



Projeto Final de Avaliação – Integração de Dados (5 Módulos)

Título do Projeto

Implementação de um Sistema Híbrido de Integração e Governança de Dados para uma Visão 360º do Cliente

Objetivo

O objetivo deste projeto é permitir que o aluno demonstre a sua proficiência em todas as áreas abordadas na disciplina de Integração de Dados: desde os fundamentos e processos ETL, passando pela qualidade dos dados e seleção de ferramentas, até aos contextos avançados de integração e governança.

Cenário de Negócio

Uma empresa de serviços financeiros encontra-se a modernizar o seu sistema de gestão de clientes. O seu objetivo é desenvolver uma Visão 360º do Cliente unificada, integrando dados provenientes de três fontes distintas:

- 1. Sistema Legado (Batch):** Base de dados histórica de clientes (dados estruturados).
- 2. Sistema de Suporte (API/Web Service):** Informações de contactos e moradas (dados semi-estruturados).
- 3. Sistema de Transações (Streaming):** Fluxo contínuo de eventos de transações financeiras (dados em tempo real).

Requisitos e Tarefas (Correspondência aos 5 Módulos)

O projeto deve ser desenvolvido em cinco fases, cada uma associada a um módulo da disciplina.

Fase 1 – Análise e Design Conceptual

(Módulo 1: Fundamentos de Integração de Dados)

Tarefa: Definir o problema, os objetivos e a arquitetura conceptual da solução.

1. **Definição de Requisitos:** Identifique e descreva os principais desafios de integração presentes neste cenário (ex.: latência, diversidade de fontes, consistência).
2. **Arquitetura:** Proponha um modelo de arquitetura de integração e justifique a sua escolha.
3. **Glossário:** Elabore um glossário com 5 termos técnicos essenciais (ex.: Data Mart, Metadata, Latência) relevantes para este projeto.

Fase 2 – Mapeamento e Processo ETL

(Módulo 2: Processos ETL)

Tarefa: Detalhar o processo de Extração, Transformação e Carga/Carregamento/Load para a fonte do Sistema Legado.

4. **Mapeamento de Dados:** Seleccione 5 campos críticos do Sistema Legado (ex.: ID_Cliente_Legado, Data_Nascimento, Status_Conta). Crie uma tabela de mapeamento para o Data Mart de destino, indicando o tipo de transformação necessária (ex.: conversão de formato, agregação, derivação).
5. **Pseudocódigo de Transformação:** Apresente o pseudocódigo para uma transformação complexa (ex.: cálculo da idade a partir da Data_Nascimento e normalização do Status_Conta de 5 valores possíveis para 3 categorias).
6. **Estratégia de Carga/Carregamento/Load:** Descreva a estratégia de Carga/Carregamento/Load adequada (ex.: Full Load, Incremental Load – CDC) para atualizar o Data Mart a partir do Sistema Legado.

Fase 3 – Qualidade e Consistência de Dados

(Módulo 3: Qualidade e Consistência de Dados)

Tarefa: Implementar Data Profiling, regras de Data Cleansing e validações de consistência.

7. **Data Profiling:** Descreva o processo de profiling aplicado aos dados de morada provenientes da API do Sistema de Suporte. Identifique as dimensões de qualidade mais críticas (ex.: Completude, Validade).

8. **Regras de Limpeza:** Defina 3 regras de Data Cleansing aplicadas ao campo Nome (ex.: remoção de caracteres especiais, uniformização de maiúsculas/minúsculas).
9. **Consistência:** Proponha uma regra de resolução de conflitos quando o NIF do cliente difere entre o Sistema Legado e o Sistema de Suporte. Indique qual das fontes deve ser considerada como Golden Record.

Fase 4 – Seleção e Justificação de Ferramentas

(Módulo 4: Ferramentas de Integração)

Tarefa: Selecionar e justificar as ferramentas tecnológicas adequadas.

10. **Ferramentas ETL:** Escolha uma ferramenta ETL comercial (ex.: Informatica PowerCenter, Talend) e uma ferramenta Open Source (ex.: Apache NiFi, Pentaho Data Integration). Justifique ambas com base em critérios como escalabilidade, curva de aprendizagem e custo.
11. **Ferramenta de Streaming:** Selecione uma plataforma para processamento em tempo real das transações (ex.: Apache Kafka + Spark Streaming/Flink) e justifique a escolha.
12. **MDM:** Indique uma ferramenta de Master Data Management (MDM) adequada para gerir a Visão 360º do Cliente e explique como esta se integra com as ferramentas ETL e de streaming.

Fase 5 – Integração em Contextos Avançados

(Módulo 5: Integração em Contextos Avançados)

Tarefa: Desenhar a integração em tempo real e os mecanismos de Governança de Dados.

13. **Design de Streaming:** Detalhe o pipeline de integração para o fluxo das Transações Financeiras. Descreva como o Change Data Capture (CDC) seria utilizado e como a Arquitetura Kappa (ou Lambda) seria aplicada.
14. **Integração de Variedade:** Caso as transações sejam disponibilizadas em XML, descreva o processo de parsing e normalização necessário para integrá-las no Data Mart relacional.
15. **Governança de Dados:** Defina os papéis de Data Owner e Data Steward para o domínio “Cliente” e descreva uma política de governança que assegure conformidade com o RGPD, especialmente no tratamento de dados como NIF e morada.

Anexos

Foram criados três ficheiros, que anexo:

- *clientes_master_data.csv*: Simula a Base de Dados de Clientes (dados estruturados).
- *interacoes_web_stream.json*: Simula o Stream de Interações Web (dados semi-estruturados).
- *mock_data_description.md*: Descreve as falhas intencionais introduzidas nos dados (NIF inválido, sentiment score fora do domínio, registo órfão), que deverão identificar e tratar na Fase 3 do projeto.

Estes ficheiros permitem que possam testar a vossa lógica de parsing, normalização e, crucialmente, a sua capacidade de implementar regras de validação de dados e de lidar com a consistência entre fontes.

Nota: Poderão criar/utilizar documentos criados por vocês desde que devidamente explicados e idênticos aos que foram partilhados.

Conteúdos a entregar:

O aluno deve entregar um único **Relatório Técnico** em formato PDF contendo:

- **Relatório Técnico:**
 - Capa e Índice.
 - Introdução e Enquadramento do Cenário.
 - Desenvolvimento das 5 Fases (com diagramas e tabelas).
 - Conclusão.
- **Anexo:** Ficheiros de código utilizados para simular os processos.

Critérios de Avaliação

Critério	Módulo Abrangido	Peso (%)	Descrição
Design Conceptual	Módulo 1	15%	Clareza na definição de requisitos e adequação da arquitetura proposta.
Processo ETL	Módulo 2	20%	Rigor no mapeamento de dados e correção do pseudocódigo de transformação.
Qualidade de Dados	Módulo 3	20%	Adequação das regras de cleansing, profiling e resolução de inconsistências.
Seleção de Ferramentas	Módulo 4	15%	Justificação técnica e conhecimento das ferramentas ETL, Streaming e MDM.
Integração Avançada	Módulo 5	20%	Detalhe no design de streaming (CDC/Kappa) e definição dos aspectos de Governança.
Relatório e Apresentação	Geral	10%	Organização, clareza, rigor técnico e cumprimento das normas de formatação.

Nota: A aprovação na disciplina requer uma nota mínima de 50% no Projeto Final.

Data de Entrega e Defesa/Apresentação:

16 de dezembro de 2025