Doc Testes Unitários

Usando a função limpeza de dados, efetuei testes unitários considerando o caminho feliz.

Resultado: Passou.

Os testes considerando um dataframe “sujo” considerou os seguintes cenários:

1. Erro de digitação no nome da coluna;
2. Erro no formato da data ( uso de – a mais);
3. Erro no formato do dataframe ( shape diferente);
4. Número negativo no sale\_value.

Usando pandas.testing, obtivemos erros para o cenário 1, 2, 3, sendo o único cenário contemplado na função o cenário 4.

Solução para erros:

1. Implementar validação do nome das colunas. Em alguns casos, pela experiência, demonstrou-se que gerar uma função que observa o nome das colunas de maneira dinâmica é mais efetivo e evita problemas. Não é sugerido utilizar a posição, pois o autor do arquivo pode inverter e se esquecer de nos avisar.
2. Tratar o campo date. Retirar caracteres de separação e manter apenas os números evita problemas de formatação.
3. Validação do shape e avaliação se é necessário reshape.

Data Quality

Crie métricas de observabilidade para o processo ETL que você desenvolveu anteriormente(não é necessário implementação):

Monitore o tempo que leva para os dados serem extraídos, transformados e carregados. –

*Implementado*

Implemente alertas para qualquer falha ou anomalia durante o processo ETL.

*Usar lib logging, tratar exceções e enviar para ferramenta de monitoramento como grafana ou splunk*

Descreva como você rastrearia um problema no pipeline, desde o alerta até a fonte do problema.

1. *Identificar logs da exceção para descobrir qual o tipo de erro, qual a função e arquivo onde está o problema.*
2. *Rastrear os dados que foram utilizados e geraram aquele cenário. Comparar os dados com os dados esperados.*
3. *Realizar testes com o cenário do caminho feliz para identificar se houve mudança no código e ninguém avisou.*
4. *Verificar se o problema ocorre novamente com determinado cenário, como horário, tipo de arquivo específico, nome do arquivo.*