[1]UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
CENTRO DE CIÊNCIAS EXATAS E NATURAIS
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO
TRABALHO DE CONCLUSÃO DE CURSO (2024)

IMPLEMENTAÇÃO DE CONSULTAS SQL EM DADOS SIMULADOS DE ATAQUES CIBERNÉTICOS UTILIZANDO DATABRICKS COMMUNITY

Rudinilly Rodrigues Nogueira[1], Paulo Henrique Lopes Silva[2]

RESUMO

A segurança dos dados tem se tornado um dos maiores desafios no mundo digital atual, à medida que o volume de informações processadas cresce exponencialmente. Com a sofisticação das ameaças cibernéticas, torna-se essencial o uso de estratégias eficazes para proteger as infraestruturas de TI e prevenir ataques. Ferramentas e técnicas de análise em bases de dados surgem como uma abordagem poderosa para identificar vulnerabilidades e antecipar ações maliciosas, permitindo uma resposta mais ágil e precisa. Pensando nisso, este trabalho tem como objetivo a implementação de consultas SQL na extração de informações em dados simulados de ataques cibernéticos, utilizando a plataforma Databricks Community. Por meio de uma análise detalhada em uma base de dados sintética obtida no Kaggle, contendo registros de tentativas de ataque, vulnerabilidades exploradas, ações de mitigação e outros indicadores de segurança, busca-se identificar os tipos de ataque mais frequentes, os navegadores e sistemas mais vulneráveis.

Palavras-chave: SQL; Databricks Community; Extração; Consultas.
[2]

1 INTRODUÇÃO

Com o avanço da digitalização, os dados tornaram-se essenciais em todas as esferas empresariais, impulsionando decisões estratégicas, inovações tecnológicas e o crescimento econômico, como dito por Humby, “Os dados são o novo óleo” (HUMBY, 2006). No entanto, essa valorização dos dados também atraiu a atenção de cibercriminosos, que veem nesse ativo uma oportunidade para realizar ataques cibernéticos cada vez mais sofisticados. Esses ataques podem variar desde tentativas de acesso não autorizado a sistemas corporativos até o sequestro de dados críticos por meio de *ransomwares*, causando prejuízos financeiros, danos à reputação e compromissos legais para as empresas. Assim, a necessidade de ferramentas e estratégias robustas para garantir a segurança e integridade dos dados tornou-se imperativa na atual esfera mundial.

O Brasil segue em primeiro lugar na América Latina como alvo de ataques hackers e segundo na esfera mundial, atrás apenas dos Estados Unidos. No segundo semestre de 2023 o Brasil sofreu 357.422 ataques, o que foi um aumento de 8,86% do primeiro semestre do ano passado. Entre os principais setores atingidos no Brasil, a telecomunicação sem fio segue sendo líder com 82.065 ataques, seguindo com transporte de cargas com 25.620 registros, e

[1]*Discente do Bacharelado em Ciência da Computação da UFERSA - campus Mossoró.*
[2]*Professor Associado do Departamento de Computação da UFERSA - campus Mossoró.*

processamento de dados, com 25.130. Esses dados são do Relatório de Inteligência de Ameaças da NetScout, que é a líder global em soluções de cibersegurança. Em comparação com os demais países da América Latina o Brasil sofreu cerca de 4,3 vezes mais ataques que a Argentina, segunda colocada, com 82.749 ataques, e cerca de 4,6 mais que o Peru, terceiro mais atacado, com 74.531 (MAIA, 2024).

Com a pandemia de covid-19 muitos colaboradores passaram a atuar seus trabalhos remotamente, ou seja, fora do perímetro de segurança da empresa. Por conta disso, muitos desses colaboradores estão sujeitos a ataques. Profissionais trabalhando isolados são alvos fáceis de atacantes maliciosos, que os manipulam para tirar vantagem por meio da engenharia social. É uma técnica empregada por criminosos virtuais para induzir usuários desavisados a enviar dados confidenciais, infectar seus computadores com *malware* ou abrir *links* para sites infectados. Com a flexibilidade proporcionada pelo trabalho remoto, muitos profissionais passaram a trabalhar em ambientes como hoteis, cafés e bares. Essa possibilidade aumenta as ameaças relacionadas à segurança digital, uma vez que os *hackers* podem explorar vulnerabilidades de dispositivos de rede, como roteadores mal configurados em lugares públicos (REDAÇÃO, 2023).

1.1 Fundamentação Teórica

Dentre os ataques mais comuns da atualidade, os que mais se destacam são *Phishing*, *Ransomware*, *DDoS* dentre outros. *Phishing* é uma forma de ataque cibernético em que o atacante tenta enganar as vítimas para que elas revelem informações pessoais sensíveis, como senhas, números de cartão de crédito ou credenciais de login. Isso é geralmente feito por meio de e-mails ou mensagens falsas que se passam por comunicações legítimas de empresas, bancos ou outras entidades confiáveis. Se bem-sucedido, o *phishing* pode levar ao roubo de identidade, perda financeira, ou acesso não autorizado a sistemas corporativos (IBERDROLA).

Ransomware é um tipo de *malware* que, uma vez instalado no sistema da vítima, criptografa os dados do usuário e exige um pagamento (resgate) para fornecer a chave de descriptografia. O termo "*ransom*" refere-se ao resgate exigido pelo criminoso para liberar os dados sequestrados, muitas vezes, o *ransomware* é distribuído por meio de anexos maliciosos em e-mails de *phishing*. Pode resultar em perda de dados, interrupção das operações de negócios, custos significativos associados ao pagamento do resgate (caso a vítima decida pagar), e gastos com recuperação de dados e sistemas (IBERDROLA).

Um ataque de negação de serviço distribuído (*DDoS*) é uma tentativa maliciosa de interromper o tráfego normal de um servidor, serviço ou rede, sobrecarregando o alvo ou sua infraestrutura circundante com uma onda massiva de tráfego da internet. Esse tráfego geralmente vem de uma rede de dispositivos infectados, chamada de *botnet*. Pode causar a indisponibilidade de serviços online, perda de receita para empresas dependentes da *web*, e danos à reputação. Em casos mais graves, pode comprometer a segurança da infraestrutura de TI (IBERDROLA).

Alguns outros que vale destacar são, *Man-in-the-Middle* (MitM), Neste ataque, o atacante intercepta e possivelmente altera a comunicação entre duas partes que acreditam estar se comunicando diretamente uma com a outra. Pode ser usado para roubar informações sensíveis como credenciais de login e números de cartão de crédito. SQL Injection consiste em um ataque onde o atacante insere ou "injeta" código SQL malicioso em um campo de

entrada para manipular o banco de dados e acessar informações que normalmente não seriam acessíveis (IBERDROLA, 2021).

Compreender a diversidade e a sofisticação dos ataques cibernéticos é essencial para a proteção das infraestruturas digitais. No entanto, para além da identificação desses ataques, a análise de dados desempenha um papel crucial na antecipação de ameaças, na detecção de padrões de comportamento malicioso, e na formulação de estratégias de defesa mais eficazes. Ao coletar, processar e analisar grandes volumes de dados, é possível extrair insights valiosos que ajudam na prevenção e resposta a incidentes de segurança. Nesse contexto, o uso de ferramentas avançadas de análise de dados, como o *Databricks Community* (CALANCA, 2023), permite explorar e interpretar esses dados de forma mais eficiente, potencializando a capacidade de identificar tendências emergentes e fortalecer a segurança cibernética (BASTOS, 2024).

Análise de dados envolve o processamento de grandes quantidades de informações que podem ser estruturadas, semiestruturadas ou não estruturadas. A capacidade de gerenciar diferentes tipos de dados, como textos, conteúdos multimídia, e formatos como XML (MICROSOFT) e JSON (JSON), é crucial para os sistemas modernos de análise. Após a extração dos dados dos bancos de dados, eles são frequentemente importados para ferramentas de visualização como Power BI (EBAC) ou Tableau (ALURA) para análise mais aprofundada. Essas ferramentas permitem identificar padrões, *insights* e tendências que são fundamentais para a tomada de decisões e para revelar informações valiosas escondidas nos dados. Armazenamento em nuvem como o *Amazon S3* e *Azure Data Lakes* oferece escalabilidade e acessibilidade, facilitando a recuperação e utilização dos dados por analistas sempre que necessário (H. S. J. P. K. N. P. 2023) .

Big Data, é como chamamos o processo de análise de grandes quantidades de dados, armazenados remotamente, o desenvolvimento de sistemas de defesa contra ataques cibernéticos baseados em *Big Data* é viável e muito vantajoso, pois não há a necessidade de arcar com custos de licenciamento de software, há também uma grande variedade de tecnologias e de detalhes de arquitetura que permitem a implementação de diferentes soluções para criar esses sistemas (ANDRADE, 2020).

Quem trabalha nessa área, é o cientista de dados, ele é responsável pelos processos e práticas relacionados à coleta, organização, análise e interpretação dos dados, o que resulta em adotar as melhores práticas de gerenciamento, definição de processos e melhores práticas que resultarão na sua proteção contra acessos, uso e modificações não autorizados (JUNQUEIRA, 2024). Essa área de trabalho vem ganhando muita relevância pois é esse profissional que possui as habilidades para lidar corretamente com os dados e por consequência definir as melhores ferramentas para identificar ameaças e proteger o sistema de dados (JUNQUEIRA, 2024).

Uma dessas ferramentas, é o *Databricks Community*, uma plataforma poderosa e acessível para análise de dados que combina as vantagens do *Apache Spark* com uma interface de usuário intuitiva, permitindo o processamento e análise de grandes volumes de dados de forma eficiente. Esta plataforma oferece uma infraestrutura em nuvem escalável, ideal para a realização de análises complexas sem a necessidade de investimentos em *hardware* (CALANCA, 2023).

1.2 Objetivo Geral

Dessa forma o problema abordado envolve a crescente sofisticação e frequência de ataques cibernéticos. Com a crescente geração de dados por parte das organizações, aumenta o desafio de mitigar ameaças em tempo hábil. Onde os ataques se aproveitam de vulnerabilidades em sistemas de rede, comprometendo dados sensíveis e infraestruturas críticas. Diante disso, são necessárias ferramentas para identificar informações em atividades maliciosas, permitindo a antecipação e prevenção dos ataques. No entanto, lidar com grandes volumes de dados não estruturados e de alta velocidade exige tecnologias avançadas como o *Databricks Communit*y, que oferece uma abordagem eficiente para o processamento e análise dessas informações.

Portanto, o objetivo deste trabalho é, através da implementação de consultas SQL, extrair informações de ataques cibernéticos de uma base de dados, como tráfego de rede mais utilizados por cada ataque, sistemas mais vulneráveis, período de ocorrência dos ataques, dentre outros.

1.3 Objetivos Específicos

Como objetivos específicos, temos:

- Obter uma base de dados em que seja possível extrair informações.
- Fazer upload dessa base de dados em plataforma de análise de dados.
- Realizar processo de preparação dos dados.
- Realizar análise exploratória utilizando consultas SQL, extraindo informações relevantes presentes nessa base de dados.

Ao realizar esses passos, será possível ter resultados que demonstram que é possível a implementação de consultas SQL em uma base de dados e que isso é uma ferramenta poderosa e que pode ajudar na segurança da web.

2 DESENVOLVIMENTO

Na seção de desenvolvimento deste trabalho, serão descritos os métodos e ferramentas utilizados para identificar padrões e tendências em dados de ataques cibernéticos. Cada subseção aborda detalhadamente o que foi feito.

2.1 Método

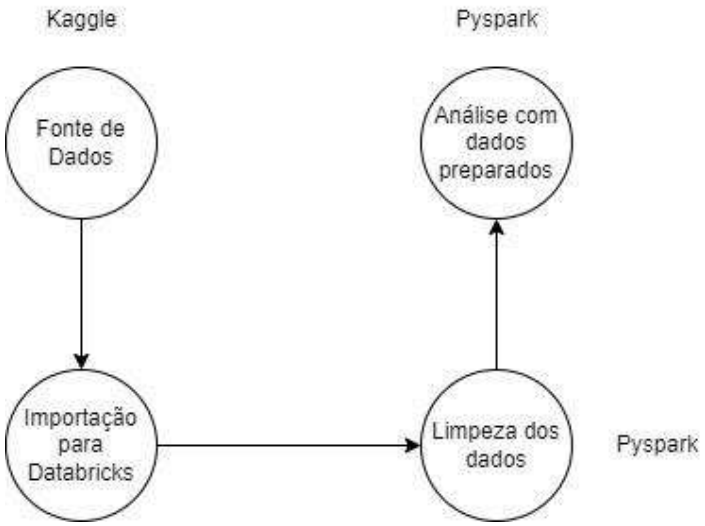
Nessa seção, os passos para a análise de dados serão apresentados; O objetivo é deixar claro quais foram os passos seguidos até a obtenção dos resultados. Ela aborda todas as etapas com uma breve descrição, com isso, o desenvolvimento foi dividido em etapas que são descritas brevemente a seguir:

1. Obtenção da base de dados: A base de dados sintética será obtida da plataforma Kaggle (CATUNDA, 2022), contendo registros detalhados de ataques cibernéticos
2. Preparação dos dados: Limpeza e transformação dos dados para assegurar a sua qualidade, utilizando a plataforma *Databricks Community*.
3. Análise exploratória: Conduzida no ambiente Databricks, utilizando consultas e visualização dos resultados para examinar informações dos ataques.

4. Uso de PySpark: A limpeza e preparação dos dados será realizada com PySpark (DATABRICKS) para garantir eficiência no processamento dos dados.
5. Resultados: Apresentação dos resultados em tabelas e gráficos que evidenciam as tendências e padrões detectados.

Conforme ilustrado na Figura 1, o processo de análise de dados inicia-se com a obtenção dos dados na plataforma Kaggle, seguido pela importação para o ambiente Databricks. Posteriormente, utiliza-se PySpark para a limpeza e preparação dos dados, permitindo sua análise de forma eficiente.

Figura 1 - Diagrama do fluxo



Fonte: Autoria própria

Com a aplicação dessas etapas de análise de dados, espera-se identificar padrões que auxiliem na antecipação e prevenção de futuros ataques cibernéticos, bem como contribuir para o aprimoramento das defesas em sistemas de segurança da informação.

2.2 Databricks

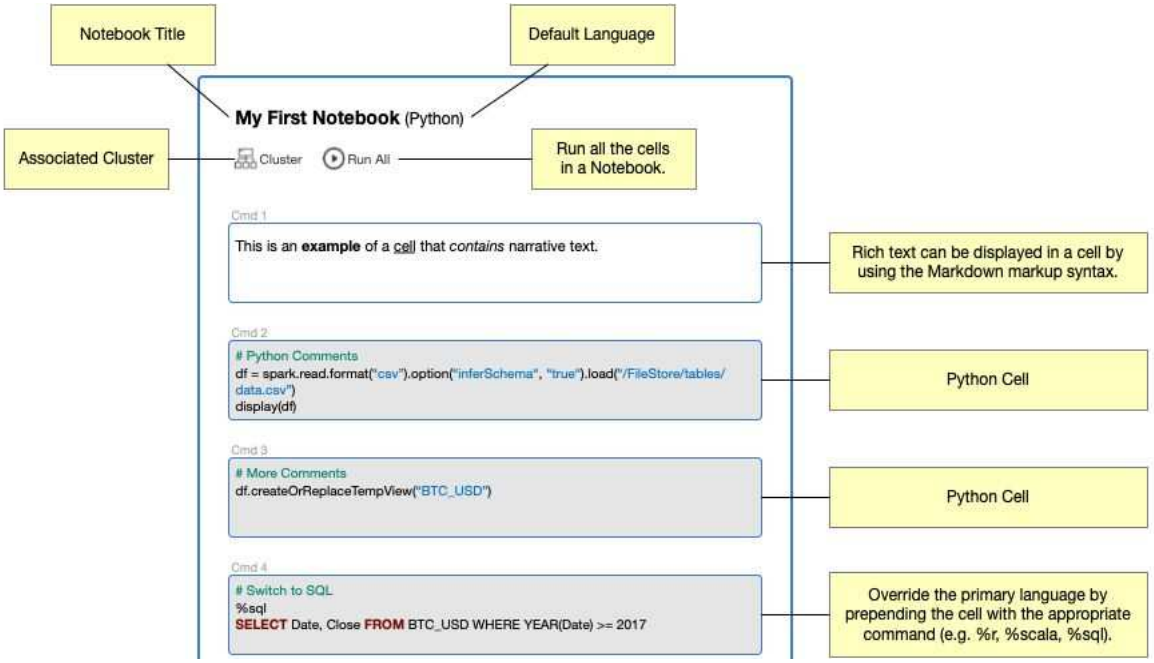
O *Databricks Community* é uma plataforma fácil de usar que oferece uma variedade de recursos para se trabalhar com dados. Entre os principais recursos do *Databricks Community* estão a capacidade de manipular dados estruturados e não estruturados, a integração com diversas fontes de dados, e o suporte a várias linguagens de programação como Python, SQL e Scala. Além disso, o Databricks facilita a criação de pipelines de dados, o desenvolvimento de modelos de *machine learning* e a visualização de dados, permitindo aos analistas explorar tendências e padrões ocultos de maneira colaborativa e eficiente. Ao utilizar os recursos oferecidos por ele, é possível acelerar o ciclo de análise, tornando-o uma ferramenta essencial para projetos de análise de dados em cibersegurança.

2.2.1 Notebooks

Os *notebooks* do *Databricks* são uma ferramenta central para criar e gerenciar fluxos de trabalho em ciência de dados e *machine learning*. Eles oferecem suporte a múltiplas linguagens de programação, incluindo Python, SQL, Scala e R, permitindo aos usuários desenvolver código, analisar dados e visualizar resultados de maneira eficiente. Uma característica destacada é a capacidade de coautoria em tempo real, facilitando a colaboração

entre equipes. Além disso, os notebooks possuem controle automático de versões, integração com repositórios *Git* e ferramentas de visualização de dados incorporadas, tornando o processo de desenvolvimento mais organizado e eficiente (DATABRICKS).

Figura 2 – Imagem explicativa de um Notebook



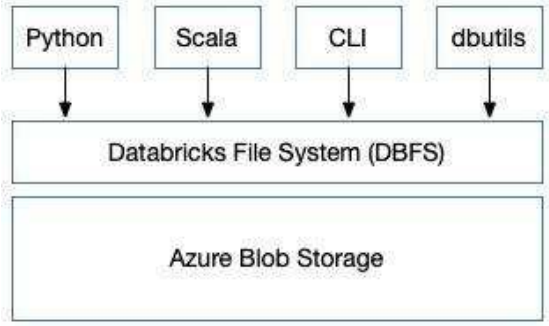
Fonte: Taygan - https://www.taygan.co/blog/2018/12/02/azure-databricks.

A Figura 2 apresenta um exemplo de notebook no *Databricks*, com várias anotações que explicam seus elementos principais. O título do notebook é "My First Notebook (Python)", e a linguagem padrão usada no *notebook* é Python, indicada ao lado do título. O cluster associado ao *notebook*, necessário para a execução dos comandos, também é exibido, assim como um botão "Run All", que permite executar todas as células de uma vez. O *notebook* contém uma célula com texto narrativo formatado com a sintaxe Markdown, além de blocos de código Python que executam operações de processamento de dados. Há também uma célula SQL, identificada pelo comando ”%sql”, que executa consultas e pode gerar visualizações de dados, conforme ilustrado no gráfico mostrado na imagem.

2.2.2 DBFS

O *Databricks File System* (DBFS) é o sistema de arquivos distribuído da plataforma *Databricks*, projetado para armazenar dados de maneira eficiente. Ele oferece uma interface de arquivos similar ao sistema de arquivos local, permitindo que os usuários leiam e gravem dados diretamente em *clusters Databricks*. O DBFS facilita a integração com serviços de armazenamento em nuvem, como *Amazon S3* e *Azure Data Lake*, permitindo o armazenamento escalável e seguro dos dados. Além disso, os dados armazenados no DBFS podem ser facilmente acessados e manipulados em *notebooks Databricks*, sendo essencial para pipelines de processamento de dados, análise e *machine learning* (DATABRICKS).

Figura 3 - Representação DBFS



Fonte: Taygan - https://www.taygan.co/blog/2018/12/02/azure-databricks

A Figura 3 demonstra que os dados podem ser acessados usando a API do Sistema de Arquivos do Databricks (Databricks File System API), API do Spark, CLI do Databricks, Utilitários do Databricks (dbutils) ou APIs de arquivos locais.

2.2.3 Pyspark.pandas

O pyspark.pandas é uma API dentro do *PySpark* que permite aos usuários trabalhar com *DataFrames* de uma maneira muito semelhante à biblioteca pandas do Python, mas com a escalabilidade e desempenho do Spark. Ele foi desenvolvido para facilitar a transição entre pandas, amplamente utilizado para análise de dados em escala menor, e *PySpark*, que é mais eficiente ao lidar com grandes volumes de dados distribuídos. O pyspark.pandas oferece funções e operações familiares do pandas, como manipulação de dados, agregações e transformações, mas permite que essas operações sejam realizadas de maneira distribuída em *clusters Spark*, o que é útil para grandes conjuntos de dados (DATABRICKS, MICROSOFT). O uso da biblioteca será demonstrado posteriormente, onde será explicado a análise de dados.

Foi escolhida essa API porque ela é mais adequada quando o objetivo é realizar análise de dados em larga escala, porém o *Databricks* tem uma opção nativa chamada DBUtils, uma coleção de utilitários que ajudam a interagir com o ambiente *Databricks* de forma eficiente. Ele é utilizado principalmente para realizar operações em arquivos, parâmetros, *jobs* e outros recursos diretamente dentro de *notebooks*, facilitando a automação e o gerenciamento de processos.

2.3 Análise de Dados

Nesta seção, será abordado como o processo de análise de dados foi utilizado para obtenção dos resultados.

2.3.1 Obtenção dos dados

Para a realização deste trabalho, utilizou-se uma base de dados sintética de ataques cibernéticos, obtida a partir da plataforma Kaggle, uma das maiores repositórias de *datasets* públicos para análise de dados e *machine learning*. A base de dados Obtida tem o título "*Cyber Security Attacks",* ela contém 40 mil registros e 24 colunas e pode ser encontrada através do link https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks onde ela se encontra disponível em um arquivo CSV (ROCK CONTENT). A análise exploratória realizada nesta base de dados não utilizou todas as colunas disponíveis, somente as listadas a seguir, Attack Type que são os tipos de ataque realizados, Severity Level que significa o nível de risco que o ataque representa, Traffic Type que são os tráfegos utilizados para os ataques,

Protocol que são os protocolos de rede, Action Taken que são as ações tomadas para cada ataque e Log Source que é a origem do registro.

2.3.2 Preparação dos Dados

Após a obtenção dos dados, a etapa de preparação foi realizada com o objetivo de assegurar a integridade e a consistência da base utilizada. Após a análise da quantidade de dados nulos, foi decidido por preencher esses dados com valores que indicassem que não há dado a ser mostrado, pois se esses registros que contém dados nulos fossem excluídos, seria perdido metade dos registros disponíveis, o que seria inviável.

Para a análise exploratória, foi utilizado um método que possibilita o uso da linguagem SQL, para isso, outro ajuste é necessário, a renomeação das colunas, os novos nomes das colunas utilizadas na análise seguiram o padrão snake case que adiciona um "_" no lugar dos espaços entre as palavras, exemplo "Attack Type" para "attack_type"

3. RESULTADOS

A etapa final do trabalho consiste na análise exploratória dos dados, cujo foco é a extração de informações relacionadas aos ataques cibernéticos presentes na base de dados. Esse processo permite identificar comportamentos recorrentes, níveis de severidade mais comuns, e outros insights relevantes para a segurança da informação.

3.1 Distribuição de Ataques

Com esta consulta, busca-se descobrir quais os tipos de ataque presentes e a proporção da distribuição entre esses ataques. A Figura 4 nos mostra a construção da consulta e a forma de plotar o gráfico.

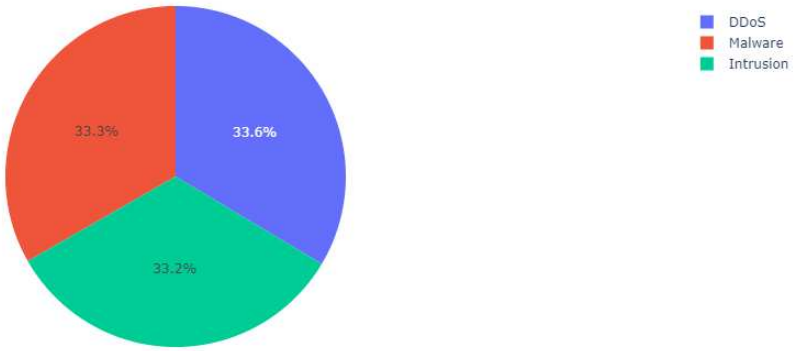
Figura 4 - Código da consulta e plotagem

```
df_attacks = ps.sql('''
SELECT attack_type, count(*) as occurence_count
FROM {DF}
GROUP BY attack_type
''', DF=df_filled)

df_attacks = df_attacks.set_index('attack_type')

df_attacks.plot.pie(y="occurence_count")
```

Fonte: Autoria própria

Para realizar a plotagem do gráfico foi preciso selecionar um dos campos retornados na consulta para ser o identificador do valor, como o objetivo é saber proporção entre os ataques então o index é o nome do ataque. A figura 5 nos mostra o resultado da consulta em formato de gráfico onde cada tipo de ataque está representado por uma cor e a proporção em porcentagem.

Figura 5 - Distribuição de ataques



Fonte: Autoria própria.

Podemos observar no gráfico que foram encontrados apenas 3 tipos de ataque, DDoS representado em azul, Malware em vermelho e Intrusion em verde, a quantidade de ataques DDoS se mostra pouco acima dos outros dois o que nos indica que não há uma predominância de nenhum dos ataques.

3.2 Distribuição dos ataques por tráfego

Com essa consulta, o objetivo é descobrir a quantidade de cada tipo de ataque por tráfego, foram utilizadas 3 funções SUM com condicionais para cada tipo de ataque e o agrupamento por tipo de tráfego, abaixo a figura 6 com o código.

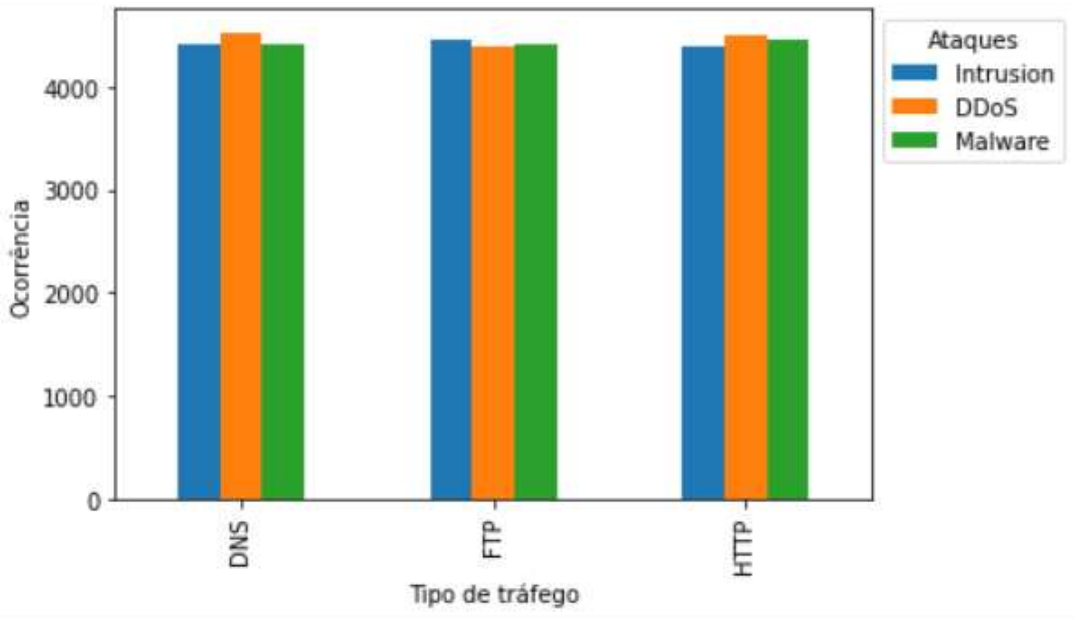
Figura 6 - Código da consulta e plotagem

```
df_attack_proportion = ps.sql('''
    SELECT
        traffic_type,
        SUM(CASE WHEN attack_type = 'Intrusion' THEN 1 ELSE 0 END) AS Intrusion,
        SUM(CASE WHEN attack_type = 'DDoS' THEN 1 ELSE 0 END) AS DDoS,
        SUM(CASE WHEN attack_type = 'Malware' THEN 1 ELSE 0 END) AS Malware
    FROM {DF}
    GROUP BY traffic_type
''', DF=df_filled)
df_most_common_attack_pd = df_most_common_attack.to_pandas()
df_most_common_attack_pd = df_most_common_attack_pd.set_index('traffic_type')
df_most_common_attack_pd.plot.bar()
plt.legend(title='Ataques', bbox_to_anchor=(1.0, 1), loc='upper left')
plt.xlabel("Tipo de tráfego")
plt.ylabel("Ocorrência")
```

Fonte: Autoria própria

Para a plotagem do gráfico, foi necessário passar o dataset para o formato pandas, depois definir o index sendo o tipo de tráfego, o resultado são 3 colunas para cada tipo de tráfego como demonstra a figura 7.

Figura 7 - Distribuição dos ataques por tráfego



Fonte: Autoria própria

Podemos observar no resultado que para os tráfegos HTTP e DNS, o mais utilizado é o DDoS já para o FTP o Intrusion é mais utilizado e que também não tem uma grande diferença entre os ataques e os tráfegos.

3.3 Distribuição de ataques por navegador

Esta consulta busca descobrir os navegadores utilizados para os ataques e a quantidade de ataques feita através de cada, a informação de qual navegador foi utilizado está no campo device information, onde a primeira informação é o navegador utilizado e logo em seguida uma "/" para separar do restante da informação, a figura 8 demonstra a construção da consulta.

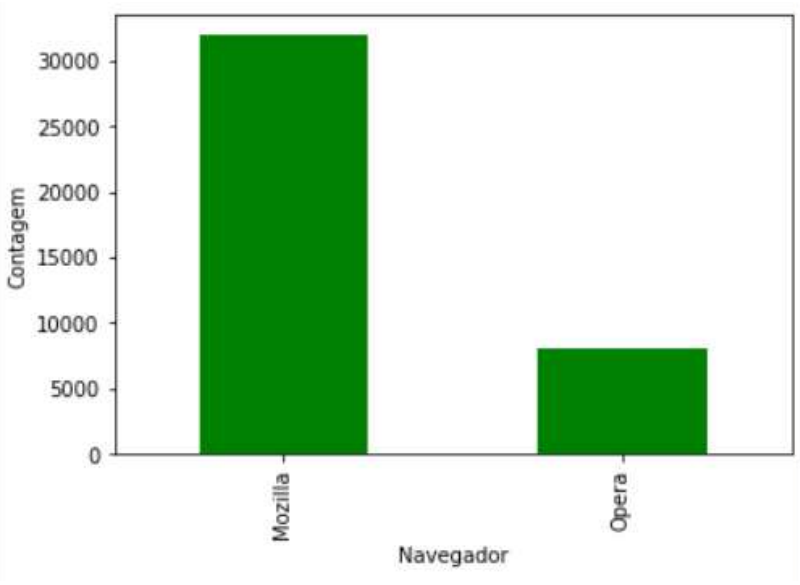
Figura 8 - Código da consulta e plotagem

```
df_browser = ps.sql('''
    SELECT
        SPLIT(device_information, '/')[0] AS browser_name,
        count(*) as Ocorrencia
    FROM {DF}
    GROUP BY browser_name
''', DF=df_filled)

df_browser = df_browser.to_pandas()
df_browser = df_browser.set_index('browser_name')
df_browser.plot(kind="bar", color="Green", legend=False)
plt.xlabel("Navegador")
plt.ylabel("Contagem")
```

Fonte: Autoria própria

A extração do navegador utilizado foi feita utilizando a função split e selecionando a posição “0” onde está a parte desejada. A plotagem foi feita em um gráfico de barras com o index sendo o “browser_name” assim como nomeado na consulta e convertido para o padrão pandas, abaixo a figura 9 com o resultado

Figura 9 - Distribuição dos ataques por navegador



Fonte: Autoria própria

O resultado da consulta nos mostra que foram registrados apenas 2 navegadores na base de dados o Mozilla e o Opera. É possível perceber que o Mozilla concentra a grande maioria dos ataques registrados, o que pode significar que o nível de segurança neste navegador é muito baixo.

3.4 Distribuição de ataques por mês

Umas das colunas contém a informação da data em que ocorreu o ataque, a coluna Timestamp, como o próprio nome já diz ela armazena esse dado no formato timestamp, então fica fácil de fazer a extração que se busca fazer nessa consulta, que é a distribuição dos ataques por mês, para isso foi utilizada a função MONTH. A demonstração da consulta está na figura 10 logo a seguir.

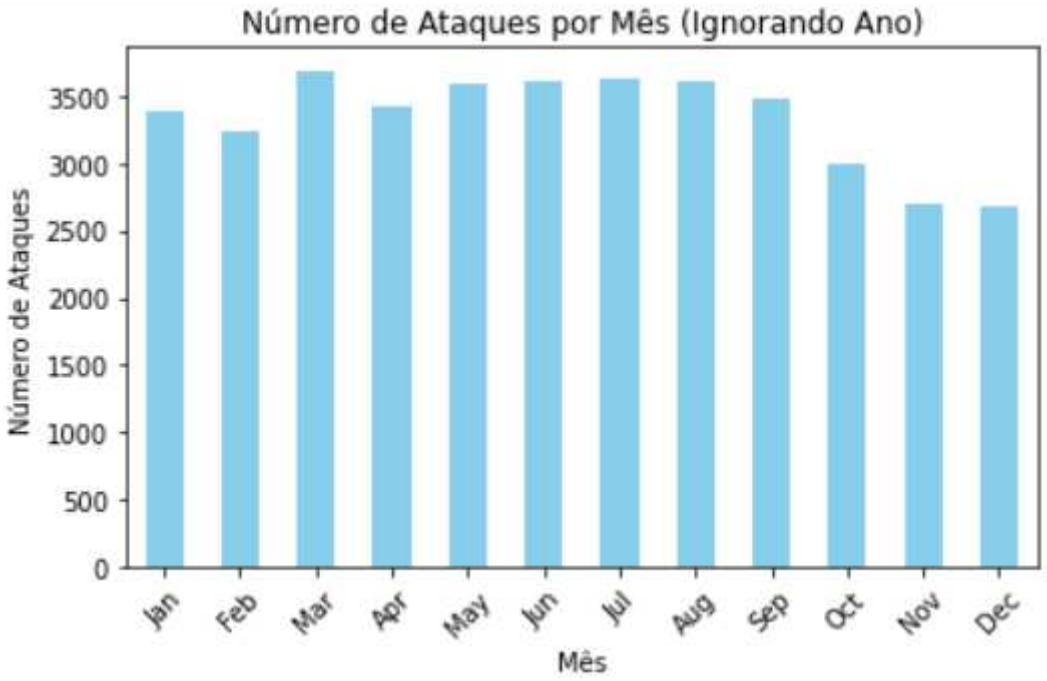
Figura 10 - Código da consulta e plotagem

```
df_monthly_attacks = ps.sql('''
    SELECT
        MONTH(Timestamp) AS month,
        COUNT(*) AS attack_count
    FROM {DF}
    GROUP BY month
    ORDER BY month ASC
''', DF=df_filled)
# Converte para Pandas para facilitar a plotagem
df_monthly_attacks = df_monthly_attacks.to_pandas()
df_monthly_attacks = df_monthly_attacks.set_index('month')
# Renomeia os índices para o nome dos meses (opcional)
df_monthly_attacks.index = ["Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"]
# Plotagem
plt.figure(figsize=(10, 6))
df_monthly_attacks.plot(kind="bar", color="skyblue", legend=False)
plt.xlabel("Mês")
plt.ylabel("Número de Ataques")
plt.title("Número de Ataques por Mês (Ignorando Ano)")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Fonte: Autoria própria

Para a plotagem, foi novamente convertido para o formato pandas para facilitar a plotagem, definido o “month” como o index e em seguida foi feito uma lista com as iniciais dos meses, para melhor visualização. O resultado da consulta segue na figura 11.

Figura 11 - Distribuição dos ataques por mês



Fonte: Autoria própria

O resultado da consulta mostra que o mês de março teve mais ataques que nos demais, que nos últimos meses do ano os ataques decaem, o que pode significar que com o decorrer do tempo os métodos de combate a esses ataques vão melhorando.

3.5 Distribuição de ataques por sistema operacional

Uma das informações da coluna “device_information” é o sistema operacional que sofreu ataque, o foco dessa consulta vai ser nos sistemas Windows e Linux, a seguir na figura 12 o código.

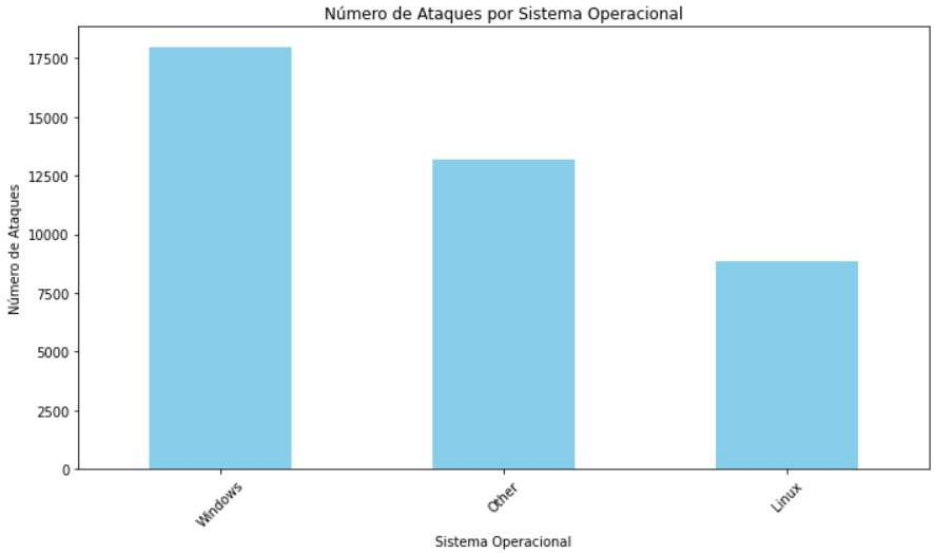
Figura 12 - Código consulta e plotagem

```
df_os_attacks = ps.sql('''
    SELECT
        CASE
            WHEN device_information LIKE '%Windows%' THEN 'Windows'
            WHEN device_information LIKE '%Linux%' THEN 'Linux'
            ELSE 'Other'
        END AS operating_system,
        COUNT(*) AS attack_count
    FROM {DF}
    GROUP BY operating_system
    ORDER BY attack_count DESC
''', DF=df_filled)

# Converte para Pandas para facilitar a plotagem
df_os_attacks = df_os_attacks.to_pandas()

# Plotagem
plt.figure(figsize=(10, 6))
df_os_attacks.set_index('operating_system')['attack_count'].plot(kind='bar', color="skyblue")
plt.xlabel("Sistema Operacional")
plt.ylabel("Número de Ataques")
plt.title("Número de Ataques por Sistema Operacional")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Fonte: Autoria própria

Para fazer a seleção utilizou o “%” no início e no fim da palavra, para indicar que não importa o que vem antes ou depois, contanto que tenha a palavra, a plotagem segue igual às demais. O resultado da consulta foi apresentado na figura 13 a seguir.

Figura 13 - Distribuição de ataques por sistema operacional



Fonte: Autoria própria

O resultado da consulta nos mostra que o sistema Windows teve mais ataques que os outros sistemas o que indica uma falta de segurança por parte dele e que o Linux teve menos ataques em relação a outros sistemas, porém os ataques dessa coluna central está distribuída entre outros sistemas, o que indica o foco nesses dois sistemas operacionais.

4. CONSIDERAÇÕES FINAIS

Durante o desenvolvimento deste trabalho foram enfrentados alguns desafios técnicos, principalmente na pesquisa da base de dados e sobre quais consultas fazer para que trouxesse resultados satisfatórios. Inicialmente, foi preciso procurar uma base de dados para se fazer a análise, pois a aquisição de dados de diferentes fontes e com parâmetros de comparação semelhantes seria inviável devido ao tempo de desenvolvimento, por isso foi necessário a busca por uma base de dados já pronta, para isso foi utilizado a plataforma kaggle.

Dentre as dificuldades encontradas, estava a escolha dos parâmetros que seriam interessantes para a análise, como a base de dados é feita com dados simulados, muitos dos valores dos parâmetros são distribuídos proporcionalmente, o que dificultou o planejamento da análise.

Um ponto que o projeto não aborda é a criação de um modelo em *machine learning* que receba os resultados e aprenda a identificar os padrões, e isso pode ser uma melhoria para trabalhos futuros, expandindo para uma análise em tempo real.

Em relação ao propósito inicial deste trabalho, os objetivos foram alcançados, destacando-se principalmente a implementação de consultas e a apresentação dos resultados que retornaram informações relevantes sobre os ataques de forma simples e objetiva.

REFERÊNCIAS

ALURA. **O que é Tableau?**. Alura. Disponível em: <https://www.alura.com.br/artigos/o-que-e-tableau?srsltid=AfmBOoreNXGf835J89vdaQEvdRAHkxVdwSz-X9CkspiLMM0RpycKIPEk>. Acesso em: 16 outubro 2024.

ANDRADE, Luiz Claudio Oliveira de. **O uso do Big Data na prevenção de ataques cibernéticos**. 2020. Trabalho de Conclusão de Curso (TCC) – Escola de Comando e Estado-Maior do Exército, Rio de Janeiro, 2020. Disponível em: <https://bdex.eb.mil.br/jspui/bitstream/123456789/7601/1/MO%206241%20-%20LUIZ%20CLAUDIO.pdf>. Acesso em: 19 agosto 2024.

BASTOS, Athena. **Análise de dados: uma ferramenta para criar melhores estratégias de negócio**. 2024. Disponível em: <https://www.alura.com.br/empresas/artigos/analise-de-dados>. Acesso em: 17 jul. 2024.

CALANCA, Paulo. **Databricks: o que é e para que serve?**. 2023. Disponível em: <https://www.alura.com.br/artigos/databricks-o-que-e-para-que-serve>. Acesso em: 22 julho 2024.

CATUNDA, Heitor. **O que é o Kaggle? Entenda e saiba como começar a usá-lo**. Hashtag, 2022. Disponível em: <https://www.hashtagtreinamentos.com/kaggle?gad_source=1&gclid=Cj0KCQjwsoe5BhDiA

RIsAOXVoUtYtih89Ty6vPZ1M5goIdALwwI7xFMQ6yjiH_BwcNmCnzLX4nLowaQaAnS3EALw_wcB>. Acesso em: 15 julho 2024.

DATABRICKS. **Glossário PySpark. Databricks**. Disponível em: <https://www.databricks.com/br/glossary/pyspark#:~:text=PySparkSQL%20é%20uma%20biblioteca%20PySpark,um%20wrapper%20do%20PySpark%20Core>. Acesso em: 16 outubro 2024.

DATABRICKS. **Introduction to Databricks notebooks**. Databricks Documentation. Disponível em: <https://docs.databricks.com/notebooks/index.html>. Acesso em: 19 setembro 2024.

DATABRICKS. **pyspark.pandas API Reference**. Databricks Documentation. Disponível em: <https://docs.databricks.com/spark/latest/spark-sql/pandas-on-spark.html>. Acesso em: 27 setembro 2024.

DATABRICKS. **What is the Databricks File System (DBFS)?**. Databricks Documentation, [s.d.]. Disponível em: <https://docs.databricks.com/dbfs/index.html>. Acesso em: 27 setembro 2024.

EBAC. **O que é Power BI?**. EBAC Online. Disponível em: <https://ebaconline.com.br/blog/o-que-e-power-bi#:~:text=O%20Power%20BI%20é%20uma%20ferramenta%20de%20avaliação%20e%20visualização,de%20maneira%20simples%20e%20intuitiva>. Acesso em: 16 outubro 2024.

HIREMATH, Shivashankar; SHETTY, Eeshan; PRAKASH, Allam Jaya; SAHOO, Suraj Prakash; PATRO, Kiran Kumar; RAJESH, Kandala N. V. P. S.; PŁAWIAK, Paweł. **A new approach to data analysis using machine learning for cybersecurity**. Big Data Cogn. Comput. 2023. Disponível em: <https://www.mdpi.com/2504-2289/7/4/176>. Acesso em: 26 agosto 2024.

IBERDROLA. **Ataques cibernéticos: Quais são os principais e como se proteger deles?**. 2021. Disponível em: <https://www.iberdrola.com/inovacao/ciberataques#:~:text=DEFINIÇÃO%20DE%20CIBERATAQUE,prejudicar%20pessoas,%20instituições%20ou%20empresas>. Acesso em: 22 jul. 2024.

JSON. **JSON: JavaScript Object Notation**. Disponível em: <https://www.json.org/json-pt.html>. Acesso em: 8 outubro 2024.

JUNQUEIRA, Dyogo. **Cientista de dados ganha relevância na segurança cibernética**. 2024. Disponível em: <https://www.cisoadvisor.com.br/security-room-posts/cientista-de-dados-ganha-relevancia-na

-seguranca-cibernetica/#:~:text=A%20ciência%20de%20dados%20deve,prevenir%20a%20exposição%20dos%20dados>. Acesso em: 26 agosto 2024.

MAIA, ElijonasS. **Ataques hackers aumentam 8,8% no Brasil e país segue como 2º mais atacado do mundo**. CNN, Brasília, 2024. Disponível em: <https://www.cnnbrasil.com.br/nacional/ataques-hackers-aumentam-88-no-brasil-e-pais-segue-como-2o-mais-atacado-do-mundo/>. Acesso em: 13 agosto 2024.

MICROSOFT. **Work with pandas on PySpark**. Microsoft Learn. Disponível em: <https://learn.microsoft.com/en-us/azure/databricks/spark/latest/spark-sql/pandas-on-spark/>. Acesso em: 27 setembro 2024.

MICROSOFT. **XML para iniciantes**. Disponível em: <https://support.microsoft.com/pt-br/office/xml-para-iniciantes-a87d234d-4c2e-4409-9cbc-45e4eb857d44>. Acesso em: 8 outubro 2024.

REDAÇÃO. **Ataques cibernéticos no trabalho remoto mais que triplicaram durante a pandemia**. 2023. Disponível em: <https://securityleaders.com.br/ataques-ciberneticos-no-trabalho-remoto-mais-que-triplicam-durante-a-pandemia/>. Acesso em: 15 agosto 2024.

RIPARI, César. **Por que dados são considerados o novo petróleo?**. 2022. Disponível em: <https://administradores.com.br/noticias/por-que-dados-sao-considerados-o-novo-petroleo>. Acesso em: 22 julho 2024.

ROCK CONTENT. **O que é CSV? Entenda como funciona o formato e aprenda a abrir arquivos!**. Disponível em: <https://rockcontent.com/br/blog/csv/>. Acesso em: 8 outubro 2024.