TÉCNICAS COLOCADAS EM PRÁTICA

1. Pré-processamento de dados de texto
2. Manipulação de dados em um Data Lake
3. Consultas em um banco de dados NoSQL, no caso, MongoDB
4. Visualização utilizando-se de alguns gráficos

SOBRE O PROJETO

É fato que qualquer empresa que tenha presença online pode estar interessada em obter informações valiosas a partir dos comentários dos usuários sobre seus produtos ou serviços, seus concorrentes, o mercado ou até mesmo suas próprias preferências.

Neste projeto, teremos acesso a dados textuais que simulam comentários de usuários.

Os dados textuais são considerados não estruturados, o que torna difícil armazená-los e analisá-los por meio de um Data Warehouse.

Para resolver esse problema, utilizaremos um Data Lake, que permitirá o carregamento dos dados em seu formato bruto.

Em seguida, faremos o pré-processamento dos dados durante a análise, de forma a fornecer informações valiosas para os tomadores de decisão.

SOBRE OS DADOS UTILIZADOS

Para este projeto, utilizaremos um conjunto de dados fictícios que simulam posts de comentários de usuários em redes sociais.

Este dataset (em anexo no projeto) foi gerado no site:
https://www.mockaroo.com (link seguro)

DICIONÁRIO DE DADOS

O Dicionário de Dados é uma descrição detalhada de cada variável presente no nosso conjunto de dados. Aqui está o dicionário do dataset utilizado no Projeto 8:

1. **'_id'** - identificador do comentário (sequência de nºs e letras)
2. **'status'** - se o comentário é público ou privado
3. **'creationDate'** - data de criação do posts
4. **'allowComments'** - se permitiu comentários
5. **'title'** - título
6. **'description'** - descrição (conteúdo do post)
7. **'tags'** - tags usadas como indexador nos comentários
8. **'Category'** - categorias (politics, lifestyle, etc)
9. **'filteredPicture'** - informa se a imagem foi filtrada

Além destas variáveis, foram criadas outras para proporcionar insights e análises, dentre estas estão:

1. **'hour'** - correspondente a hora do comentário
2. **'year'** - ano
3. **'length_of_heading'** - tamanho do título do post
4. **'length_of_desc'** - tamanho da descrição

AUTOR DO CÓDIGO:
Jonatas A. Liberato
Agradecimento: DSA