

DADOS DOS ALUNOS(A)

Nomes dos integrantes: Rodrigo Bertozzi de Castro(1096500), Lara Meneses(1209842), Lídia Fernandes Magalhães(718129), Vanessa Pais(1500283)
Atividade: Trabalho Prático 6

Valor da atividade: **30 pontos**

INSTRUÇÕES

Esta atividade deve ser entregue em modelo de relatório em formato .zip. Deverá constar no arquivo:

- **Relatório do trabalho em formato .pdf**
- **Cabeçalho com:**
 - **Nome completo do aluno**
 - **Matrícula**
 - **Nome da atividade**
- **Todos os scripts desenvolvidos e executados corretamente**
- **Apresentações visuais dos dados solicitados**

DESCRIÇÃO DO PROBLEMA

Uma universidade precisa coletar dados de seus alunos das diversas unidades e cursos para construir um sistema analítico que permita o armazenamento e analisar em tempo real as notas de cada estudante, curso, calcular médias e, num futuro, gerar previsões sobre as disciplinas para que ela consiga tomar decisões itens passíveis de ajustes para melhorar a eficiência e produtividade dos alunos.

As informações passadas pela empresa contratante são:

- a. Nome da instituição: FACULDADE PRODATA
- b. Dados dos cursos ofertados:
 - i. ENGENHARIA DE COMPUTAÇÃO
 1. Código do curso: 1
 2. Períodos: 10
 3. Carga horária prevista por disciplina: **até 92 horas**
 4. Disciplinas por período: 8
 5. Média de alunos por turma: 50
 - ii. CIÊNCIA DA COMPUTAÇÃO
 1. Código do curso: 2
 2. Períodos: 8
 3. Carga horária prevista por disciplina: **até 64 horas**
 4. Disciplinas por período: 8
 5. Média de alunos por turma: 62
 - iii. SISTEMA DE INFORMAÇÕES
 1. Código do curso: 3
 2. Períodos: 8
 3. Carga horária prevista por disciplina: **até 64 horas**
 4. Disciplinas por período: 7
 5. Média de alunos por turma: 48
 - iv. JOGOS DIGITAIS
 1. Código do curso: 4
 2. Períodos: 6
 3. Carga horária prevista por disciplina: **até 48 horas**
 4. Disciplinas por período: 6
 5. Média de alunos por turma: 32
 - v. TECNÓLOGO EM REDES DIGITAIS
 1. Código do curso: 5
 2. Períodos: 4
 3. Carga horária prevista por disciplina: **até 48 horas**
 4. Disciplinas por período: 6
 5. Média de alunos por turma: 28
- c. Dados necessários dos alunos:
 - i. Matrícula (número aleatório)
 - ii. Nome (Nome e sobrenome)
 - iii. Sexo (M | F)
 - iv. Cadastro geral da pessoa (um código único, similar ao CPF)
 - v. Data de nascimento (dd/mm/aaaa)
 - vi. E-mail para contato (nome.sobrenome@trabalhopuc.br)
 - vii. Endereço (pode ser cidade e estado)
 - viii. Telefone celular (+55 XX XXXXX-XXXX)
 - ix. Curso (código do curso)
 - x. Data de ingresso (dd/mm/aaaa)

- xi. Data prevista de conclusão (dd/mm/aaaa)
- xii. Data de conclusão (dd/mm/aaaa)
- xiii. Status da matrícula (ativo | Trancado | Cancelado)

d. Dados de disciplinas dos cursos:

- i. Código da disciplina (deve iniciar em 1, não pode repetir).
- ii. Nome da disciplina (Sugestão: busque informações sobre os cursos fornecidos acima para que os dados sejam mais coesos).
- iii. Tipo: Teórica | prática
- iv. Nota de corte: 70 pontos
- v. Carga horária prevista: Gerar um valor aleatório de acordo com a carga horária prevista por disciplina. O número deve ser **SEMPRE** maior que 16 horas e menor que a carga horária prevista por disciplina).
- vi. Frequência mínima (deve ser equivalente à 75% da carga horária)

e. Dados de relacionamento entre alunos, cursos e disciplinas:

- i. Código do curso
- ii. Código da disciplina
- iii. Período (mm/aaaa)
- iv. Matrícula do aluno
- v. Nota do aluno
- vi. Frequência do aluno
- vii. Status (aprovado | reprovado | exame especial)
- viii. Motivo reprovação (nota | frequência)

f. Dados de atividades realizadas pelo aluno:

- i. Matrícula do aluno
- ii. Código do curso
- iii. Código da disciplina
- iv. Período
- v. Código da atividade (número sequencial de acordo com a atividade)
- vi. Nome da atividade (Exercício | Trabalho | Avaliação)
- vii. Valor da atividade
- viii. Nota obtida pelo aluno

Para facilitar o desenvolvimento desse trabalho, considere que:

- Um aluno não pode estar matriculado simultaneamente em mais de um curso, mas podem fazer mais de um curso (para cada ingresso, um novo código de matrícula deverá ser gerado);
- Um aluno não pode repetir o mesmo curso;
- Considere que cada período contenha **entre 6 e 10 disciplinas** (precisa estar nesse intervalo).
- A idade mínima para um aluno ingressar na universidade é a partir dos 18 anos;

- O aluno **não pode trancar/cancelar uma disciplina**, mas sim, o seu curso como um todo.
- A nota das atividades deve ser um valor inteiro. A soma do valor das atividades de cada disciplina não pode ultrapassar 100 pontos.
- O nome das atividades deve ser sucedido por X, onde X indica o código da atividade (exemplo: Atividade 1, Prova 2, Trabalho 3...)

ATIVIDADE

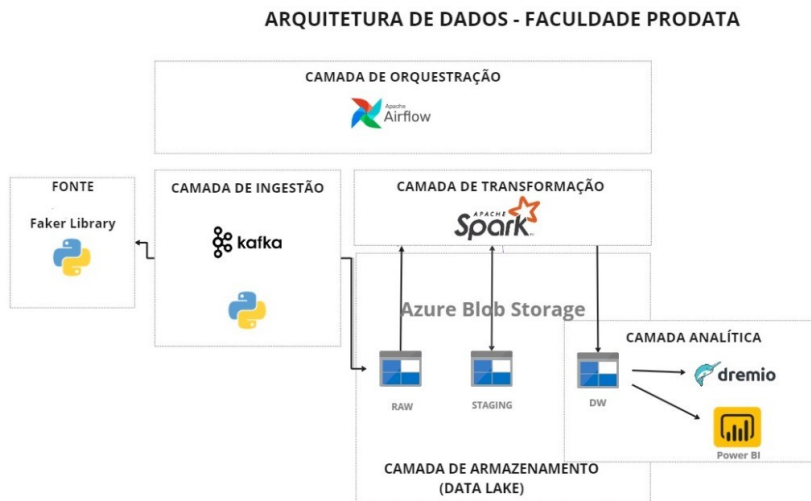
1. Desenvolva um sistema de produção de dados que gere as informações de alunos e notas a partir de 2015 até a presente data da realização desse trabalho. Esses dados devem ser gerados a cada 50 minutos. Garanta que **não haja duplicidade de dados** (utilize as camadas em tratamento para eliminar dados duplicados que podem ser levados para a camada de consumo).
2. Construa consultas que respondam as seguintes perguntas:
 - a. Total de alunos cadastrados na instituição
 - b. Total de alunos únicos na instituição
 - c. Média de aprovação de alunos por curso
 - d. Tempo médio de conclusão de curso
 - e. TOP 10 disciplinas que possuem maior índice de alunos reprovados
 - f. TOP 10 disciplinas que possuem o menor índice de alunos reprovados
 - g. Faixa etária dos alunos que possuem mais notas altas e baixas
 - i. Para gerar as faixas etárias, considere o range a seguir:
 - ii. 18 a 25 anos (pré-adulto)
 - iii. 26 a 30 anos (jovem adulto I)
 - iv. 31 a 35 anos (jovem adulto II)
 - v. 36 a 40 anos (adulto médio I)
 - vi. 41 a 50 anos (adulto médio II)
 - vii. 50 a 60 anos (adulto médio III)
 - viii. Acima de 61 anos (adulto idoso)
 - h. Quais são os 10 melhores alunos de todos os tempos
 - i. Quantidade de alunos reprovados por frequência
 - j. Quantidade de alunos reprovados por nota
 - k. Curso que possui o maior índice de aprovação
 - l. Curso que possui o menor índice de aprovação
 - m. Quais são as disciplinas que possuem o maior percentual de atividades com notas acima da média
 - n. Quais são as disciplinas que possuem o menor percentual de atividades com notas acima da média
 - o. Quais são as disciplinas que possuem o maior percentual de atividades com notas abaixo da média

- p. Quais são as disciplinas que possuem o menor percentual de atividades com notas abaixo da média


















O QUE DEVE SER ENTREGUE?

Deverá ser construído um ambiente analítico contendo:

- Desenho da arquitetura do ambiente



- Scripts de automatização padronizados (com tags que identifiquem o ambiente, grupo e escopo da atividade)
- Data lake para armazenamento de dados (o nome deve ser **datalake-dados-faculdade**)
 - Camada de dados brutos (**raw**). **Nessa camada, os dados devem ser salvos em formatos .json, .xml ou .csv.** Separe-os de acordo com a lógica de negócio do ambiente.
 - Camada de dados em tratamento (**staging**). **Nessa camada, os dados devem ser salvos em formato .parquet ou .delta.** Separe-os de acordo com a lógica de negócio do ambiente.
 - Camada de dados para consumo (**dw**). Nessa camada, os dados devem ser salvos em formato .parquet ou .delta. Separe-os de acordo com a lógica de negócio do ambiente.
- Mapeamento das fontes de dados na plataforma DREMIO

- ✓  stgaccount1096500
 - >  datalake-1096500
- ✓  datalake-dados-faculdade
 - ✓  consume
 - ✓  dados_faculdade
 - >  dados_alunos
 - >  dados_alunos_curso_disciplinas
 - >  dados_atividades_alunos
 - >  dados_cursos
 - >  dados_disciplinas
 - ✓  raw
 - ✓  dados_faculdade
 - >  dados_alunos
 - >  dados_alunos_curso_disciplinas
 - >  dados_atividades_alunos
 - >  dados_cursos
 - >  dados_disciplinas

- Scripts de ingestão de dados
caminho do arquivo: Scripts/Airflow/Ingestao/
- Dags no Apache Airflow
caminho do arquivo: Scripts/Airflow/Dags
- Produtor/consumidor de dados no Apache Kafka
caminho do arquivo: Scripts/Kafka/
- Todos os scripts SQL solicitados
arquivo: scripts_consultas.sql
- Apresentação numérica e gráfica dos resultados (pode ser Excel ou PowerBI)
arquivos: consultas_resultados/