumenta o impacto no data warehouse, o que pode afetar outras ferramentas.
- Um bom número de threads é a quantidade de modelos em paralelos que seu projeto esteja construindo.
- Porém o número de queries concorrentes que um usuário está permitido para executar pode limitar a configuração de threads.

Dúvidas

Demo

...

Carregando dados no DW

DBT - Sources

DBT - Sources

- É uma referência a uma tabela de uma fonte de dados externa ao data warehouse. Com isso é possível referenciar/documentar as tabelas que deram origem aos dados crus.
- São arquivos YAML dentro da pasta Models (models/*.yml).
- Permite a visualização das tabelas de origem na documentação.
- Serve como um único lugar para configuração das tabelas de uma determinada fonte.
- Pode verificar a recência de atualização de cada tabela da fonte de dados.

DBT - Sources

- Para referência é utilizada a função `{{ source('<nome_fonte>', '<tabela>') }}` dentro dos modelos.

Demo

...

DBT Sources

Dúvidas

DBT - Modelos

DBT - Models

- São consultas SQL (select) em arquivos .sql que são armazenados dentro da pasta models (models/*.sql). Cada arquivo corresponde a um objeto no data warehouse.
- Usados para limpar, transformar e enriquecer dados brutos.

DBT - Modelos - Conceitos

- **Baseado em SQL:** Escreva transformações como você faria qualquer consulta SQL.
- **Reutilizável:** Os modelos podem se referir a outros modelos.
- **Materialização:** Como os modelos são transformados em objetos de banco de dados (views, tabelas, etc.).
- **Modelagem**: formatar os dados de seu formato original (cru) para o formato final (confiável).
    - Raw/Bronze/Source
    - Cleaned/Silver/Staging/Intermediate
    - Trusted/Gold/Final

DBT - Modelos - Nomenclaturas

- **Staging (stg):** Tabela com dados limpos e normalizados (1 staging para cada 1 source)
- **Intermediate (int):** Tabelas com referência a outras tabelas staging (nunca tabelas source).
- **Final (fct/dim):** Tabelas fatos (fct) ou dimensões (dim).

Dúvidas

Demo

...

DBT Modelos

DBT - Exposures

DBT - Linhagem - Exposures

- Exposures em DBT é um recurso que permite aos usuários declarar e documentar os pontos em que a base de código DBT é exposta a, ou é usada por, sistemas externos

DBT - Exposures

- Melhor documentação e transparência.
- Rastreabilidade: entenda a cadeia de dependências até o consumo final.
- Facilita a colaboração entre equipes de dados e equipes de negócios.

Demo

...

DBT Exposures

Dúvidas

DBT - Documentação

DBT - Documentação - Contexto

- Documentações bem construídas auxiliam usuários a responderem suas dúvidas sobre os dados sozinhos
- Permite que novos membros do time comecem a produzir com mais rapidez (onboarding)
- Documentação fica comumente desatualizada em relação ao código quando são realizadas em ferramentas/processos separados.

DBT - Documentação - Objetivos

- Facilitar o onboarding de novos integrantes na equipe
- Minimizar o tempo que um novo membro leva para começar a contribuir
- Diminuir a necessidade de comunicação síncrona para a passagem de conhecimento
- Aumenta as chances de ter férias tranquilas
- “Documentação como referência, não pessoas”
- Usuários podem responder suas próprias perguntas (de onde vem o dado, como ele é calculado, etc.)
- Geração de documentação de maneira automática e na mesma ferramenta
- Linhagem dos dados

DBT - Documentação - Benefícios

- Gerada automaticamente
- Atualizada automaticamente
- Codificação e documentação na mesma interface
- Codificação e documentação na mesma sintaxe

DBT - Documentação

- Gráfico de Linhagem
- Descrições (fonte, modelos, colunas, testes, etc.)
- Codígo SQL compilado de cada modelo

DBT - Documentação - Blocos

- Blocos de documentação são arquivos markdown utilizados para uma documentação mais robusta.
- É possível utilizar {{ doc(“<bloco_documentação>”) }} para chamá-los.
- Ao usar essa função, você pode fazer referência a blocos de documentação definidos em outro lugar no projeto, tornando seu código mais limpo e a documentação centralizada.

DBT - Documentação

- Fontes de dados (sources)
- Destinos de dados (exposures)
- Arquivos estáticos (Seeds)
- Linhagem dos modelos (dependência entre os modelos)
- Testes aplicados
- Descrição dos objetos (tabelas, esquemas, colunas, etc.)
- Códigos SQL compilados
- Blocos de código em markdown
- Disponibilização em formato HTML

DBT - Linhagem - DAGs

- Sources (verde)
- Modelos (azul)
- Exposures (laranja)

Dúvidas

Demo

...

DBT Documentação

DBT - Testes

DBT - Testes - Objetivos

- Em desenvolvimento: garante que se esteja produzindo o que se espera
- Em produção: avisa quando alguma coisa expectativa falha, antes que seja disponibilizado para a área de negócio

DBT - Testes - Contexto

- Testes são suposições/afirmações/expectativas que se tem sobre os dados.
- A maioria dos profissionais de dados realizam testes manuais:
  - Exportar e validar um arquivo CSV
  - Rodar queries manualmente e repetidamente (e.g. count (*), count(distinct *), etc.)
- Tarefas repetitivas devem ser automatizadas sempre que possível
- Testes manuais não escalam
- Testes melhoram a cobertura e confiança no código

DBT - Testes

- Engenheiros de software possuem uma grande cultura de aplicar testes. Enquanto engenheiros de dados não a possuem, ainda que os dados estejam constantemente mudando.
- Sem testes, os problemas/bugs serão provavelmente descobertos em produção, o que diminui a confiança da área de Business na área de TI.
- Várias horas são perdidas todos os dias corrigindo problemas descobertos em produção (geralmente são problemas críticos que entram na frente de outras demandas).

DBT - Testes

- DBT vai procurar no projeto por testes especificados em arquivos YAML.
- DBT auxilia desenvolvedores a escalar testes entre projetos de forma rápida e fácil.
- DBT disponibiliza 4 testes já pré-desenvolvidos (testes genéricos)
  - Unique
  - Not null
  - Values accepted
  - Relationship

Demo

...

DBT Testes

Dúvidas