

Escola Superior de Ciência e Tecnologia EI - 2022/2023

Processamento de Informação – 1º momento de avaliação



Alunos:

Miguel Magalhães (2021103166) Gonçalo Silva (2020101055) Gonçalo Pereira (2020101880)

Docente: Prof. José Monteiro

1 de Dezembro de 2023



ÍNDICE DE CONTEÚDOS

Índice (de Conteúdos	2
Índice o	de Ilustrações	3
Resumo	o (problema, objetivos, dificuldades)	4
Introdu	ıção	5
Materia	ais e recursos (linguagem utilizada e a base de dados)	6
Lingu	agem Utilizada:	6
Base o	de Dados:	6
Manip	pulação de Dados:	6
Forma	ato de Saída:	6
Análise	· · · · · · · · · · · · · · · · · · ·	7
Extraç	ção (LINHAS 51 ATE 79 NO CODIGO)	8
1.	Leitura do Ficheiro Excel:	8
2.	Limpeza dos Nomes das Colunas:	8
3.	Escrita no Ficheiro XML:	8
4.	Mensagem de Confirmação:	8
Arma	zenamento (LINHAS 81 ATE 102 NO CODIGO)	10
1.	Configuração do Banco de Dados:	10
2.	Estabelecimento da Conexão com o Banco de Dados:	10
3.	Inserção dos Dados na Tabela do Banco de Dados:	10
4.	Commit e Fechamento da Conexão:	10
Testes		11
Passo	1: verificar a funcionalidade de um menu muito simplificado	11
Passo	2: Escolha do arquivo desejado	11
Passo	3: Escolha da folha do qual deseja ser extraída	11
Passo	4: Observar resultados	12
Conclu	são	13
Bibliog	rafia	14



ÍNDICE DE ILUSTRAÇÕES

Figura 1.: Extração.	9
Figura 2.:Armazenamento	10
Figura 3.: Passo 1: Verificar a funcionalidade de um menu muito simplificado	11
Figura 4.: Passo 2: Escolha do arquivo desejado.	11
Figura 5.: Passo 3: Escolha da folha do qual deseja ser extraída	11
Figura 6.: Output no XML	12
Figura 7.: Output no phpmyadmin	12



RESUMO (PROBLEMA, OBJETIVOS, DIFICULDADES)

O projeto visa superar desafios na gestão eficaz de dados heterogêneos, focando em ficheiros MS Excel. O objetivo é desenvolver um modelo robusto que extraia dados automaticamente, adapte-se a variações nas estruturas e formatos, e transfira esses dados para um Sistema de Gerenciamento de Banco de Dados (SGBD) de maneira eficiente. As dificuldades incluem lidar com variações nas estruturas dos ficheiros, normalização de dados e coordenação entre membros da equipe. A abordagem colaborativa entre 2 a 3 estudantes é fundamental para integrar habilidades diversas. O projeto também prioriza a implementação de testes de integridade para garantir a confiabilidade da base de dados resultante. O sucesso do projeto não só resolverá desafios técnicos específicos, mas também contribuirá para estratégias eficazes na gestão de dados heterogêneos em ambientes acadêmicos e profissionais.



Introdução

A administração eficaz de dados provenientes de diversas fontes, notadamente ficheiros MS Excel, representa um desafio premente na contemporaneidade. A complexidade decorre da diversidade de formatos, estruturas e tipos de dados, criando barreiras substanciais para a integração eficiente dessas informações em um Sistema de Gerenciamento de Banco de Dados (SGBD). Este projeto é uma resposta direta à necessidade de desenvolver um modelo capaz de extrair dados de maneira automática, facilitando a transição harmoniosa para o ambiente do SGBD.

No centro desse desafio encontra-se a variação nas estruturas dos ficheiros Excel ao longo do tempo, incorporando mudanças nos campos, folhas e formatos. Adicionalmente, a normalização de dados emerge como uma dificuldade crucial, especialmente quando se trata de campos sujeitos a alterações de nome ao longo dos anos. A habilidade de lidar eficazmente com essas complexidades é imperativa para garantir a consistência e a integridade dos dados ao serem transferidos para o SGBD.

A primeira dificuldade reside na adaptação à heterogeneidade nas estruturas dos ficheiros Excel, uma tarefa que requer consideração cuidadosa das mudanças nos campos, folhas e formatos ao longo do tempo. Em seguida, a normalização de dados apresenta-se como um desafio intricado, demandando a identificação e resolução de questões relacionadas a campos que experimentam alterações de nome ao longo dos anos. Além disso, a coordenação efetiva entre 2 a 3 estudantes, integrando suas habilidades diversas, é uma faceta crítica do projeto. Por fim, a elaboração e implementação de testes robustos de integridade tornam-se essenciais para assegurar que os dados na base final do SGBD permaneçam íntegros e fidedignos.

Este projeto visa abordar de forma abrangente essas dificuldades por meio da colaboração entre membros da equipe, uma análise minuciosa das fontes de dados e a implementação cuidadosa de um modelo automatizado para a transferência de dados ao SGBD. Ao enfrentar esses desafios, não apenas buscamos resolver questões técnicas específicas, mas também aspiramos contribuir para uma compreensão mais ampla de estratégias eficazes na gestão de dados heterogêneos, tanto em ambientes acadêmicos quanto profissionais.



MATERIAIS E RECURSOS (LINGUAGEM UTILIZADA E A BASE DE DADOS)

LINGUAGEM UTILIZADA:

• Python: O script foi escrito principalmente em Python para manipulação de dados, leitura de ficheiros Excel, limpeza de dados, escrita em XML, interação com o sistema operacional (via os para manipulação de diretórios) e interação com banco de dados (usando a biblioteca mysql.connector para MySQL).

BASE DE DADOS:

• MySQL: O código está configurado para interagir com um banco de dados MySQL. Ele se conecta a um banco de dados local (localhost) usando as credenciais fornecidas ('root' como usuário). O nome do banco de dados é especificado como 'dcw'. Além disso, há um código SQL para inserir dados do DataFrame do Pandas no banco de dados.

MANIPULAÇÃO DE DADOS:

- Pandas: A biblioteca Pandas é utilizada para carregar os dados do ficheiro Excel em um DataFrame. Ela também é usada para limpar os nomes das colunas, substituindo espaços por sublinhados e removendo caracteres especiais.
- Manipulação de Ficheiros: A biblioteca os é usada para listar ficheiros em um diretório e manipular caminhos de ficheiros.

FORMATO DE SAÍDA:

• XML: O código está configurado para escrever dados em um ficheiro XML. Utiliza a função to_xml do Pandas para criar um ficheiro XML a partir do DataFrame.

Essencialmente, o código lida com a manipulação de dados em ficheiros Excel, realiza limpeza e transformações nesses dados utilizando Pandas, e então, opcionalmente, armazena esses dados limpos tanto em um ficheiro XML quanto em um banco de dados MySQL.



ANÁLISE

O objetivo era identificar possíveis relações entre os conjuntos de dados, áreas de melhoria e, de forma geral, determinar a melhor abordagem para integrá-los numa base de dados e administrar os dados por meio de um sistema de base de dados. Isso envolveria garantir a integridade dos dados e realizar normalização para evitar redundâncias.

O conteúdo dos ficheiros na pasta de conjuntos de dados, observamos que todos os ficheiros são em formato Excel, a maioria deles relacionados ao "Inquérito ao Emprego no Ensino Superior Público" (IEESP) e ao inquérito estatístico "Registo Biográfico de Docentes do Ensino Superior" (Rebides). Todos esses ficheiros têm um nome geral, que é "Lista Nominativa de Docentes do Ensino Superior".

Ao explorar os ficheiros da "Lista Nominativa de Docentes do Ensino Superior", notamos que, nos anos letivos de 2003/04 a 2015/16, os documentos referem-se exclusivamente à lista Rebides, ou seja, uma lista de docentes do ensino superior em instituições privadas (inquérito Rebides). No entanto, a partir do ano letivo 2016/17 até 2018/19, todos os dados foram incluídos em uma única lista, que manteve os dados do inquérito Rebides referentes a esse ano, mas também incluiu o inquérito IEESP. Isso significa que tanto os docentes do ensino superior em instituições públicas quanto privadas estão presentes, mas organizados em folhas separadas.

De 2003/04 a 2015/16:

Ficheiro Excel com 3 folhas: capa, carreira Rebides e habilitações Rebides.

De 2016/17 a 2018/19:

Ficheiro Excel com 8 folhas: capa, categoria (ou carreira) IEESP, habilitações IEESP, habilitações aeestada, nota metodológica IEESP, categoria (ou carreira) Rebides, habilitações Rebides e nota metodológica Rebides.

É importante que certos campos nas folhas mudaram de nome ao longo dos anos, indicando uma atualização na forma como esses dados foram interpretados para se adaptar a um sistema mais recente. Também notamos que alguns campos podem conter valores nulos em registos específicos.

É crucial identificar esses aspetos para encontrar uma solução viável para a normalização da base de dados, garantindo que não haja redundâncias, campos desnecessários ou tabelas em excesso. Devemos determinar o que pode ser consolidado entre as diferentes páginas e ficheiros de todos os anos letivos fornecidos.



Existem 4 ficheiros Excel relacionados ao "Observatório do Emprego Científico e Docente" para os anos de 2019, 2020, 2021 e 2022. Esses ficheiros consistem em 5 páginas, que incluem a capa, contrato, habilitações mais elevadas, atividades e nota metodológica. Embora haja semelhanças com os ficheiros mencionados anteriormente, eles parecem conter dados de tipos diferentes em alguns campos.

EXTRAÇÃO (LINHAS 51 ATE 79 NO CODIGO)

Vou fornecer uma explicação detalhada para as linhas de código dentro da função "Ler_escrever". Essa função realiza a leitura de um ficheiro Excel e escreve os dados em um ficheiro XML. Aqui está uma explicação passo a passo:

1. Leitura do Ficheiro Excel:

A função utiliza a biblioteca pandas para ler o conteúdo do ficheiro Excel especificado pelo caminho input_file e nome da planilha sheet_name. O DataFrame resultante é armazenado na variável df.

2. LIMPEZA DOS NOMES DAS COLUNAS:

As linhas seguintes realizam a substituição de espaços e caracteres especiais nos nomes das colunas do DataFrame. Isso é feito para garantir consistência e evitar problemas com caracteres especiais em nomes de colunas.

3. ESCRITA NO FICHEIRO XML:

A função utiliza o método to_xml do DataFrame para converter os dados em formato XML. O ficheiro XML resultante é salvo no caminho especificado pela variável output. Os argumentos root_name e row_name definem o nome do elemento raiz e o nome dos elementos de linha no ficheiro XML.

4. MENSAGEM DE CONFIRMAÇÃO:

Uma mensagem é exibida indicando que todos os dados foram salvos no ficheiro XML.

Esta função realiza a leitura do ficheiro Excel, realiza algumas operações de limpeza nos dados e, em seguida, escreve os dados no formato XML.



```
# Funcão para ler um ficheiro Excel e escrevê-lo para um ficheiro XML

def Len_escrever(input_file, sheet_name): # Add search_column as a parameter

# Carregar a folha do ficheiro Excel

df = pd.read_excel(input_file, sheet_name, engine='openpyxl')

# Substituir espaços e caracteres especiais nos nomes das colunas - Gabriel Fernando

df.columns = df.columns.str.replace('\', '\')

df.columns = df.co
```

Figura 1.: Extração.



ARMAZENAMENTO (LINHAS 81 ATE 102 NO CODIGO)

No nosso código, as linhas responsáveis pelo armazenamento dos dados no banco de dados MySQL estão dentro da função Ler_escrever, especificamente no trecho onde ocorre a conexão com o banco de dados e a inserção dos dados nas respetivas colunas da tabela habilitacoesnaofinal.

Explicação passo a passo:

1. Configuração do Banco de Dados:

Antes da função Ler_escrever, há a configuração do banco de dados (db_config) que contém informações como o host, usuário e nome do banco de dados.

2. ESTABELECIMENTO DA CONEXÃO COM O BANCO DE DADOS:

As linhas conn = mysql.connector.connect(**db_config) e cursor = conn.cursor() são responsáveis por estabelecer a conexão com o banco de dados.

3. Inserção dos Dados na Tabela do Banco de Dados:

O loop for index, Row in df.iterrows(): itera sobre as linhas do DataFrame df e insere os dados correspondentes nas colunas da tabela habilitacoesnaofinal do banco de dados.

4. COMMIT E FECHAMENTO DA CONEXÃO:

Após a inserção dos dados, conn.commit() efetua o commit para persistir as mudanças no banco de dados, e conn.close() fecha a conexão.

Esta parte do código é responsável por integrar os dados extraídos do ficheiro Excel no DataFrame (df) à tabela habilitacoesnaofinal no banco de dados MySQL.

```
# Configuração da base de dados

db_config = {
    'nost: 'localhost',
    'user': 'root',
    'database': 'dcw'

# Conn = mysql.connector.connect("db_config)

cursor = conn.cursor()

print("connection made 1")

for index, Row in df.iterrows():

dados = [Row.get(col) if pd.notna(Row.get(col)) else None for col in [ROW1, ROW2, ROW3, ROW4, ROW5, ROW6, ROW7, ROW8, ROW10, ROW11, ROW

query = "INSERT INTO habilitacoesnaofinal (IDDOcente, Nome_docente, Estab, Estab_Nome, UO, UO_Nome1, Grau_nome, CodigoCurso, NomeCurso, Outro

print("connection made 2")

valores = tuple(dados)

print("connection made 3")

cursor.execute(query, valores)

print("connection made 4")

conn.commit()

print("connection so commited")

conn.close()

print("connection closed")
```

Figura 2.:Armazenamento



TESTES

Ao decorrer da resolução de todo o código, foram efetuados diversos testes par comprovar a sua funcionalidade. Os testes mais frequentados irão ser representados nas seguintes figuras e passos:

PASSO 1: VERIFICAR A FUNCIONALIDADE DE UM MENU MUITO SIMPLIFICADO

Figura 3.: Passo 1: Verificar a funcionalidade de um menu muito simplificado.

PASSO 2: ESCOLHA DO ARQUIVO DESEJADO

```
File Edit Selection View Go Run Terminal Help 

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

11. DGEEC DSEE_DEES_ListalECDES_IEESP_2022.xlsx

12. ListasPublicasRebides_RB8904.xlsx

13. ListasPublicasRebides_RB8906.xlsx

14. ListasPublicasRebides_RB8906.xlsx

15. ListasPublicasRebides_RB8906.xlsx

16. ListasPublicasRebides_RB890708.xlsx

17. ListasPublicasRebides_RB8909.xlsx

18. ListasPublicasRebides_RB89010.xlsx

19. ListasPublicasRebides_RB891112.xlsx

20. ListasPublicasRebides_RB81112.xlsx

21. ListasPublicasRebides_RB1112.xlsx

22. ListasPublicasRebides_RB1123.xlsx

23. ListasPublicasRebides_RB1123.xlsx

24. ListasPublicasRebides_RB1123.xlsx

25. Carreinatividades1415

36. Habilitacoes1415

37. Lapa

28. Carreinatividades1415

38. Habilitacoes1415

39. Habilitacoes1415

30. Habilitacoes1415

30. Habilitacoes1415

30. Habilitacoes1415
```

Figura 4.: Passo 2: Escolha do arquivo desejado.

PASSO 3: ESCOLHA DA FOLHA DO QUAL DESEJA SER EXTRAÍDA

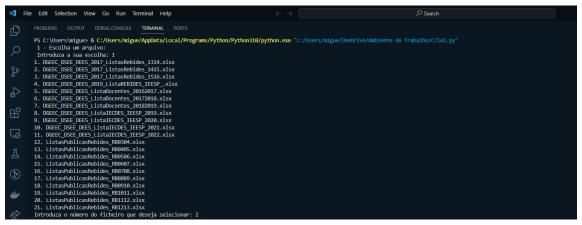


Figura 5.: Passo 3: Escolha da folha do qual deseja ser extraída.



PASSO 4: OBSERVAR RESULTADOS

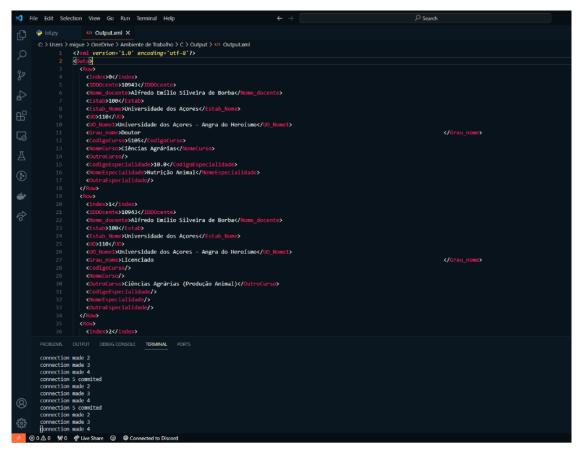


Figura 6.: Output no XML.

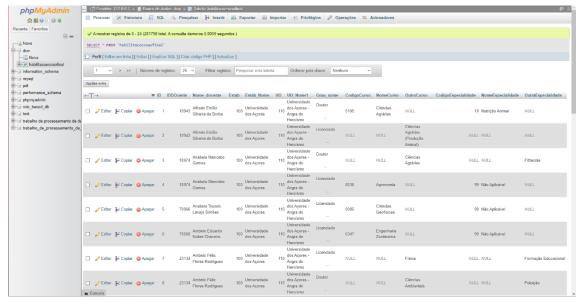


Figura 7.: Output no phpmyadmin.



CONCLUSÃO

Ao finalizar este projeto, teremos desenvolvido uma solução inovadora para extração e transferência automática de dados de ficheiros MS Excel para um SGBD. A análise aprofundada das fontes de dados permitiu a criação de um modelo capaz de lidar com a complexidade inerente à diversidade dessas fontes.

O processo de desenvolvimento do modelo não apenas destacou as habilidades técnicas dos membros da equipe, mas também demonstrou a importância da colaboração em projetos desafiadores. A implementação efetiva do modelo, seguida por testes de integridade de dados, assegurará que a base de dados resultante seja confiável e pronta para processamento de informações.

A entrega da documentação técnica não apenas cumpre um requisito obrigatório, mas também serve como um valioso recurso para futuros desenvolvimentos e manutenções. A apresentação final será uma oportunidade para compartilhar insights, desafios superados e resultados alcançados, consolidando o aprendizado adquirido ao longo deste projeto colaborativo.



BIBLIOGRAFIA

Como conectar Python AO banco de dados MySQL fácil. (2021, April 18). DebugEverything. https://blog.debugeverything.com/pt/como-python-ao-banco-de-dados-mysql

Jonatas Silva do Espirito Santo. (2021, March 21). *Python: Manipulando dados armazenados Em bancos MySQL. Medium.* https://jonates.medium.com/python-manipulando-dados-armazenados-em-bancos-mysql-45ba477b19d1

Python tutorial. (2010, September 5). W3Schools Online Web Tutorials. https://www.w3schools.com/python/

Web scraping com Python: Tutorial com 3 Exemplos Práticos. (2023, October 5). Hashtag Treinamentos | Cursos de Excel, VBA, Python e mais!. https://www.hashtagtreinamentos.com/web-scraping-python

(2022, March 20). YouTube. https://www.youtube.com/watch?v=iZ8sTC2tG-w