Tópicos em Engenharia de Software: Relatório Técnico

Edson Júnior Campolina Silva

https://github.com/EdsonCampolina/TES_Pratica4

Introdução

Nessa atividade foi gerada uma base de dados no arquivo users.json utilizando o gerador.py, que gerava usuário utilizando a biblioteca *Faker*. A partir dessa base de dados foi necessário consertar a primeira função que possuía valores incorretos. Após consertar essa função foi necessário criar uma nova função para calcular a média de idades. Testes em big data são cruciais para identificar problemas de qualidade, integridade e desempenho, assegurando a confiabilidade dos resultados e garantindo que as operações de armazenamento e análise de dados sejam precisas e eficazes.

1 Desenvolvimento de Testes Unitários

1.1 Função avgAgeCountry

Para essa função foi calculada a idade média para cada país e foi testado a idade média de cada um desses países. Além disso, também foi testada a idade média geral. Isso foi testado tanto em meses quanto em anos

- Arquivo JSON vazio.
- Valores de idade ausentes ou nulos.
- Campo ‘country‘ ausente ou nulo.

1.2 Outras Funções de Processamento

A função extra que foi desenvolvida foi a função de calcular a idade em meses ao invés de calcular somente em anos

1.3 Função de Transformação

A função de transformação multiplicava a idade por 12 para garantir que estava em meses. Ela permitia testar outros cenários, caso a idade deixe de ser informada em anos e começe a ser informada em meses.

2 Objetivos dos Testes

Objetivos dos testes:

- Testar idade média de cada um dos países
- Testar idade média total
- Testar idade média em meses de cada um dos países
- Testar idade média em meses total
- Esses testes evitavam dados inconsistentes, como nulo, e caso alguns dos dados fossem alterado os testes iriam avisar

3 Reflexão sobre Testes em Big Data

Escrever testes para ambientes de Big Data apresenta desafios significativos, sobretudo devido à complexidade e ao volume massivo de dados processados. A necessidade de testar funções que lidam com grandes conjuntos de dados é crucial para garantir a precisão e a confiabilidade dos resultados. Ao trabalhar com PySpark, otimizar o desempenho se torna uma prioridade fundamental, pois até pequenas melhorias podem ter um impacto significativo na velocidade de processamento. Um teste de desempenho nesse contexto exigiria a definição clara de métricas de avaliação, como tempo de processamento, consumo de memória e utilização de recursos, com a finalidade de identificar gargalos e pontos de estrangulamento no sistema.

4 Conclusão

A atividade ressaltou a importância crucial dos testes em ambientes de Big Data, especialmente ao lidar com grandes volumes de dados. Além disso, enfatizou-se que os testes não apenas garantem a precisão dos resultados, mas também contribuem para a confiabilidade e a eficiência operacional do sistema de processamento de dados em larga escala.