

1. Título do Projeto

Projeto Final: Engenharia e Análise de Dados com T-SQL

2. Visão Geral e Objetivos

Este projeto foi desenhado para simular um ciclo completo de engenharia e análise de dados, com a restrição de utilizar **exclusivamente T-SQL** para a manipulação de dados. Os alunos deverão encontrar um dataset público massivo, projetar um banco de dados otimizado, construir um processo de ETL e, finalmente, **transformar os dados brutos em insights de negócio acionáveis**. O objetivo é demonstrar maestria em modelagem, ETL, performance, segurança e, crucialmente, na criação de análises que respondam a perguntas de negócio.

3. Temas e Fontes de Dados (Datasets)

O foco do projeto é a análise de dados reais. Os grupos devem escolher um tema e um ou mais datasets públicos que permitam a criação de um banco de dados rico e complexo.

Sugestões de Temas:

- **Análise Esportiva:** Histórico de campeonatos de futebol, estatísticas de jogadores, resultados de corridas.
- **Saúde Pública:** Análise de dados do DataSUS sobre doenças, internações, vacinação.
- **Dados Governamentais:** Análise de despesas públicas, licitações, dados do censo (IBGE).
- **Mercado Financeiro:** Séries históricas de ações, dados de companhias abertas (CVM).
- **Ciência e Meio Ambiente:** Dados de biodiversidade, registros climáticos, qualidade do ar.

Onde Encontrar Datasets:

- Kaggle: <https://www.kaggle.com/datasets>
- Portal Brasileiro de Dados Abertos: <https://dados.gov.br>
- Google Dataset Search: <https://datasetsearch.research.google.com>
- IBGE: <https://www.ibge.gov.br/estatisticas/>
- DataSUS: <https://datasus.saude.gov.br>

4. Requisitos Obrigatórios

1. **Processo de ETL com T-SQL:** Entregar os scripts T-SQL (Stored Procedures, etc.) usados para o ETL.
2. **Modelagem:** Entregar o Diagrama Entidade-Relacionamento (DER) completo.
3. **Escopo:** Mínimo de **20 tabelas** no modelo final.
4. **Volume:** Mínimo de **200.000 registros** na tabela fato principal.
5. **Documentação:** Criar um "Dicionário de Dados" e um manual de uso para as rotinas.
6. **Versionamento:** Todo o código SQL e a documentação no GitHub.
7. **Gerenciamento:** Quadro Kanban (Trello, GitHub Projects) para gestão das tarefas.
8. **Programação T-SQL:** Implementar no mínimo 5 Views, 5 Stored Procedures (sendo 3 de ETL/CRUD e **2 analíticas**), 2 Functions e 2 Triggers.
9. **Segurança (DCL):** Criar no mínimo 3 perfis de acesso (Roles) com permissões distintas.
10. **Plano de Análise:** Entregar um documento descrevendo **quais perguntas de negócio serão respondidas** com os dados. (Ex: "Qual o perfil de gasto dos deputados por estado?", "Qual a correlação entre vacinação e queda de internações?").
11. **(Opcional) Dashboard de BI:** Conectar a base de dados a uma ferramenta de visualização (Power BI, Tableau, Looker Studio, Excel) e criar um painel que consuma os dados das Stored Procedures analíticas.

5. Métodos de Avaliação

- **Avaliação Contínua (40%):** Qualidade das apresentações semanais e commits no GitHub.
- **Qualidade das Entregas (30%):** Análise dos artefatos (modelo, scripts de ETL e SQL, plano de análise).
- **Projeto e Apresentação Final (30%):** Avaliação do projeto consolidado e defesa final.

6. Cronograma Detalhado (13 Semanas)

Fase 1: Concepção, Análise de Dados e Modelagem (Semanas 1-3)

- **Semana 1: Kick-off e Análise Exploratória**
 - **Foco:** Definição do tema, encontrar e analisar o dataset bruto. Criação do repositório/Kanban.
- **Semana 2: Requisitos e Plano de Análise**
 - **Foco:** Com base na análise, criar o Modelo Lógico e o **documento do Plano de Análise**, definindo as perguntas de negócio.
- **Semana 3: Modelagem Física**
 - **Foco:** Finalizar o Modelo Físico com tipos de dados e constraints do SQL Server.

Fase 2: Estruturação, ETL e Carga (Semanas 4-6)

- **Semana 4: Construção do Banco (DDL)**
 - **Foco:** Criar todos os scripts DDL e executar a estrutura do banco.
- **Semana 5: Desenvolvimento do Processo de ETL em T-SQL**
 - **Foco:** Codificar as Stored Procedures de ETL. Testar a extração e transformação com uma amostra dos dados.
- **Semana 6: Execução do ETL e Consultas Exploratórias**
 - **Foco:** Executar o ETL completo. Criar consultas DQL para validar a carga e explorar os dados.

Fase 3: Otimização e Análise (Semanas 7-10)

- **Semana 7: Views e Índices**
 - **Foco:** Criar views para simplificar a análise e índices para otimizar as futuras consultas analíticas.
- **Semana 8: Stored Procedures Analíticas**
 - **Foco:** Implementar as **Stored Procedures que respondem às perguntas do Plano de Análise**, gerando métricas e relatórios.
- **Semana 9: Triggers e Transações (DTL)**
 - **Foco:** Criar triggers para auditoria/validação e controlar transações.
- **Semana 10: Segurança (DCL)**
 - **Foco:** Implementar a estratégia de segurança com logins, usuários e perfis.

Fase 4: Finalização e Entrega (Semanas 11-13)

- **Semana 11: Documentação e (Opcional) Dashboard**
 - **Foco:** Finalizar a documentação. Iniciar a criação do dashboard de BI (opcional).
- **Semana 12: Preparação da Entrega Final**
 - **Foco:** Gravar o vídeo de demonstração e preparar a apresentação final.
- **Semana 13: Apresentação Final**
 - **Foco:** Defesa do projeto, demonstrando o processo de ETL e os insights gerados pelas análises.

7. Checklist de Avaliação do Projeto

Este checklist será usado para guiar a avaliação final do projeto, cobrindo os aspectos técnicos e de qualidade.

1. Modelagem e Estrutura (DDL)

- ☐ **Qualidade do DER:** O diagrama é claro, completo e representa corretamente as regras de negócio?
- ☐ **Normalização:** As tabelas estão adequadamente normalizadas (até 3FN)? As decisões de desnormalização (se houver) são justificadas?
- ☐ **Tipos de Dados:** Os tipos de dados escolhidos são os mais eficientes para os dados armazenados?
- ☐ **Constraints:** Foram usadas constraints (PRIMARY KEY, FOREIGN KEY, UNIQUE, CHECK, DEFAULT) para garantir a integridade dos dados?

- [] **Scripts DDL:** Os scripts são bem organizados, comentados e idempotentes (podem ser executados várias vezes sem gerar erros)?

2. Engenharia de Dados (ETL & DML)

- [] **Extração:** O método de extração (BULK INSERT, etc.) é robusto e eficiente?
- [] **Transformação:** A lógica de limpeza e transformação dos dados (tratamento de nulos, conversão de tipos, etc.) é completa e correta?
- [] **Carga (DML):** A carga dos dados nas tabelas finais é performática e garante a integridade referencial?
- [] **Modularidade do ETL:** O processo de ETL está bem estruturado em Stored Procedures, facilitando a manutenção e a reexecução?

3. Análise de Dados (DQL)

- [] **Complexidade das Consultas:** As Stored Procedures analíticas usam JOINS complexos, agregações (GROUP BY), subqueries ou CTEs de forma eficaz?
- [] **Relevância das Análises:** As análises respondem diretamente às perguntas definidas no "Plano de Análise"?
- [] **Performance (DQL):** As consultas analíticas são otimizadas? Foi demonstrada a melhoria com o uso de índices?
- [] **Clareza dos Resultados:** Os resultados retornados pelas procedures são claros e fáceis de entender?

4. Programação e Controle (DTL & DCL)

- [] **Transações (DTL):** O uso de BEGIN TRAN, COMMIT, ROLLBACK é aplicado corretamente em operações críticas para garantir a atomicidade?
- [] **Controle de Acesso (DCL):** A implementação de Roles e permissões (GRANT, DENY) é granular e segue o princípio do menor privilégio?
- [] **Lógica de Negócio:** Stored Procedures, Functions e Triggers encapsulam regras de negócio de forma correta e segura?

5. Qualidade Geral e Profissionalismo

- [] **Documentação:** O Dicionário de Dados e o manual do projeto são completos e úteis?
- [] **Organização do Código:** Os scripts no GitHub estão bem organizados em pastas, com nomes claros e um README.md instrutivo?
- [] **Versionamento (Git):** O histórico de commits é consistente, com mensagens claras que descrevem a evolução do projeto?
- [] **Apresentação:** A apresentação final e o vídeo de demonstração são claros, profissionais e demonstram domínio sobre o projeto?
- [] **Conhecimento dos Dados:** O grupo demonstra profundo conhecimento sobre o dataset escolhido e os insights que ele pode gerar?

8. Estrutura da Apresentação Semanal (Weekly)

O objetivo da weekly é ser um **relatório de progresso rápido e objetivo**, não uma aula. O foco é em mostrar o que foi feito e os resultados alcançados.

- **Duração Máxima:** 5 a 7 minutos por grupo.
- **Formato Sugerido:**
 1. **Progresso da Semana (1 min):** Mostrar o quadro Kanban. Quais tarefas foram movidas para "Concluído"?
 2. **Demonstração do Resultado (3-4 min):** Foco no "show, don't tell".
 - *Fase de Modelagem:* Mostrar o diagrama ER atualizado.
 - *Fase de DDL:* Executar um SELECT em tabelas recém-criadas para mostrar que existem.
 - *Fase de ETL:* Mostrar a contagem de linhas (SELECT COUNT(*)) antes e depois de uma carga.
 - *Fase de Análise:* Executar uma Stored Procedure analítica e mostrar a tabela de resultados.
 3. **Commits Relevantes (1 min):** Mostrar a tela de histórico de commits do GitHub da semana e explicar brevemente 1 ou 2 commits importantes.
 4. **Plano e Impedimentos (1 min):** Qual a principal meta para a próxima semana? Existe algum obstáculo?

9. Estrutura da Apresentação Final

A apresentação final deve ser tratada como um **relatório técnico para stakeholders de um projeto real**. O objetivo é "vender" o projeto, demonstrando seu valor, a qualidade da engenharia e os insights gerados.

- **Duração:** 20 a 25 minutos por grupo.

- **Roteiro e Papéis Sugeridos:**

1. Visão de Negócio e Arquitetura (5 min) - Apresentador 1 (Arquiteto de Dados)

- **O Problema:** Apresentar o dataset escolhido e as perguntas de negócio definidas no Plano de Análise.
- **A Solução:** Apresentar o Diagrama Entidade-Relacionamento (DER) final. Justificar as principais decisões de modelagem (normalização, tipos de dados) e como a arquitetura suporta as análises propostas.

2. Engenharia de Dados em Ação: O ETL (7 min) - Apresentador 2 (Engenheiro de Dados)

- **O Desafio:** Descrever os principais desafios encontrados nos dados brutos (dados sujos, formatos inconsistentes, etc.).
- **A Implementação:** Explicar o pipeline de ETL construído em T-SQL. Mostrar trechos das Stored Procedures de extração, transformação e carga. Demonstrar a execução de uma etapa chave do ETL.

3. Análise e Geração de Valor (8 min) - Apresentador 3 (Analista de Dados)

- **Os Resultados:** Esta é a parte principal. Executar ao vivo as Stored Procedures analíticas.
- **Os Insights:** Para cada resultado, explicar o que ele significa e como responde a uma das perguntas de negócio.
- **(Opcional) O Dashboard:** Se o grupo fez o dashboard de BI, mostrá-lo em ação, explicando os gráficos e como eles facilitam a visualização dos insights.

4. Conclusão e Próximos Passos (2 min) - Apresentador 1 (Arquiteto de Dados)

- **Resumo:** Sumarizar as principais conclusões e o valor gerado pelo projeto.
- **Futuro:** Sugerir possíveis evoluções (novas análises, outros datasets a serem integrados).
- Abertura para perguntas.