Instituto Federal de Educação, Ciência e Tecnologia

Câmpus Campinas

Professores: Bianca Pedrosa
Samuel Martins

Projeto Interdisciplinar

1. Especificação

O objetivo do projeto interdisciplinar é a aplicação das técnicas de Ciência de Dados aprendidas nas diferentes disciplinas do semestre, a partir de um problema apoiado por um conjunto de dados (dataset). O projeto requer o desenvolvimento de uma solução completa (ponta-a-ponta) de dados usando ferramentas da AWS. A primeira etapa da solução envolve tarefas de Engenharia de Dados (ETL/pipeline de dados) enquanto a segunda etapa foca no treinamento e avaliação de modelos de Machine Learning.

2. Regras Gerais

- O tema do projeto é de livre escolha e deve ser levantado pelos os alunos integrantes do grupo.
- O projeto deverá ser feito em **grupos de até 3 (três) pessoas**.
- Ao final do semestre, os alunos devem elaborar uma apresentação do projeto, mostrando os resultados das respectivas atividades requeridas por cada uma das disciplinas cursadas.
- Todos os grupos devem apresentar o trabalho. A entrega pelo Moodle é só uma formalidade e não garante avaliação. Os alunos que não participarem da apresentação, a princípio, terão nota 0.
- Cada professor exigirá atividades relativas a conteúdos específicos de sua disciplina.
- Uma parte da nota final será definida em conjunto pelos professores. A parte restante será avaliada separadamente para cada disciplina, conforme os critérios de avaliação apresentados neste documento.

3. Apresentação

- Cada grupo terá no máximo 15 minutos para apresentar o trabalho.
- Todos os membros do grupo deverão estar presentes na data de sua apresentação.
- Todos os grupos deverão estar presentes nos dias de apresentação.
- A banca de avaliação, formada pelos professores das disciplinas, poderá arguir o grupo ou diretamente qualquer membro do grupo.
- A banca é soberana e responsável por resolver os casos previstos nestas orientações e definir os casos omissos.
- As datas das apresentações serão definidas durante o semestre juntamente com professores e alunos de todas as disciplinas.

Data final de entrega: 19/11/2023
Data da apresentação: 21 e 23/11/2023

4. Entregáveis

Link do **repositório no github** (ou sistema similar) contendo:

- Relatório, feito no README do repositório, que descreva o problema abordado, soluções, resultados, discussões e trabalhos futuros.
- Códigos desenvolvidos (p. ex., notebooks)
- Apresentação (slides)
- Todos os demais artefatos produzidos (*scripts, imagens de arquiteturas, workflows*).
- **Este material será avaliado e contará parte da nota final.**

Os professores disponibilizarão um exemplo de repositório que deverá ser usado para o desenvolvimento do projeto.

5. Critérios de Avaliação

Por ser um projeto interdisciplinar, parte da nota será definida em comum acordo pela banca definida, sendo a outra parte avaliada especificamente para cada disciplina. Os critérios são apresentados a seguir:

a. Etapa 1: Engenharia de Dados [6 pontos na disciplina D2TEC]

- Arquitetura do ETL **[1 ponto]**
    - Diagrama da arquitetura completa AWS utilizada, com explicação detalhada do pipeline de dados

- Metadados (descrição dos dados, colunas, tabelas, etc) **[1 ponto]**
    - Quais são os principais atributos e seus tipos? O que eles representam?

- Infraestrutura: **[1 ponto]**
    - Região da nuvem, recursos (número e configuração das máquinas), permissões.

- Scripts **[3.0 pontos]**
    - Do start up e clean up da arquitetura de nuvem usada, bem como o deploy do endpoint.

    - Os scripts de 10 consultas SQL para análise de dados. Esses scripts devem incluir comandos de consulta úteis em análise de dados, tais como:
        - Junções e suas variações (join, left join, right join)
        - Agregações (Group by, having, max, min, avg, sum)
        - funcões analíticas (partition, rank, etc)

  Obs: Apresentar as consultas SQL mais relevantes do trabalho. Não repetir comandos, por exemplo, não pode usar as mesmas funções e operadores e só mudar o nome das tabelas. Consultas simples usadas apenas para apresentar uma amostra dos dados não são consideradas na contagem.

b. Etapa 2: Machine Learning [6 pontos na disciplina D2APR]

- Análise Exploratória de Dados: **[1 ponto]**
    - Como as principais variáveis se distribuem?
    - Correlação de variáveis;is
    - Discussão dos principais achados da análise exploratória de dados;
- Limpeza e preparação da base de dados: **[1.5 pontos]**
    - Exemplos:
        - Remoção de duplicidade e/ou outliers;
        - Preenchimento de dados faltantes;
        - Feature scaling;
        - Class imbalance; etc
        - Discussão sucinta sobre a razão de cada etapa de limpeza e pré-processamento considerada;
- Treinamento e Validação de modelos: **[2.5 pontos]**
    - Comparar ao menos 3 algoritmos de classificação diferentes;

    - Cross-Validation;
        - Métricas consideradas para o problema;
        - Discussão dos resultados;
        - Há overfitting ou underfitting?
    - Fine-tunning
    - Avaliação no conjunto de teste:
    - Avaliar os melhores modelos no conjunto de teste;
- Discussão dos resultados. **[1 ponto]**
    - Trabalhos Futuros:
    - Discussão sobre estratégias/ideias/sugestões para a melhoria dos modelos;

c. Entregáveis e Apresentação [4 pontos nas duas disciplina]

- Artefatos entregados: **[2 pontos]**
    - Todos os artefatos serão incluídos no github. O github tem que estar bem organizado em pastas contendo notebooks e dados. A documentação do github, README, deve conter uma introdução com a descrição do problema e uma visão geral da solução, além de descrição dos dados, diagrama da arquitetura em nuvem e todos os outros itens solicitados na ETAPA 1. Os itens da ETAPA2 devem ser reunidos em notebooks, que também estarão disponíveis no Github.
    - Apresentação oral e slides: **[1 ponto]**
- Arguição: **[1 ponto]**

6. Submissão

- **APENAS UM MEMBRO DO GRUPO** precisará enviar o link para o projeto.
- A submissão do projeto será feita via atividade específica no Moodle.

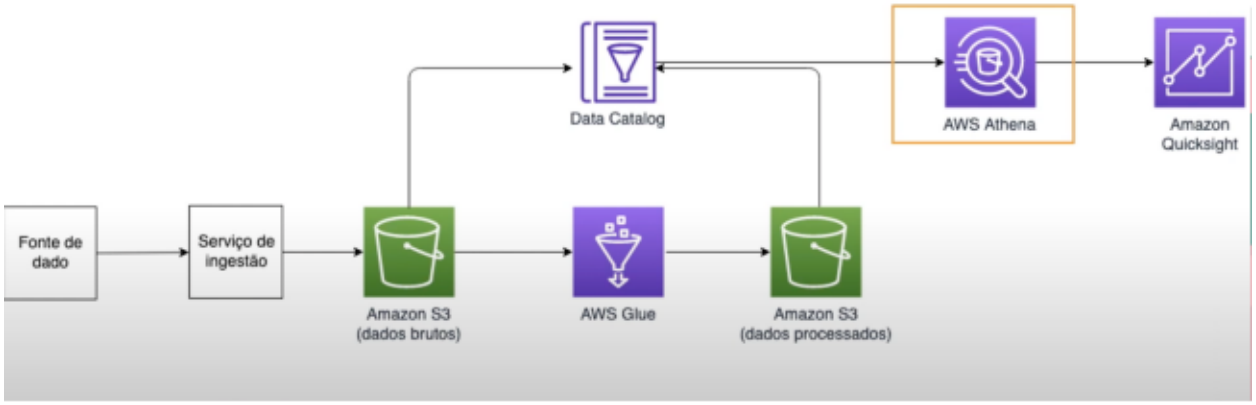
7. Referências

A seguir, disponibilizamos alguns recursos para apoiar o desenvolvimento do trabalho.

1. Laboratórios do curso AWS Data engineering.
2. Demos da série AWS LATAM transformando dados em insights
3. Um exemplo do Kaggle que apresenta uma **boa documentação e fundamentação do SQL**: https://www.kaggle.com/biancapedrosa/data-analysis-using-sql
4. Exemplos de trabalhos de ex-alunos, hospedados no GitHub:
    - Trabalho 1 (NBAStats)

- Trabalho 2 (fórmula 1)
- Trabalho 3 (crédito residencial)
- Trabalho 4 (diabetes)
- Trabalho 5 (fake news)

5. Exemplo de Diagrama de infraestrutura/Worflow de trabalho



No site visual paradigm tem templates para diagramas AWS