Materiais:

Q: Onde encontro os materiais de apoio da disciplina?

R: O material de apoio consta no ícone de folha a4 ao lado do título da disciplina.

Ferramentas:

Q: Quais as vantagens de usar ferramentas pagas de ETL, podendo desenvolver os fluxos diretamente em Python? Documentação? Sustentação? Performance?

R: As ferramentas de ETL pagas permitem que pessoas que não sabem uma linguagem de programação mesmo possam desenvolver inúmeros ETLs para padronizar os seus dados. As ferramentas também permitem que consigamos automatizar grande parte desse processo como o que você mesmo indicou que faz, como conexões de banco, *requests* e o próprio agendamento. Como as ferramentas são construídas para esse propósito também é natural enxergarmos um ganho de performance. Além disso, como pode não ser uma linguagem de programação, é possível ter uma curva de aprendizagem mais fácil permitindo que mais pessoas possam dar sustentação.

Aqui tem um documento bem legal que traz uma discussão exatamente como a lançada: https://gustavomaiaaguiar.wordpress.com/2010/05/10/por-que-utilizar-uma-ferramenta-de-etl/

Traz alguns prós de *hard code* como o que você mantém e prós da utilização de ferramentas de ETL.

Conceitos e Exercícios:

Q: Pude observar aos 46min15s da parte 1 da aula que o arquivo itens_pedidos.csv ocupa 9 MB e o itens_pedidos.xlsx ocupa 4,5 MB. Me chamou a atenção que o arquivo excel ocupa metade do arquivo csv para os mesmos dados, achei curioso! Isto é um comportamento padrão? Poderia por gentileza explicar como pode esta diferença tão significativa?

R: Em termos gerais a diferença de tamanho entre esses dois tipos de arquivo é que o Excel (XLSX) sempre vai compactar o arquivo de forma interna enquanto arquivos CSV não são compactados.

Você verá arquivos XLSX menores se houver muitos dados repetidos. Na verdade, o XLSX extrai todos os valores de texto, os armazena em uma tabela de pesquisa e, em seguida, os substitui por um número de referência menor na tabela de pesquisa. Isso significa que ele só precisa usar espaço uma vez por *string* de texto (um pouco mais de "uma vez" por causa das referências, mas ainda assim muito menos espaço).

Um arquivo CSV listará todas as ocorrências por completo, o que ocupa muito mais espaço. Eu vi arquivos aumentarem para 10 vezes o tamanho depois de salvar em CSV. A única vez que você deve esperar uma economia com um CSV é se houver muita formatação e não muito texto reutilizado. Em seguida, o CSV remove a formatação e o produto é menor.

Q: Dentro da Ciência de Dados, a área de Arquitetura e Gerência de Bancos de Dados visa principalmente...

R: É consenso na literatura que a área de Ciência de Dados abrange amplo espectro de aplicações, reunindo sob si inúmeras outras áreas mais específicas. O conhecimento dessa abrangência é de fundamental importância para os profissionais que escolhem essa área, visto que, na maior parte dos casos, as estratégias para implementação de projetos relacionados à Ciência de Dados devem contemplar atividades em diversas dessas sub-áreas, como é o caso da Arquitetura e Gerência de Bancos de Dados.

**FAQ gerado com base em comentários até o dia 02/03/2022.*