INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Técnicas de encriptação e anonimização para Big Data

DAVID MIGUEL COUTINHO MARQUES
Outubro de 2023

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Técnicas de encriptação e anonimização para Big Data

David Marques

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Sistemas de Informação e Conhecimento

Orientador: Ana Maria Dias Madureira Pereira (AMD)
Co-Orientador: Jorge Fernandes Rodrigues Bernardino

Júri:
Presidente:

Vogais:

Porto, 12 de outubro de 2023

Dedicatória

Um projecto final de graduação é uma longa viagem, que inclui uma trajectória permeada por numerosos desafios e muitos percalços pelo caminho, mas apesar do processo, recolhe contribuições de várias pessoas, indispensáveis para encontrar o melhor resultado final. A realização deste projecto só foi possível com o apoio, energia e força de várias pessoas, a quem dedico especialmente este projecto. Especialmente aos meus dois orientadores professora Doutora Ana Madureira e professor Doutor Jorge Bernardino, que sempre se mostraram disposta a ajudar e sempre com um interesse permanente, uma visão crítica e oportuna, exigindo de uma forma enriquecedora para obter o melhor de mim. Finalmente, o meu agradecimento mais profundo a aos meus pais que sempre me apoiaram e incentivaram em todos os meus desafios a que me propus. Pelo tempo que não lhes dediquei, pela formação que me permitiram adquirir e por toda a ajuda incondicional.

Resumo

O objetivo desta dissertação é investigar várias estratégias que podem ser utilizadas para encriptar e ocultar as identidades em grandes dados (Big Data). Uma vez que os sistemas de *Big Data* estão a tornar-se mais predominantes numa vasta gama de indústrias, a proteção da confidencialidade dos dados sensíveis tem surgido como uma das preocupações mais prementes nos dias de hoje. A encriptação e a anonimização de dados são duas formas típicas utilizadas para preservar a privacidade dos dados em contextos que fazem uso intensivo de *Big Data*.

Esta dissertação começa com uma visão exaustiva e análise detalhada dos diferentes métodos de encriptação e de anonimização que são amplamente utilizados na área de *Big Data*. Nesta visão geral, tanto os métodos clássicos de encriptação como a encriptação simétrica e assimétrica, assim como as estratégias modernas de encriptação como a encriptação *homomorphic* e a encriptação baseada em atributos, são discutidos em detalhe. De forma semelhante, uma grande variedade de estratégias de anonimização, tais como *k-anonymity*, *l-diversity*, e *t-closeness*, são examinadas em pormenor.

Após a conclusão da avaliação, estas estratégias de encriptação e anonimização são analisadas e contrastadas, tendo em consideração os seus benefícios, inconvenientes e aplicabilidade a uma variedade de situações de *Big Data*. No decurso da investigação, são considerados fatores como a sobrecarga computacional, a utilidade dos dados, e a resistência ao ataque.

Além disso, a dissertação contém uma secção prática na qual as técnicas de encriptação e de anonimização escolhidas são implementadas e avaliadas num conjunto de dados de *Big Data*. A avaliação inclui métricas de desempenho tais como tempo de encriptação/desencriptação de dados, perda de dados durante a anonimização, e a capacidade da anonimização para preservar a utilidade dos dados.

Os resultados desta dissertação contribuem para um melhor conhecimento das estratégias de encriptação e anonimização de *Big Data*, e também dão aos investigadores e profissionais informações sobre como escolher as abordagens mais adequadas para proteger dados sensíveis em ambientes que utilizam de *Big Data*. Ao considerar estratégias de encriptação e de anonimização para uma variedade de casos de utilizam *Big Data*, os resultados ressaltam a importância de encontrar um equilíbrio adequado entre a proteção de dados e o seu potencial uso.

Palavras-chave: Big Data, Data encryption, Network security, Data encryption

Abstract

The aim of this thesis is to investigate various strategies that can be used to encrypt and hide the identities of *Big Data*. As *Big Data* systems are becoming more prevalent in a wide range of industries, protecting the confidentiality of sensitive data has emerged as one of the most pressing concerns today. *Data encryption* and *Data anonymisation* are two typical ways used to preserve data privacy in contexts that make intensive use of *Big Data*.

This thesis starts with an exhaustive overview and detailed analysis of the different encryption and anonymization methods that are widely used in the area of Big Data. In this overview, both classical encryption methods like symmetric and asymmetric encryption, as well as modern encryption strategies like homographic encryption and attribute-based encryption are discussed in detail. Similarly, a wide variety of anonymization strategies, such as k-anonymity, l-diversity, and t-closeness, are examined in detail.

Upon completion of the evaluation, these encryption and anonymisation strategies are analysed and contrasted, taking into consideration their benefits, drawbacks and applicability to a variety of *Big Data* situations. The research takes into account factors such as computational overhead, data utility and resistance to attack.

In addition, the thesis includes a practical part where the chosen encryption and anonymisation techniques are implemented and evaluated on a set of big data. The evaluation includes performance metrics such as data encryption/decryption time, data loss during anonymisation, and the ability of anonymisation to preserve data utility.

The results of this dissertation contribute to a better understanding of *Big Data* encryption and anonymisation strategies, and also provide researchers and practitioners with information on how to choose the most appropriate approaches to protect sensitive data in environments using *Big Data*. By considering encryption and anonymisation strategies for a variety of *Big Data* use cases, the results highlight how critical it is to find a healthy balance between data protection and potential use.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

Capítulo 1

Introdução

Este capítulo pretende descrever o contexto e a motivação para explorar as diversas estratégias de encriptação de dados para as abordagens dos sistemas de Big Data. Além disso, os objetivos desta dissertação de mestrado serão especificados. Finalmente, a abordagem e a metodologia serão descritas, assim como o resto da estrutura do documento.

1.1 Motivação

A inspiração para a criação esta dissertação sobre técnicas de encriptação e anonimização de *Big Data* surge do grande volume dos dados na sociedade moderna. A rápida progressão da tecnologia levou a uma dependência crescente de dados em diversos campos, desde as finanças e os cuidados de saúde aos meios de comunicação social e ao comércio eletrónico.

Como a quantidade de dados a ser gerada e processada continua a crescer exponencialmente, a necessidade de medidas de segurança robustas para proteger a informação sensível tornou-se primordial. As violações de dados e da privacidade tornaram-se uma preocupação significativa, com consequências potencialmente graves para os indivíduos, organizações e a sociedade no seu todo.

O domínio de *Big Data*, que engloba o tratamento e o exame de conjuntos volumosos de dados, coloca obstáculos distintos no que diz respeito à salvaguarda de dados. A grande escala e complexidade de *Big Data* exigem metodologias especializadas para garantir a segurança dos dados ao longo de toda a sua vida útil, abrangendo a transmissão, armazenamento e processamento.

A encriptação e a anonimização são técnicas geralmente utilizadas para salvaguardar o sigilo e a confidencialidade da informação. O processo de encriptação implica a conversão de dados num texto cifrado que só pode ser descodificado com a chave correspondente. Por outro lado, a anonimização envolve a alteração ou remoção de informações identificáveis dos dados para salvaguardar a privacidade dos indivíduos.

A importância da salvaguarda de dados no domínio de *Big Data* é da maior importância e não deve ser subestimada. As ramificações das violações de dados ou do acesso não autorizado a informações sensíveis têm o potencial de se alargar significativamente, abrangendo perdas financeiras, danos à reputação, obrigações legais, e violações dos direitos de privacidade [1].

A presente dissertação procura investigar e avaliar diversas metodologias de encriptação e anonimização de *Big Data*, tendo em consideração os seus pontos fortes, limitações, e aplicabilidade em vários cenários. A aplicação e avaliação empírica destas metodologias sobre conjuntos de *Big Data* reais oferecerá perspetivas e recomendações significativas para

profissionais e académicos no processo de escolha das abordagens mais apropriadas para garantir a confidencialidade dos dados em contextos de *Big Data*. Este estudo visa dar uma contribuição académica no campo da segurança e privacidade de dados no contexto de *Big Data*, explorando a complexa interação entre a segurança e a utilidade dos dados.

1.2 Objetivos

O objetivo deste projeto de dissertação de mestrado é comparar e avaliar as várias técnicas de encriptação e anonimização utilizadas para os sistemas de *Big Data*. Especificamente, pretende examinar as diferenças e semelhanças entre os vários algoritmos de encriptação, fazendo um estudo empírico dos algoritmos de encriptação mais populares e uma avaliação dos benefícios e desvantagens oferecidos por cada um deles. A análise dos vários métodos de encriptação deve incluir certas características essenciais, tais como a Encriptação Simétrica e a Encriptação Assimétrica, bem como o tempo de encriptação, e identificar os benefícios e desvantagens de cada um. A fim de se chegar a uma conclusão sobre a proposta para este projeto, um estudo cuidadoso dos algoritmos, e as suas ações sobre um conjunto de dados foram observadas.

Será desenvolvido um cenário para demonstrar a aplicação prática destas técnicas. Este cenário envolverá um grande conjunto de dados contendo informações sensíveis, que devem ser partilhados de forma segura entre múltiplas partes, mantendo simultaneamente a confidencialidade. Para proteger os dados, serão implementadas as técnicas de encriptação e anonimização selecionadas, e a sua eficácia será avaliada com base em fatores como a segurança dos dados, a proteção da privacidade e a usabilidade dos dados.

Este documento procura fornecer uma análise abrangente das técnicas de encriptação e de anonimização de *Big Data*, destacando os seus benefícios, desvantagens, e aplicações no mundo real utilizando como exemplo um caso prático.

O presente projeto de dissertação de mestrado tem como objetivo.

- Comparar e avaliar várias técnicas de encriptação e anonimização utilizadas em sistemas de Big Data, incluindo algoritmos de encriptação simétrica e assimétrica.
- Realizar um estudo empírico dos algoritmos de encriptação mais populares, avaliando as suas características, benefícios e desvantagens, incluindo o tempo de encriptação.
- Identificar os benefícios e desvantagens de cada técnica de encriptação e anonimização em termos de segurança dos dados, proteção da privacidade e usabilidade dos dados.
- Desenvolver um cenário prático que envolva um elevado conjunto de dados contendo informações sensíveis, onde serão implementadas as técnicas de encriptação e anonimização selecionadas.
- Avaliar a eficácia das técnicas de encriptação e anonimização implementadas no cenário prático, analisando a segurança dos dados, a proteção da privacidade e a usabilidade dos dados.
- Propor conclusões e recomendações com base nos resultados obtidos, incluindo possíveis melhorias nas técnicas de encriptação e anonimização de Big Data, e áreas de investigação futura.

1.3 Estrutura

Antes de prosseguir com an análise dos capítulos subsequentes, é pertinente realizar uma análise das principais questões abordadas em cada capítulo, com o intuito de fornecer uma visão abrangente do trajetória planeada da investigação e da obtenção de conhecimentos.

Capítulo 1: Introdução

Este capítulo introdutório da dissertação começa por fornecer uma visão geral dos antecedentes e dos fundamentos que sustentam a investigação. Nesta análise, é aprofundado as motivações subjacentes a este estudo e os objetivos específicos que se pretende atingir. Para além disso, será apresentada uma análise abrangente da dissertação, delineando o quadro geral e os temas explorados ao longo da investigação.

Capítulo 2: Metodologia do estudo

O capítulo seguinte debruçar-se-á sobre os métodos de investigação utilizados neste estudo específico. O presente estudo incluirá uma análise dos inquéritos de investigação que orientaram o curso da ação, bem como as bases de dados e as frases de pesquisa utilizadas. Os critérios de exclusão e os procedimentos de extração de informação utilizados serão avaliados de forma a garantir a integridade dos dados recolhidos.

Capítulo III: Os fundamentos teóricos

Este capítulo explora os fundamentos teóricos necessários para compreender os aspetos contextuais do *Big Data* e da segurança da informação. O objetivo deste estudo é examinar os princípios fundamentais relativos ao processamento de dados no âmbito do *Big Data*, os instrumentos correspondentes utilizados para este fim e a importância de salvaguardar a segurança da informação neste domínio.

Capítulo IV: Análise de Valor

O quarto capítulo deste estudo centra-se na análise de valor, com ênfase específica na identificação de oportunidades e na avaliação do valor associado à solução sugerida. Neste estudo, faremos uma análise das oportunidades encontradas. Além disso, será feita uma análise do valor percebido associado a essas possibilidades.

Capítulo V: Execução da solução

Este capítulo fornecerá uma visão global da implementação da solução, incluindo uma análise das tecnologias e das estruturas utilizadas. Neste discurso, será abordado os conceitos de encriptação homomórfica, centrando especificamente na estrutura conhecida como SEAL. Além disso, será investigado o domínio da Computação Multipartidária Segura, bem como os princípios subjacentes à Privacidade Diferencial.

Capítulo VI: Conclusão

No capítulo final, sintetizamos toda a investigação, salientando os objetivos alcançados e os resultados obtidos. Além disso, serão examinadas as perspetivas de investigação futura, culminando com um resumo pormenorizado dos contributos deste estudo, concluindo assim a dissertação.

Capítulo 2

Metodologia de Pesquisa

Um dos objetivos de uma revisão sistemática da literatura é compilar uma lista de todas as obras publicadas sobre um determinado assunto. Ao contrário do processo de revisão convencional, que procuram resumir as conclusões de diversos estudos, uma revisão sistemática aplica critérios definidos e rigorosos para descobrir, avaliar criticamente e sintetizar a literatura relevante sobre um determinado assunto. A revisão sistemática utilizada neste projeto foi inspirada na metodologia PRISMA [1]. A principal finalidade desta revisão é estabelecer uma sólida base para que o leitor compreenda os conceitos existentes relacionados à criptografia em sistemas de Big Data em larga escala, ao mesmo tempo que realça a necessidade de novas investigações neste campo.

Tabela 2.1: Questões de investigação

	Questões de investigação
RQ1	Quais as melhores técnicas de encriptação existentes?
RQ2	Quais as técnicas tradicionais de encriptação que podem ser utilizadas em sistemas de Big Data ?
RQ3	Que tipos de encriptação existem para sistemas de Big Data?
RQ4	Quais as técnicas de anonimização existentes para Big Data?
RQ5	Que técnicas de anonimização podem ser utilizadas em sistemas de Big Data ?

2.1 Questões de investigação

O objetivo deste estudo é identificar as melhores técnicas de encriptação para sistemas de *Big Data* (RQ1) através da realização de uma análise abrangente das técnicas de encriptação existentes. Procura avaliar os prós e os contras de várias técnicas de cifragem, tais como a cifragem simétrica e assimétrica, e determinar a sua adequação para proteger dados em grande escala em sistemas de *Big Data*.

Além disso, este trabalho procura identificar técnicas tradicionais de encriptação aplicáveis aos sistemas de *Big Data* (RQ2). Investigar métodos de encriptação bem estabelecidos que têm sido amplamente utilizados em cenários tradicionais de segurança de dados e avaliar a sua aplicabilidade e eficácia no contexto dos sistemas de *Big Data*, que frequentemente envolvem grandes volumes, alta velocidade, e diversos tipos de dados.

[1] http://www.prisma-statement.org

Além disso, esta investigação visa identificar as variedades de encriptação mais adequadas aos sistemas de *Big Data* (RQ3). Irá analisar as vantagens, desvantagens, e aplicabilidade de várias abordagens de encriptação, tais como cifra em bloco e encriptação homomórfica, no contexto dos sistemas de *Big Data*, tendo em conta a escalabilidade, desempenho, e requisitos de segurança.

Para além da encriptação, esta investigação procura identificar técnicas de anonimização de *Big Data* (RQ4). Investigar várias técnicas de anonimização, tais como generalização, supressão e substituição, que podem ser aplicadas em sistemas com volume elevado de dados para salvaguardar a privacidade de dados sensíveis.

O objetivo final deste estudo é identificar técnicas de anonimização aplicáveis aos sistemas de *Big Data* (RQ5). Examinar a viabilidade de vários métodos de anonimização para assegurar dados em grande escala em sistemas de *Big Data*, tendo em conta fatores como a utilidade dos dados, a preservação da privacidade, e o cumprimento dos regulamentos de proteção de dados. Os pontos fortes e as limitações destas técnicas serão analisados para determinar a sua aplicabilidade em cenários do mundo real de *Big Data*.

2.2 Bases de dados

A primeira coisa a fazer para realizar uma revisão sistemática da literatura é localizar os diferentes tipos de dados que serão necessários para a investigação. As bases de dados que foram utilizadas nesta análise são detalhadas no Tabela 2.2. Na fase subsequente, teremos em conta o facto de que algumas destas fontes de dados se sobrepõem.

Tabela 2.2: Base de dados

Identificador	Base de dados	URL
DS1	ACM Digital Library	`https://dl.acm.org/`
DS2	IEEE Explore	`https://ieeexplore.ieee.org/`
DS3	Scopus	`https://www.scopus.com/home.uri`
DS3	DBLP	`https://dblp.org/`

2.3 Termos de investigação

Este estudo examinou não só o número de estudos concluídos para cada um dos subdomínios explorados, mas também a influência da sua combinação sobre o número de descobertas. Para o efeito, foram selecionados os seguintes domínios e subdomínios, bem como as respetivas palavras-chave, para inclusão na base de dados. Utilizámos as ferramentas das bases de dados acima mencionadas, que nos permitem conjugar palavras usando e ou, para fazer uma pesquisa mais sofisticada.

```
TITLE-ABS-KEY (
1 ("Data Encryption" OR "Data Anonymisation" OR "Data Anonymization")
2 OR
3 ("Big Data Encryption" OR "Big Data Anonymisation" OR "Big Data
    Anonymization")
}
```

2.4 Avaliação da qualidade: critérios de exclusão

Esta avaliação foi regulada por um conjunto de critérios que serão utilizados para determinar se um determinado parâmetro deve ser investigado para inclusão nesta avaliação. Em certos casos, considerações tais como se o artigo foi publicado em formato de livro serão consideradas, enquanto noutros, a ênfase será colocada na qualidade global e relevância do material. Quando os sistemas forem avaliados, será dada prioridade aos que foram desenvolvidos e testados na medida possível. As medidas descritas em baixo revelam os critérios de exclusão, que irão ser utilizados nesta investigação:

Critérios de Exclusão:

- **EC1** Artigos não escritos em inglês ou português, uma vez que a proficiência linguística pode afetar a exatidão da compreensão e análise;
- **EC2** Artigos não publicados nos últimos 4 anos (a partir de 2018);
- **EC3** Artigos que não estão disponíveis em texto integral ou que requerem pagamento pelo acesso, a menos que sejam críticos para a investigação;
- **EC4** Artigos que não fornecem informação substancial ou análise sobre técnicas de encriptação ou anonimização para *Big Data*;
- **EC5** Artigos que não apresentam resultados ou avaliação experimental.

2.5 Extração da informação

A seleção de artigos para inclusão na revisão é uma componente crucial na metodologia PRISMA [2]. A abordagem de pesquisa deve ser exaustiva, replicável, e adequada à questão de pesquisa específica que está a ser tratada.

Uma vez realizada a pesquisa, o passo seguinte é avaliar a relevância e elegibilidade da pesquisa resultante. Isto implica examinar os títulos e resumos de cada pesquisa para avaliar se esta satisfaz os requisitos de inclusão. Os critérios de inclusão baseiam-se na conceção da investigação, população, intervenção ou exposição de interesse, e os resultados são monitorizados.

Após a primeira triagem, os textos completos das restantes investigações são normalmente examinados para avaliar a sua admissibilidade. Isto requer uma avaliação mais abrangente da conceção, métodos e resultados do estudo, que deve ser conduzida por pelo menos dois revisores independentes para reduzir o enviesamento.

A última fase da metodologia PRISMA é a extração de dados dos estudos incluídos. Isto requer a identificação de informação crítica de cada investigação, incluindo a conceção do estudo, a demografia, a intervenção, as medidas de resultados, e os resultados.

Como se pode ver na Figura 2.1, foi descoberto um total de 1802 artigos dentro das bases de dados que tinham sido selecionadas anteriormente. Os resultados foram ordenados com a base de dados IEEE em primeiro lugar. Apenas 121 documentos permaneceram após a aplicação dos critérios de exclusão de EC1 a EC3, o que resultou na eliminação de 1679. Após uma análise dos títulos e resumos dos documentos, 334 documentos foram eliminados da análise. Na fase de determinação da elegibilidade, os restantes artigos foram analisados

[2] http://www.prisma-statement.org

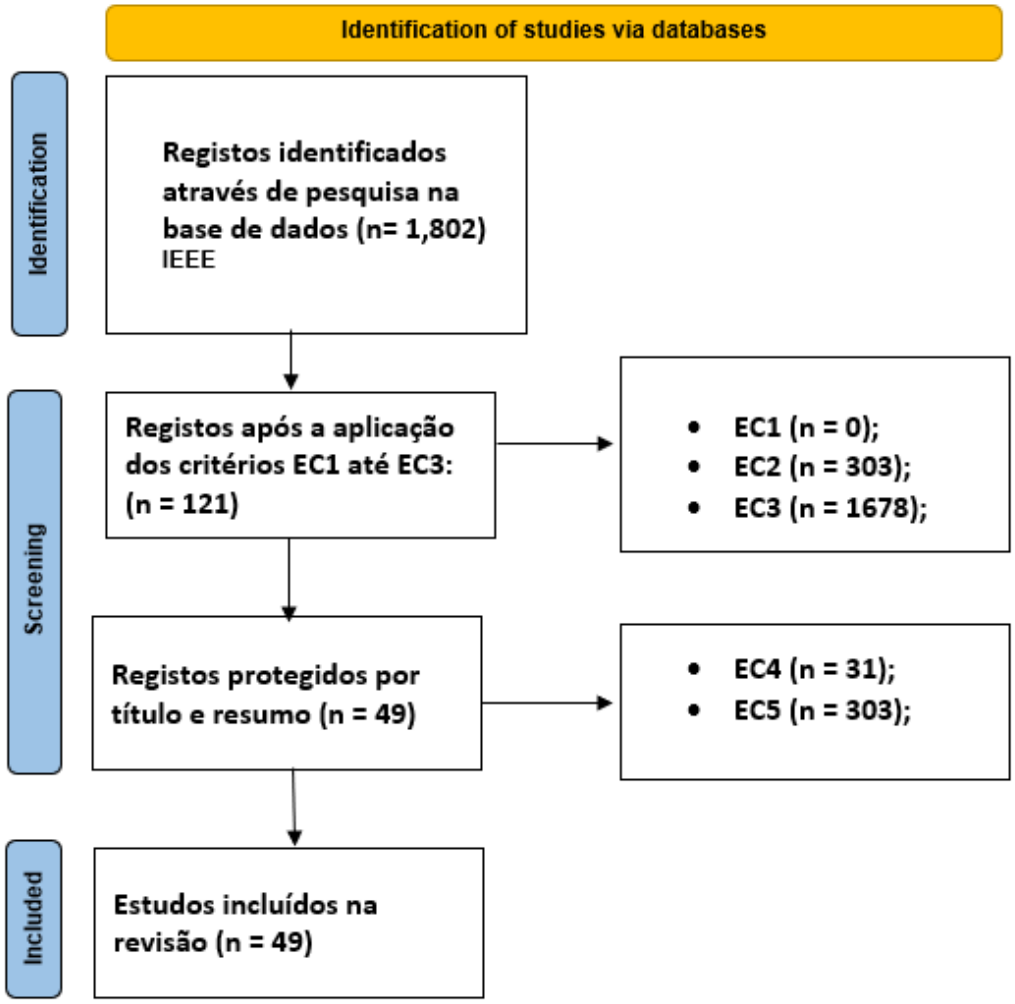

Figura 2.1: Diagrama Prisma

segundo os critérios de inclusão e exclusão. Para a fase de inclusão, apenas 49 artigos satisfaziam os critérios.

2.6 Resumo

O capítulo sobre a metodologia de investigação assume uma importância vital. O primeiro passo envolve a identificação de questões de investigação que servem para orientar a investigação e abordagem estratégica. Estas questões funcionam como pontos fulcrais, fornecendo orientação na procura de respostas e resoluções significativas para questões relacionadas com a segurança dos dados no domínio dos *Big Data*.

O processo de seleção de uma base de dados tem uma importância significativa. A seleção de fontes de informação adequadas é crucial para a aquisição de dados relevantes e inclusivos. Além disso, foram criadas frases de pesquisa, de forma a garantir uma cobertura abrangente dos componentes essenciais relacionados com a segurança dos dados.

A avaliação da qualidade dos dados é uma fase essencial e fulcral em qualquer projeto de investigação. Neste estudo, foi descrito os critérios de exclusão rigorosos que foram utilizados para garantir a inclusão apenas de artigos credíveis e relevantes na análise. Os critérios acima mencionados constituem a base fundamental para garantir a integridade e a validade dos resultados.

A utilização da abordagem PRISMA para a extração de informação é um aspeto notável discutido neste capítulo. A utilização desta abordagem sistemática permite a recolha de informações de uma forma metódica e organizada a partir de muitas fontes de investigação e de dados. A utilização do quadro PRISMA facilita o estabelecimento de consistência nos procedimentos analíticos, garantindo assim que todos os dados são submetidos a uma abordagem padronizada.

O capítulo da Metodologia de Investigação tem como objetivo estabelecer as bases metodológicas da investigação e oferecer os princípios orientadores que informam todo o trabalho futuro. Com base nos fundamentos acima mencionados, os capítulos subsequentes analisam em profundidade as abordagens de encriptação e anonimização no domínio de *Big Data*.

Capítulo 3

Big Data, Encriptação e Anonimização: fundamentos teóricos

Nas últimas décadas, o *Big Data* tem sido uma das tecnologias mais significativas e em rápido desenvolvimento, facilitando a acumulação, o armazenamento e a análise em tempo real de grandes volumes de dados. No entanto, a utilização crescente desta informação também coloca desafios significativos à segurança e privacidade dos dados.

Este capítulo abordará os fundamentos dos sistemas de *Big Data* e a importância da segurança da informação neste contexto. Além disso, será investigado as técnicas de encriptação e anonimização, incluindo os algoritmos de encriptação mais populares e as técnicas de anonimização de dados.

Este capítulo começa com uma panorâmica dos conceitos fundamentais de *Big Data* e da sua importância na análise de grandes volumes de dados. De seguida, será discutida a importância da segurança da informação nos sistemas de *Big Data*, tendo em conta as ciberameaças, as violações de dados e outras preocupações relacionadas com a privacidade.

As técnicas de encriptação e anonimização, que são essenciais para garantir a segurança e a privacidade dos dados em Big Data, também serão investigadas. Serão descrito os algoritmos de encriptação mais frequentemente utilizados neste contexto, bem como técnicas de anonimização de dados.

Por fim, será apresentado um resumo das técnicas de segurança da informação mais importantes em *Big Data*, enfatizando as suas vantagens e desvantagens, e discutido algumas considerações importantes sobre a implementação dessas técnicas em vários ambientes de *Big Data*.

Em conclusão, este capítulo fornece uma visão geral dos principais conceitos e técnicas associados à segurança da informação de Big Data.

3.1 Conceitos de Big Data

Embora a noção de *Big Data* seja relativamente nova, o início dos conjuntos de *data sets* massivos começou nos anos 60 e 70, quando o mundo dos grandes dados de informação estava apenas a emergir com a construção dos primeiros instalações de base de dados relacionais.[2]. Big Data são definidos como dados com uma maior diversidade, vindo em quantidades crescentes e a um ritmo mais rápido. Isto é referido como os três Vs:

- Volume: As organizações adquirem dados de várias fontes, incluindo como transacções, dispositivos da Internet das Coisas (IoT), equipamento industrial, vídeos, fotos,

áudio, e meios de comunicação social. No passado, teria sido demasiado dispendioso armazenar todos estes dados, mas opções de armazenamento mais económicas, tais como Data Lake, Hadoop, e a nuvem, aliviaram esta carga;

- Velocidade: Com a expansão da Internet das Coisas, as organizações devem processar os dados a um ritmo sem precedentes e de forma atempada. As etiquetas RFID, os sensores e os contadores inteligentes necessitam de lidar com estes dados em tempo quase real;
- Variedade: Dados estruturados e quantitativos em bases de dados convencionais coexistem com documentos de texto não estruturados, e-mails, vídeos, ficheiros de áudio e transações financeiras;

Vários outros V's foram reconhecidos para além dos três V's de *Big Data* (volume, velocidade, e variedade). Cada um destes aspetos é essencial para a compreensão, interpretação, e análise de grandes volumes de dados [3][4]:

- Veracidade: Este termo está relacionado com a qualidade e precisão dos dados. A importância da veracidade deriva do facto de que o valor *Big Data* está dependente da qualidade e precisão dos dados;
- Valor: O valor do negócio que pode ser obtido a partir dos dados. O objetivo final de *Big Data* é extrair perceções relevantes e fazer escolhas baseadas em dados e produzir valor empresarial;
- Variabilidade: Está relacionada com a falta de fiabilidade dos dados e a sua alterabilidade ao longo do tempo;
- Complexidade: Refere-se à dificuldade em processar e compreender os dados como resultado da sua quantidade, estrutura e variabilidade;
- Escalabilidade: Refere-se à capacidade de gerir grandes volumes de dados e de se expandir à medida que o volume de dados aumenta;
- Agilidade: Refere-se à capacidade de se ajustar rapidamente às mudanças de dados e às necessidades comerciais;
- Interconectividade: Refere-se à capacidade de ligar dados de muitas fontes e sistemas para fornecer uma perspetiva holística dos dados;

Para além destas qualidades, *Big Data* também implica a utilização de novas tecnologias, incluindo como bases de dados *NoSQL*, computação em *Cloud* e técnicas de *Machine Learning*, para gerir, processar e analisar volumes massivos de dados [5]. O conceito de *Big Data* continua a se expandir à medida que novas tecnologias e aplicações surgem, no entanto, esses 10 critérios fornecem uma compreensão sólida do que *Big Data* envolve.

Elevados volumes de dados referem-se às grandes quantidades de informação geradas todos os dias a partir de uma variedade de fontes, incluindo redes sociais, sensores e transações financeiras, entre outras. Estas informações são frequentemente complexas e não estruturadas, o que dificulta o processamento e a análise tradicionais.

Para processar, analisar e dar sentido a sistemas de *Big Data*, são necessárias tecnologias computacionais avançadas, como a aprendizagem automática e a inteligência artificial. Os sistemas de *Big Data* podem ser utilizados com estas tecnologias para obter informações

sobre o comportamento dos consumidores, otimizar os processos empresariais e até mesmo evitar a má conduta.

No entanto, com tantos dados valiosos recolhidos e armazenados, a segurança torna-se uma grande preocupação. É aqui que reside a importância da encriptação.

A encriptação é o processo de transformação da informação num formato ininteligível para impedir o acesso não autorizado. Garante que os dados sensíveis permanecem seguros e confidenciais, mesmo no caso de uma violação de dados.

À medida que o volume de dados continua a aumentar, a utilização de tecnologias de encriptação para proteger estes dados sensíveis torna-se cada vez mais crucial. As organizações que trabalham com grandes quantidades de dados são obrigadas a implementar soluções de encriptação para salvaguardar os seus dados e cumprir os regulamentos de privacidade de dados.

Em conclusão, a encriptação e sistemas de *Big Data* estão intimamente relacionados. A encriptação protege os dados sensíveis em conjuntos de dados maciços, assegurando que estes permanecem privados e a salvo de potenciais ameaças.

3.1.1 Processamento de dados em Big Data

O processamento de grandes volumes de dados é um tema importante em muitas áreas, incluindo a informática, a engenharia de software e a ciência dos dados. O processamento de quantidades significativas de dados pode ser efetuado através de várias fases que devem ser compreendidas para garantir a eficácia e a qualidade dos resultados.

Gandomi e Haider [3] descobriram na sua investigação que as principais fases do processamento de *Big Data* são a aquisição de dados, o armazenamento de dados, o pré-processamento, a análise de dados e a visualização de dados . Cada uma destas fases apresenta obstáculos únicos e necessita da aplicação de métodos e recursos específicos para ser concluída com êxito.

A recolha de dados é a primeira fase do processamento de grandes volumes de dados. Os dados podem ser recolhidos de várias fontes, como sensores, dispositivos móveis e redes sociais. Para garantir o rigor e a qualidade dos resultados, é essencial escolher as fontes de dados adequadas e recolher uma amostra representativa do universo em análise.

A segunda fase é o armazenamento dos dados. Para gerir grandes volumes de dados, são necessários sistemas de armazenamento distribuído, como o Hadoop Distributed File System (HDFS) ou o Amazon S3 [4]. Estes sistemas permitem o armazenamento eficiente e escalável dos dados, possibilitando aos utilizadores o seu acesso e processamento de forma rápida e eficiente.

Após a recolha e o armazenamento dos dados, estes devem ser submetidos a um pré-processamento, que inclui a limpeza, a transformação e a integração dos dados. O pré-processamento envolve a remoção de dados duplicados, a retificação de erros tipográficos e a normalização de dados em vários formatos. A limpeza dos dados é necessária para garantir a veracidade dos resultados obtidos.

A quarta fase do processamento de Big Data é a análise de dados. Nesta fase, são utilizadas várias técnicas de extração de dados, aprendizagem automática, análise estatística e outras técnicas para identificar padrões, tendências e conhecimentos relevantes nos dados [4]. A

análise de dados é possível utilizando aplicações e plataformas específicas, como o *Apache Spark* ou o *RStudio*.

A visualização de dados é o passo final no processamento de *Big Data*. Nesta fase, os resultados da análise de dados são apresentados aos utilizadores finais de uma forma simples e compreensível.

A tabela 3.1 resume as várias fases do processamento de *Big Data*, incluindo as suas descrições [6].

Tabela 3.1: Etapas de processamento de dados em Big Data

Etapa	**Descrição**
1. Recolha de Dados	Aquisição de dados brutos de várias fontes, como sensores, dispositivos de IoT, base de dados, redes sociais, entre outros.
2. Armazenamento	Armazenamento dos dados brutos em um sistema de base de dados (DBMS) ou em um sistema de arquivos distribuído (DFS).
3. Pré-processamento	Limpeza, filtragem, transformação e integração dos dados brutos para torná-los mais adequados para análise.
4. Análise	Utilização de técnicas de análise de dados, como mineração de dados, Machine Learning, estatística, entre outras, para extrair informações úteis dos dados.
5. Visualização	Representação gráfica das informações obtidas na etapa de análise, tornando-as mais compreensíveis e acessíveis aos utilizadores finais.
6. Tomada de decisão	Utilização das informações obtidas na análise para tomar decisões informadas.

Estas são as etapas fundamentais do procedimento de tratamento de dados de sistemas de *Big Data*, que podem ser alteradas em função das necessidades e dos objetivos de cada projeto. Cada fase é essencial para garantir que os dados são geridos corretamente e que deles são extraídas informações valiosas.

3.1.2 Ferramentas de processamento e análise de dados em sistemas de Big Data

A análise de *Big Data* tem vindo a ganhar importância para as empresas e organizações, uma vez que permite a extração de valiosos conhecimentos comerciais e operacionais. No entanto, o processamento e a análise de grandes volumes de dados apresentam desafios significativos que exigem a aplicação de ferramentas especializadas. Este capítulo aborda algumas das técnicas e instrumentos mais importantes para o processamento e análise de Big Data.

Apache Hadoop

O Apache Hadoop é uma plataforma distribuída de processamento de *Big Data* que permite o armazenamento e o processamento de enormes conjuntos de dados em clusters de computadores. Baseia-se no paradigma de programação *MapReduce*, que separa o processamento

em etapas de mapeamento e redução. A escalabilidade e adaptabilidade do Hadoop fazem dele uma das ferramentas mais utilizadas em todos os sectores de atividade [7].

Apache Spark

O Apache Spark é uma ferramenta de processamento de dados em tempo real que facilita a análise e o processamento eficientes de grandes volumes de dados. Baseia-se num modelo de programação para processamento na memória, o que o torna substancialmente mais rápido do que o Hadoop para determinadas aplicações. O Spark é frequentemente utilizado em casos de utilização como a análise de dados em tempo real, a aprendizagem automática e o processamento de fluxos de dados [7].

Apache Flink

O Apache Flink é uma ferramenta para processar dados em tempo real. Usando uma API, os utilizadores podem escrever programas que processam dados em tempo real e em massa. A tolerância a defeitos e a escalabilidade do Flink fazem dele um instrumento popular para aplicações de *Big Data* de missão crítica [7].

Apache Kafka

O Apache Kafka é uma plataforma de *streaming* distribuído que permite o processamento de fluxos de dados em tempo real. É frequentemente utilizado para a ingestão e transmissão de dados em tempo real entre aplicações ou sistemas. A tolerância a falhas e a escalabilidade do Kafka fazem dele um instrumento popular para aplicações de *Big Data* [7].

A tabela 3.2 apresenta algumas das principais ferramentas de processamento e análise de dados em sistemas de *Big Data*, suas respectivas funcionalidades e algumas empresas que as utilizam

Tabela 3.2: Ferramentas de processamento

Ferramenta	**Funcionalidades**	**Empresas que utilizam**
Apache Hadoop	Armazenamento e processamento distribuído de Big Data utilizando modelo MapReduce	Amazon, Facebook, IBM
Apache Spark	Processamento de dados em tempo real e análise de grandes volumes de dados utilizando modelo de processamento em memória	Airbnb, Netflix, Uber
Apache Flink	Processamento de dados em tempo real e em lote utilizando uma única API	Alibaba, Lyft, Zalando
Apache Kafka	Plataforma de streaming distribuída para ingestão e entrega de dados em tempo real	LinkedIn, PayPal, Uber

A escolha da ferramenta adequada dependerá das necessidades e requisitos específicos de cada projeto.

3.1.3 Importância da Segurança da Informação em Big Data

Com o volume cada vez maior de dados gerados, processados e armazenados nos sistemas de *Big Data*, a segurança da informação tornou-se uma questão de extrema importância. Para garantir a privacidade das informações pessoais, a confidencialidade dos dados comerciais e a integridade dos dados científicos, é essencial proteger os dados armazenados. As intrusões na segurança e privacidade dos dados podem resultar em perdas monetárias, danos na reputação da empresa, perda de confiança dos consumidores e violações legais.

Perigos da segurança da informação na era do Big Data

As ameaças à segurança da informação nos sistemas de *Big Data* são cada vez mais complexas e diversificadas. Ataques de *hackers*, roubo de dados, acesso não autorizado e exploração de vulnerabilidades de software são alguns exemplos. Além disso, o carácter distribuído e descentralizado dos sistemas de *Big Data* pode tornar a segurança ainda mais difícil, uma vez que existem vários pontos de entrada potenciais para ataques.

Medidas e técnicas de segurança para Big Data

Várias técnicas e precauções de segurança podem ser implementadas nos sistemas de *Big Data* para salvaguardar os dados armazenados e garantir a confidencialidade das informações de identificação pessoal. Algumas dessas estratégias incluem:

A encriptação, que é uma técnica que codifica os dados para que só possam ser decifrados por partes autorizadas. Há uma variedade de algoritmos de encriptação que podem ser utilizados para proteger os dados em sistemas de *Big Data* [8].

A anonimização, que é uma técnica que envolve a remoção de informações de identificação pessoal dos dados [8]. Esta técnica pode ser utilizada para proteger a privacidade de dados sensíveis, especialmente em sistemas de *Big Data*.

Controlo do acesso, que é uma medida de segurança que restringe o acesso aos dados apenas àqueles que estão autorizados. Ao implementar sistemas de autenticação e autorização, este objetivo pode ser alcançado [8].

A monitorização da segurança implica a análise contínua dos sistemas de *Big Data* para identificar potenciais perigos e intrusões. Isto pode ser conseguido através da implementação de ferramentas de deteção de intrusões e da análise de registos.

As ferramentas de processamento e análise de dados dos sistemas de *Big Data* são indispensáveis para a tomada de decisões baseadas em dados por parte das empresas e organizações. No entanto, existem muitos outros instrumentos disponíveis.

3.2 Encriptação

A encriptação é o processo de transformação do texto num formato que é difícil de decifrar para as pessoas não autorizadas. O texto codificado é muitas vezes conhecido como *ciphertext* e só pode ser desencriptado, ou convertido de volta à sua forma original, utilizando uma

chave ou palavra-chave secreta [1]. A encriptação é uma componente essencial e é utilizada para salvaguardar dados sensíveis tais como detalhes de cartões de crédito, transações financeiras e informações pessoais [9] .

Big Data é mais difícil de gerir devido à sua quantidade e complexidade. A segurança dos recursos de dados é essencial para qualquer empresa. Hoje em dia, a segurança é de grande importância, uma vez que existem enormes quantidades de volume de dados e por vezes informações pessoais tanto de potenciais consumidores como de uma corporação, fazendo dela um alvo altamente tentador para um atacante [10] .

Uma boa gestão de dados inclui o armazenamento e transporte seguro de dados. A segurança preocupa-se com a confidencialidade, disponibilidade e integridade [6]. Diminuindo a possibilidade de dados serem roubados ou atacados utilizando sistemas informáticos. A implementação de medidas de segurança para impedir o acesso a dados sensíveis por parte de utilizadores não autorizados.

3.3 Anonimização

A anonimização é um método para proteger a privacidade das pessoas através da eliminação de informação de identificação pessoal das coleções de dados. A anonimização é utilizada para preservar a privacidade das pessoas cujos dados estão a ser recolhidos e processados no contexto de *Big Data* [11] . Quando os conjuntos de dados incluem informações sensíveis, tais como dados pessoais, informações financeiras, e registos de saúde, a anonimização torna-se necessária. A anonimização protege a privacidade das pessoas, impedindo que informações sensíveis sejam utilizadas para as identificar [11].

Numerosos sectores, incluindo os cuidados de saúde, o banco e as redes sociais, utilizam regularmente métodos de anonimização para proteger a privacidade das pessoas, permitindo em simultaneo a análise dos dados subjacentes. A anonimização nos cuidados de saúde permite aos investigadores analisar os dados médicos, respeitando ao mesmo tempo, a confidencialidade dos doentes. Na indústria financeira, a anonimização preserva a privacidade dos clientes, facilitando a análise dos dados financeiros pelas instituições financeiras. Da mesma forma, a anonimização protege a privacidade dos utilizadores nas redes sociais, ao mesmo tempo que permite às empresas de redes sociais examinar os dados dos utilizadores [12].

Para preservar a privacidade das pessoas em sistemas de *Big Data*, foram desenvolvidos vários métodos de anonimização. Entre estas estratégias encontram-se o *k-anonymity*, *L-diversity*, *T-closeness*, e a *Differential privacy*[13]. O *k-anonymity* garante que cada registo de uma recolha de dados não pode ser distinguido de pelo menos k-1 outros registos no mesmo conjunto de dados [13]. A *L-diversity* garante que cada característica sensível numa recolha de dados tem pelo menos l valores diferentes. *T-closeness* garante que a distribuição das qualidades sensíveis numa recolha de dados e a distribuição destes atributos na população não são substancialmente diferentes. A *Differential privacy* acrescenta ruído aleatório aos dados, a fim de evitar a identificação dos sujeitos dos dados [13].

As técnicas de anonimização têm demonstrado ser eficazes na manutenção da privacidade em *Big Data*. Por exemplo, El Emam [12] realizaram um estudo que revelou que os métodos de anonimização podem ser utilizados para proteger a privacidade dos pacientes em conjuntos de dados médicos, ao mesmo tempo que permitem aos investigadores realizar estudo dos dados.

Em conclusão, a anonimização é um método crucial para salvaguardar a privacidade das pessoas em sistemas de dados de grande escala. Diversas estratégias de anonimização foram desenvolvidas e demonstraram ser bem sucedidas, permitindo ao mesmo tempo que investigadores e analistas estudem sistemas com elevado número de informações[12][13].

3.4 Encriptação vs Anonimização

Big Data e anonimização são dois conceitos essenciais de gestão de dados que são frequentemente utilizados em conjunto. Apesar de ambos os conceitos estarem relacionados com a privacidade e a segurança dos dados, têm objetivos e metodologias distintos.

Como discutido anteriormente, *Big Data* refere-se aos vastos volumes de dados gerados por diversas fontes que são demasiado complexos e não estruturados para serem analisados utilizando técnicas convencionais. Em contrapartida, a anonimização é o processo de remoção de informações pessoais dos dados para salvaguardar a privacidade dos indivíduos e cumprir os regulamentos de proteção de dados.

Embora a anonimização possa ajudar na proteção de dados sensíveis em grandes conjuntos de dados, não é o mesmo que encriptação. A encriptação é o processo de transformação de dados num formato ininteligível para impedir o acesso não autorizado.

A encriptação é uma medida de segurança de dados crucial, especialmente quando os dados estão a ser transmitidos ou armazenados. Ao encriptar os dados, as empresas podem garantir que as informações sensíveis permanecem seguras e confidenciais.

Ao eliminar as informações pessoais, a anonimização protege a privacidade dos indivíduos em grandes conjuntos de dados. É essencial notar, no entanto, que a anonimização não é à prova de falhas e, por vezes, é possível voltar a identificar indivíduos a partir de dados anónimos utilizando técnicas sofisticadas.

Capítulo 4

Técnicas de Encriptação e Anonimização

Na era digital atual, a utilização de grandes sistemas de dados está a aumentar numa variedade de sectores, incluindo os cuidados de saúde, o banco, e outros. Apesar do facto de que estes sistemas podem dar importantes perspetivas sobre enormes conjuntos de dados, se os dados que armazenam não forem salvaguardados com segurança, podem colocar sérios problemas de privacidade e segurança. As técnicas de criptografia e anonimização oferecem um meio de garantir a preservação da privacidade e segurança de dados sensíveis, enquanto ainda possibilitam a realização de análises e estudos.

Há vários algoritmos de encriptação e anonimização disponíveis para sistemas de dados de grande escala, cada um com as suas próprias vantagens e desvantagens. A *Homomorphic encryption*, por exemplo, permite cálculos sobre dados encriptados ao mesmo tempo que mantém a privacidade e permite a análise. O *K-anonymity* e a *L-diversity* são métodos de anonimato que alteram os dados para garantir que os indivíduos não possam ser reconhecidos a partir do conjunto de dados, enquanto que a *Differential privacy* protege a privacidade individual ao adicionar ruído aleatório a uma recolha de dados [14].

A computação multipartidária segura "*Secure multi-party computation (MPC)*"é uma tecnologia que permite a muitas partes fazer cálculos sobre dados sensíveis sem revelar os dados em si, tornando-a uma opção perfeita para circunstâncias em que a privacidade dos dados é crucial [14]. A máscara "*Masking*"inclui a adição de ruído aleatório aos dados para evitar a identificação individual[14], e a Criptografia Totalmente Homomórfica "*Fully Homomorphic Encryption (FHE)*"permite a realização de cálculos diretamente sobre dados encriptados, oferecendo um grau de privacidade ainda maior do que a encriptação homomórfica [14].

A compreensão das diferentes estratégias de encriptação e anonimização disponíveis para os sistemas de *Big Data* é vital para a proteção de dados sensíveis neste ambiente. É possível explorar todo o potencial de *Big Data*, desde que sejam utilizadas as ferramentas apropriadas para preservar simultaneamente a privacidade e a segurança individual.

4.1 Encriptação Homomórfica

A Encriptação Homomórfica ("*Homomorphic encryption*") é um tipo de encriptação que permite fazer cálculos sobre dados encriptados sem primeiro os decifrar. Isto é conseguido através da utilização de métodos matemáticos que permitem a realização de cálculos diretamente sobre dados encriptados, fornecendo um resultado encriptado que pode ser descodificado para recuperar o resultado do cálculo [14].

Em contraste com os métodos típicos de encriptação, que necessitam que os dados sejam descodificados antes que os cálculos possam ser efetuados, este método não necessita de descodificação. A encriptação homomórfica oferece um elevado grau de privacidade e segurança para dados sensíveis, uma vez que os cálculos podem ser efetuados em dados encriptados sem nunca revelar o texto da placa.

No sector da saúde, a Encriptação Homomórfica é utilizada para salvaguardar os dados dos pacientes, permitindo simultaneamente a análise e o estudo. Por exemplo, a encriptação homomórfica tem sido utilizada para salvaguardar registos médicos e dados genéticos, permitindo a análise e o estudo sem comprometer a privacidade dos pacientes [14]. A Encriptação Homomórfica também pode ser utilizada para salvaguardar dados financeiros sensíveis, tais como números de cartões de crédito e registos de transações, nos sistemas financeiros. A encriptação homomórfica permite cálculos seguros sobre estes dados, incluindo deteção de fraude e avaliação de risco, sem pôr em risco a privacidade de clientes individuais [14].

Para além do comércio electrónico a *Homomorphic encryption* tem utilizações potenciais noutras áreas, tais como a *Cloud*. A *Homomorphic encryption*, por exemplo, pode ser utilizada para proteger os dados dos utilizadores em aplicações baseadas na *Cloud*, permitindo que os cálculos sejam feitos sem sacrificar a privacidade.

4.1.1 Computação Multipartidária Segura

Computação Multipartidária Segura "*Secure Multiparty Computation (SMC)*"é um mecanismo criptográfico que permite a muitos participantes calcular em conjunto uma função sobre as suas entradas sem revelar as suas próprias entradas. O SMC é um método que protege a privacidade dos inputs de todos os participantes, permitindo-lhes ao mesmo tempo, calcular um resultado que depende dos seus inputs [15].

A SMC é alcançado por um protocolo seguro e distribuído que permite às partes calcular em colaboração o resultado desejado sem expor mais informação sobre os seus inputs do que é necessário para o cálculo [15]. Isto é conseguido através de um procedimento criptográfico sofisticado que inclui frequentemente muitas operações tais como encriptação, desencriptação e Computação Multipartidária Segura.

A SMC tem aplicações numa variedade de áreas onde muitas pessoas devem colaborar para calcular um resultado, mantendo a confidencialidade das suas entradas. Estes setores incluem os cuidados de saúde, finanças, e análise de dados, entre outros.

Na área dos cuidados de saúde, por exemplo, a SMC pode ser utilizado para permitir que várias instituições cooperem na análise de dados médicos, protegendo ao mesmo tempo, a privacidade dos pacientes [15]. Da mesma forma, a SMC pode ser utilizada em finanças para leilões seguros ou transações financeiras entre numerosas partes sem a exigência de uma terceira parte de confiança.

Em termos de eficiência, escalabilidade e usabilidade, a SMC é ainda uma tecnologia relativamente nova e sofisticada, e existem ainda muitos obstáculos a ultrapassar. No entanto, a SMC tem o potencial de fornecer análise e computação segura e colaborativa de dados numa vasta gama de aplicações onde a privacidade é primordial.

A SMC é aplicável a uma variedade de aplicações, incluindo análise segura de dados, Machine Learning que preservam a privacidade, e leilões seguros. Nestas situações, a SMC permite a muitos participantes calcular um resultado em colaboração, mantendo ao mesmo tempo, a confidencialidade dos seus contributos.

Utilizando a SMC, por exemplo, muitos profissionais de saúde podem cooperar com segurança na análise de dados médicos sem revelar informações sensíveis sobre doentes. Pode também ser utilizado em aplicações financeiras, tais como leilões seguros, nos quais muitos licitantes podem calcular em colaboração o resultado do leilão sem revelar as suas ofertas individuais [15].

4.1.2 Privacidade diferencial

A abordagem de Privacidade Diferencial é uma estratégia aplicada à análise de dados com o objetivo de proteger as informações individuais. Isso é alcançado ao introduzir uma quantidade controlada de ruído nos dados, tornando difícil identificar as informações de um indivíduo no conjunto de dados, enquanto ainda se preserva a relevância das informações sobre o conjunto de dados como um todo [3].

Dentro do contexto da Privacidade Diferencial, os dados passam por um processo que garante que os resultados obtidos não revelem informações significativas sobre os dados de um indivíduo. O controle da quantidade de ruído adicionada aos dados é conhecido como orçamento de privacidade [3], o qual especifica a quantidade máxima de privacidade que pode ser comprometida durante a análise dos dados.

Essa abordagem encontra aplicações em diversos setores que lidam com informações sensíveis, como saúde, finanças e ciências sociais. Por exemplo, pode ser usada para avaliar dados médicos, protegendo a identidade dos pacientes, ou para analisar informações financeiras, preservando a privacidade dos indivíduos [3].

A estratégia de Privacidade Diferencial mostra-se promissora para proteger a privacidade individual na análise de dados e possui um amplo potencial de adoção em áreas como saúde, economia e ciências sociais.

4.1.3 K-anonymity

A *K-anonymity* é uma abordagem de privacidade de dados utilizada para ocultar as identidades das pessoas dentro de um conjunto de dados. O objetivo do *K-anonymity* é garantir que nenhuma pessoa de um conjunto de dados possa ser reconhecida por uma combinação de k-1 outras pessoas. Em outras palavras, os dados são transformados ou disfarçados de tal forma que um potencial atacante não consegue distinguir entre indivíduos que compartilham o mesmo conjunto de características [16].

A *K-anonymity* é aplicável numa variedade de contextos, tais como cuidados de saúde, bancos, e investigação social. Por exemplo, pode ser utilizado para salvaguardar a privacidade da informação médica ou para assegurar que as pessoas não possam ser reconhecidas ao analisar dados financeiros [16].

A capacidade do *K-anonymity* para dar fortes garantias de privacidade, a sua aplicabilidade a uma grande variedade de tipos de dados e metodologias analíticas, e a sua relativa simplicidade de implementação em comparação com outras abordagens de privacidade estão entre as suas características principais [16]. No entanto, um dos aspetos mais difíceis de estabelecer com o *K-anonymity* é conseguir o equilíbrio adequado entre o valor dos dados e a proteção da privacidade.

4.1.4 L-diversity

O *L-diversity* é um método para proteger a privacidade dos dados sensíveis, garantindo que cada valor de atributo sensível num conjunto de dados seja representado por pelo menos L diferentes valores não sensíveis [16]. Este método visa impedir que um atacante identifique pessoas com base nas suas qualidades sensíveis, aumentando a variedade das características não sensíveis do conjunto de dados.

Pode ser utilizada em vários contextos, incluindo os cuidados de saúde e as finanças. Por exemplo, pode ser utilizada para salvaguardar a privacidade dos registos médicos ou informações financeiras, ou para avaliar dados sobre os interesses e atividades das pessoas sem revelar a sua identidade [16].

O *L-diversity* dá maiores garantias de privacidade do que o *K-anonymity*, uma vez que tem em conta não só o número de pessoas com características idênticas, mas também a variedade de atributos não sensíveis associados a cada valor de atributo sensível. Encontrar o valor adequado de L que estabeleça um equilíbrio entre privacidade e utilidade dos dados é, contudo, um dos aspectos mais difíceis de estabelecer a *L-diversity* [16].

4.1.5 T-closeness

O *T-closeness* é um método para proteger a privacidade de dados sensíveis, garantindo que a distribuição de uma característica sensível dentro de um conjunto de dados esteja "próxima"da sua distribuição na população em geral [17]. A confidencialidade em T destina-se a evitar que um agressor identifique pessoas com base nas suas qualidades sensíveis, tornando a distribuição destes atributos mais semelhante à de toda a população.

Este método oferece maiores garantias de privacidade do que o *K-anonymity*, uma vez que tem em conta não só o número de pessoas com as mesmas características, mas também a distribuição de atributos sensíveis no conjunto de dados [17]. O *T-closeness*, tal como outras estratégias de privacidade, tem as suas próprias limitações e contrapartidas, tais como a dificuldade de obter tanto a privacidade como a utilidade dos dados e o efeito da *T-closeness* na precisão e exatidão estatística.

4.1.6 Fully Homomorphic Encryption (FHE)

A *Fully Homomorphic Encryption (FHE)* é um método de encriptação que não exige a desencriptação para efetuar cálculos em dados encriptados. Quando comparado com os métodos de encriptação padrão, que requerem a desencriptação dos dados antes do cálculo, trata-se de uma melhoria significativa.

Com a FHE, os cálculos podem ser efectuados dentro do próprio texto encriptado, protegendo os dados de serem desencriptados. A proteção da confidencialidade de informações pessoais, como registos médicos e financeiros, pode beneficiar muito com isto [18].

Dada ainda estar numa fase bastante inicial e o elevado custo de computação, a FHE é atualmente impraticável para a maioria das utilizações no mundo real. No entanto, estudos recentes têm-se concentrado em melhorar a eficácia e a eficiência da FHE [18]. Outra investigação centra-se na otimização da utilização da FHE em domínios como a aprendizagem automática.

As potenciais aplicações da FHE incluem a proteção de informações sensíveis durante a *Cloud*, o armazenamento de dados e *Machine Learning* [18].

4.1.7 Máscara

A Máscara "*Masking*"é um método utilizado na criptografia para evitar que as pessoas com más intenções interceptem dados sensíveis. Funciona adicionando uma máscara aleatória aos dados transmitidos, tornando muito mais difícil para os atacantes extrair os dados reais [19]. Em sistemas criptográficos que exigem elevados graus de segurança, tais como aplicações bancárias e militares, a utilização de Máscara é frequentemente utilizada.

A Máscara é a técnica de acrescentar um valor aleatório aos dados que estão a ser encriptados no contexto da criptografia. Este valor aleatório, ou Máscara, é selecionado de tal forma que, quando os dados são desencriptados, se equilibra com outra máscara [19]. A utilização de uma máscara aleatória distinta cada vez que os dados são encriptados torna muito mais difícil para os atacantes detetar os dados originais.

Nos últimos anos, a Máscara tem sido o foco de um estudo substancial, com vários artigos que abordam a sua eficácia e usos prospetivos. O desenvolvimento de métodos de criptografia de máscara melhorados e mais resistentes a ataques tem sido um importante tema de estudo. Por exemplo, os académicos têm sugerido novos métodos para produzir as máscaras aleatórias utilizadas em esquemas de mascaramento, tais como o mascaramento baseado no ruído, que podem aumentar a aleatoriedade e imprevisibilidade das Máscara [19].

A utilização de Máscara para vários tipos de sistemas criptográficos também tem sido objeto de investigação. Um exemplo notável envolve a exploração de combinações de técnicas, como o uso de máscaras juntamente com a Encriptação Homomórfica, com o objetivo de aprimorar a segurança de sistemas criptográficos e garantir a preservação da privacidade.

Globalmente, a utilização de Máscara é um método eficaz para evitar que os atores hostis interceptem e descodifiquem informações sensíveis. Embora o Máscara não seja uma solução perfeita e tenha limites, a investigação contínua procura melhorar e expandir a sua aplicação numa vasta gama de sistemas criptográficos.

4.2 Comparação de técnicas de criptografia avançada para garantir a privacidade em ambientes de Big Data

Recentemente, temos observado um aumento significativo no uso de grandes volumes de dados em diversos setores. A recolha e o processamento de grandes volumes de dados sensíveis têm suscitado preocupações significativas em relação à confidencialidade e à segurança. A utilização de métodos de encriptação pode ser uma abordagem viável para salvaguardar a confidencialidade dos dados. No entanto, as técnicas de encriptação convencionais podem não estar equipadas para gerir as complexidades das configurações de *Big Data*. Este subcapítulo tem por objetivo efetuar uma análise comparativa de três metodologias de encriptação sofisticadas, nomeadamente a encriptação homomórfica, a encriptação totalmente homomórfica (FHE) e a computação segura multipartidária (SMPC).

A encriptação homomórfica é uma técnica criptográfica que permite a execução de operações matemáticas em dados encriptados, sem necessidade de desencriptação prévia dos dados. Esta metodologia é particularmente vantajosa em cenários em que a informação necessita de ser analisada por uma entidade externa, mantendo a confidencialidade. A eficiência da cifragem homomórfica é limitada, o que torna a abordagem impraticável para o processamento de grandes volumes de dados.

O conceito de Encriptação Totalmente Homomórfica (FHE) é uma expansão da Encriptação Homomórfica, que permite a execução de cálculos sem restrições no texto cifrado. Este atributo faz da FHE um mecanismo potente para salvaguardar a privacidade durante a análise de dados. A Encriptação Totalmente Homomórfica (FHE) tem a capacidade de facilitar a delegação segura de responsabilidades de análise de dados a entidades terceiras não fiáveis, sem necessidade de desencriptação dos dados. No entanto, a carga computacional associada à Encriptação Homomórfica Total (FHE) é considerável, o que coloca desafios práticos ao processamento de dados em grande escala.

A técnica de computação segura multipartidária (SMPC) é uma abordagem de análise de dados que permite que várias partes executem uma função de forma colaborativa nos seus respectivos dados privados, assegurando simultaneamente que os seus dados não são revelados umas às outras. A computação segura multipartidária (SMPC) é uma ferramenta valiosa em cenários em que a informação está dispersa por várias entidades e não pode ser consolidada. A Computação Multipartidária Segura (SMPC) tem o potencial de facilitar a delegação segura de tarefas de análise de dados, distribuindo a computação por várias partes, mitigando assim o risco de um único ponto de falha. No entanto, a computação segura multipartidária (SMPC) pode implicar custos computacionais significativos, especialmente quando o número de partes envolvidas aumenta.

Em resumo, a encriptação homomórfica é uma opção viável para cálculos básicos que envolvam dados limitados, enquanto a encriptação totalmente homomórfica pode ser mais adequada para cálculos complexos que envolvam conjuntos de dados extensos. Por outro lado, a computação segura multipartidária (SMPC) é a escolha ideal para cenários em que os dados estão dispersos por várias partes. A seleção de uma determinada metodologia depende das exigências precisas da operação de manipulação de dados em questão.

Tabela 4.1: Comparação de técnicas de criptografia de Big Data [7]

Técnica	Pontos fortes	Pontos fracos
Homomorphic Encryption	Permite o cálculo de dados encriptados, preservando a privacidade.	Desempenho lento com grandes conjuntos de dados. Pode não ser prático para o processamento em grande escala.
Fully Homomorphic Encryption (FHE)	Permite cálculos arbitrários sobre o texto cifrado, o que é poderoso para a análise de preservação da privacidade.	Custos computacionais significativos, difícil de utilizar para processamento em grande escala. Necessita de mais investigação.
Secure Multi-Party Computing	Permite a computação conjunta de dados privados sem revelar às partes.	Computacionalmente dispendioso, requer uma implementação cuidadosa para evitar partes maliciosas.

4.3 Comparação de técnicas de anonimização em sistemas de Big Data

A *Differential Privacy* é uma técnica de preservação da privacidade que oferece um elevado grau de proteção contra ataques avançados. A técnica tem o potencial de ser implementada em diversos conjuntos e formatos de dados e oferece uma conceptualização rigorosa da confidencialidade que pode servir de métrica para avaliar a sua eficácia. A introdução de dados estranhos pode potencialmente diminuir a precisão dos resultados, e a calibração da quantidade de dados estranhos a incorporar pode constituir um desafio, exigindo uma avaliação meticulosa dos dados e da sua aplicação pretendida.

Em contrapartida, o *K-Anonymity* oferece uma abordagem direta e eficiente para salvaguardar a privacidade pessoal. A metodologia é altamente compreensível e implementável, e apresenta versatilidade na sua aplicabilidade em diversos conjuntos e formatos de dados. No entanto, o sistema é suscetível a ataques de homogeneidade, através dos quais um adversário pode deduzir dados suplementares sobre uma pessoa aproveitando a semelhança entre indivíduos de um grupo.

O conceito de *L-Diversity* foi concebido para proteger contra o risco de divulgação de atributos, que se refere à possibilidade de deduzir os dados confidenciais de um indivíduo com base na existência de valores específicos num determinado conjunto de dados. O requisito é garantir que um determinado atributo em cada grupo de registos de um conjunto de dados contenha pelo menos L valores únicos. A metodologia supramencionada revela-se vantajosa em conjuntos de dados em que a divulgação de dados confidenciais nem sempre é abertamente reconhecível, mas pode ser deduzida a partir de padrões discerníveis. Apesar dos seus potenciais benefícios, a*L-Diversity* pode ser susceptível a ataques de conhecimento de fundo, em que um agressor possui conhecimento prévio da prevalência de um determinado atributo no conjunto de dados.

A abordagem *T-Closeness* funciona verificando se a distribuição do atributo sensível num determinado conjunto de dados não apresenta um desvio estatisticamente significativo da sua distribuição na população em geral. A abordagem apresenta uma eficácia notável no tratamento de conjuntos de dados em que o atributo de interesse pertence a dados categóricos. No entanto, o *T-Closeness* pode ser suscetível a ataques que utilizam informações auxiliares para deduzir informações sensíveis a partir de dados não sensíveis.

O *Masking* é uma metodologia que implica a substituição de informações confidenciais por dados menos sensíveis ou por dados fictícios. Esta abordagem revela-se vantajosa em situações em que não é possível eliminar informações confidenciais de um conjunto de dados ou quando não são viáveis métodos alternativos de anonimização. O mascaramento é uma técnica que pode ser utilizada para ocultar dados específicos, substituindo-os por um valor genérico ou eliminando-os completamente. As técnicas de mascaramento são susceptíveis de ataques que exploram a existência de correlações entre atributos, permitindo assim a inferência de informações sensíveis.

Em última análise, a seleção de um método de anonimização depende das exigências específicas dos dados e do grau de proteção considerado necessário. É crucial compreender os pontos fortes e as limitações de cada técnica antes de escolher a mais adequada para um determinado cenário.

Tabela 4.2: Comparação de técnicas de anonimização de Big Data

Técnica	Pontos fortes	Pontos fracos
Differential Privacy	Oferece uma forte protecção da privacidade; Pode ser aplicado a vários conjuntos e tipos de dados	Pode reduzir a exactidão dos resultados
K-Anonymity	Protecção da privacidade simples e eficaz	Vulnerável a ataques de homogeneidade
L-Diversity	Protege contra a divulgação de atributos; Útil para conjuntos de dados com informações sensíveis inferidas	Vulnerável a ataques de conhecimento de base
T-Closeness	Assegura que a distribuição do atributo sensível não é significativamente diferente da da população em geral; Particularmente eficaz para conjuntos de dados categóricos	Vulnerável a ataques que utilizam informações auxiliares
Masking	Oculta dados individuais; Útil quando os dados sensíveis não podem ser removidos	Vulnerável a ataques que utilizam correlações entre atributos

4.4 Tecnologias e Frameworks

Várias estratégias e estruturas podem ser utilizadas para implementar os algoritmos de encriptação mencionados. De destacar:

- A biblioteca SEAL: SEAL (Biblioteca Aritmética Encriptada Simples) é uma biblioteca C++ que implementa a *Homomorphic encryption* usando os esquemas BFV e CKKS de uma forma rápida e fácil de usar. A SEAL destina-se ao desenvolvimento de aplicações de *Homomorphic encryption* e foi implementada em várias aplicações [20].
- OpenMined é um projeto de código aberto que oferece uma plataforma para a aprendizagem de máquinas de preservação da privacidade utilizando métodos *Differential privacy* e *Homomorphic encryption*. OpenMined oferece uma interface baseada em Python para implementação de várias técnicas de encriptação, tais como encriptação homomórfica e computação multipartidária segura [21].
- SPDZ é um protocolo para computação multipartidária segura que permite soluções escaláveis e eficientes de análise de dados que preservam a privacidade. O SPDZ utiliza métodos como a partilha secreta e a *Homomorphic encryption* para fornecer computação segura sobre dados dispersos [21].
- PALISADE é uma biblioteca C++ que oferece uma estrutura escalável e flexível para

criação de *Homomorphic encryption* e outros primitivos criptográficos. PALISADE suporta numerosos algoritmos de *Homomorphic encryption*. PALISADE tem sido implementado numa variedade de aplicações, tais como a aprendizagem segura da máquina e o cálculo externalizado seguro [21].

Estas abordagens e estruturas fornecem formas escaláveis e eficientes de implementar vários algoritmos de encriptação em aplicações do mundo real.

4.5 Resumo

A encriptação e a anonimização são cruciais para proteger a privacidade e a segurança de dados sensíveis em sistemas de *Big Data*. Várias técnicas de encriptação, foram desenvolvidas para permitir a computação segura em dados encriptados. Abordagens de anonimização, tais como *K-anonymity*, *L-diversity*, *T-closeness*, e *Differential privacy*, foram também criadas para proteger a privacidade de dados sensíveis, permitindo em simultâneo, a análise de dados.

Vários aspetos, incluindo desempenho, escalabilidade, e usabilidade, devem ser cuidadosamente considerados ao utilizar estas abordagens em sistemas de *Big Data*. Para responder a estas preocupações, várias estruturas e bibliotecas, tais como SEAL, PALISADE, SPDZ, e OpenMined [21], foram propostas para facilitar a implementação eficiente e escalável de métodos de encriptação e anonimização em aplicações do mundo real.

A encriptação e a anonimização são cruciais para proteger a privacidade e a segurança de dados sensíveis em grandes plataformas de dados. Como a prevalência do *Big Data* continua a aumentar, é essencial continuar a desenvolver e aperfeiçoar estas metodologias para acompanhar o ambiente de ameaça em constante mudança e proteger a privacidade e a segurança de dados sensíveis.

Na atual sociedade orientada para os dados, a geração de dados está a acelerar a um ritmo sem precedentes. O crescente quantidade de dados, também conhecido como *Big Data*, oferece tanto oportunidades como desafios para empresas e organizações de todos os setores. Os elevados volumes de dados podem fornecer informações valiosas e ajudar a fomentar a inovação, mas também colocam problemas de privacidade e segurança.

Para abordar essas preocupações, foram criadas diversas técnicas de encriptação e anonimização para salvaguardar informações sensíveis em ambientes de *Big Data*. Estas técnicas foram concebidas para garantir que os dados permanecem seguros e confidenciais apesar dos perigos crescentes e dos desafios de segurança em evolução.

Em sistemas de *Big Data*, estas técnicas de encriptação e anonimização podem ser utilizadas para proteger dados sensíveis. As técnicas de encriptação, como a encriptação homomórfica, a encriptação totalmente homomórfica (FHE) e a computação multipartidária segura, permitem executar cálculos em dados encriptados sem os desencriptar previamente, protegendo assim os dados sensíveis.

Em contrapartida, as técnicas de anonimização, como a privacidade diferencial, o *K-Anonymity*, a *L-Diversity*, a *T-Closeness* e o *Masking*, destinam-se a salvaguardar a privacidade dos indivíduos em grandes conjuntos de dados, removendo informações pessoais ou introduzindo ruído nos dados.

Tabela 4.3: Técnicas de encriptação e anonimização usadas em sistemas de Big data

Técnica	Tema
Homomorphic Encryption	Encriptação
Fully Homomorphic Encryption (FHE)	Encriptação
Secure Multi-Party Computing	Encriptação
Differential Privacy	Anonimização
K-Anonymity	Anonimização
L-Diversity	Anonimização
T-Closeness	Anonimização
Masking	Anonimização

Utilizando uma combinação de técnicas de encriptação e anonimização, é possível proteger dados sensíveis em sistemas com elevado volume de dados e, ao mesmo tempo, obter informações úteis a partir dos dados.

Capítulo 5

Análise de Valor

O objetivo da análise de valor é analisar e maximizar o valor de um produto, processo, ou serviço. Implica uma avaliação exaustiva dos custos e vantagens de várias alternativas para escolher o curso de ação mais eficiente e rentável.

No contexto de sistemas de *Big Data*, a análise de valor identifica os dados mais valiosos e otimiza o custo-benefício da recolha, armazenamento e processamento de dados. Esta estratégia implica ponderar as vantagens potenciais da recolha e processamento de várias formas de dados contra as despesas relacionadas, tais como o armazenamento, processamento e a gestão de dados.

Além disso, a análise de valor pode ser utilizada para melhorar o desempenho e a eficácia dos grandes sistemas de dados. Ao examinar os custos e vantagens de várias configurações de sistemas, as empresas podem escolher a solução mais eficiente e rentável para as suas necessidades particulares.

Em geral, a análise de valor é um método eficaz para maximizar a relação custo-benefício da gestão de dados em grandes sistemas de dados. Ao identificar as fontes de dados mais importantes e otimizar as configurações do sistema, as empresas podem reduzir o custo total da gestão de dados e maximizar as vantagens de *Big Data*.

5.1 Identificação das oportunidades

Na era de *Big Data*, a quantidade de informação pessoal adquirida, processada, e armazenada está a expandir-se drasticamente. Como resultado, é essencial salvaguardar estes dados contra agressões cibernéticas, violações de dados e acesso ilegal. A encriptação e a anonimização surgiram como tecnologias essenciais para salvaguardar a privacidade de dados sensíveis. Existem várias possibilidades de fazer investigação sobre encriptação e abordagens de anonimização para volumes de dados.

Em primeiro lugar, uma investigação sobre estratégias de encriptação e anonimização para *Big Data* oferece a possibilidade de avaliar a eficácia de várias abordagens para vários tipos de dados e casos de utilização. Esta análise pode ajudar investigadores ou qualquer pessoa a fazer uma seleção bem informada sobre as estratégias a utilizar para as suas necessidades específicas. Além disso, a investigação pode identificar quaisquer limites ou dificuldades relacionadas com as metodologias atuais, tais como o compromisso entre privacidade e utilidade dos dados, e investigar potenciais soluções.

Em segundo um estudo de estratégias de encriptação e anonimização de dados de grandes dimensões pode também levar à criação de novas abordagens que estabeleçam um equilíbrio ótimo entre privacidade e utilidade. Com a crescente complexidade dos *Big Data*, há necessidade de sistemas cada vez mais sofisticados que possam gerir grandes quantidades de dados, mantendo ao mesmo tempo, um elevado grau de proteção da privacidade. Um estudo pode ajudar os investigadores a identificar métodos novos e criativos que se ajustem a estes critérios.

Em terceiro lugar, uma investigação sobre métodos de encriptação e anonimização de grandes conjuntos de dados as consequências éticas e legais da utilização de dados encriptados e anonimizados por várias razões. Por exemplo, as leis e regulamentos sobre proteção de dados podem diferir por nação ou área, e é crucial assegurar que estas políticas sejam cumpridas a fim de salvaguardar a informação pessoal. Além disso, a utilização de dados anonimizados pode levantar considerações éticas, tais como o perigo de reidentificação ou a possibilidade de discriminação.

Uma investigação sobre criptografia e estratégias de anonimização de *Big Data* pode dar às empresas e entidades reguladoras informações sobre as melhores práticas de salvaguarda de dados pessoais, utilizando ao mesmo tempo, o potencial dos grandes dados. Este aconselhamento pode ajudar as empresas a fazer escolhas bem informadas sobre como preservar os seus dados enquanto ainda os utilizam para análise e tomada de decisões.

Em conclusão, um estudo sobre estratégias de encriptação e anonimização de *Big Data* proporciona a investigadores e organizações várias perspetivas. Uma investigação pode ajudar a preservar dados pessoais, promover a privacidade e a segurança, e desbloquear todo o potencial das grandes análises de dados, avaliando a eficácia das abordagens existentes, inventando novas técnicas, analisando preocupações éticas e legais, e oferecendo aconselhamento sobre as melhores práticas.

5.2 Análise de oportunidades

Uma investigação sobre métodos de encriptação e anonimização para sistemas com grandes quantidades de dados proporciona várias oportunidades de análise, incluindo pontos fortes, fracos, oportunidades e ameaças. Aqui estão várias oportunidades de análise:

- Pontos fortes: A oportunidade de oferecer uma revisão completa do estado atual da investigação neste campo, é uma das principais vantagens de escrever um artigo científico sobre métodos de encriptação e a anonimização de *Big Data*. Este tipo de publicação pode consolidar os resultados de investigações anteriores e realçar as vantagens das metodologias contemporâneas;
- Pontos fracos: A identificação das fraquezas da literatura atual é outra componente crucial da produção de um relatório científico. Por exemplo, certos estudos podem ter limites no seu âmbito ou metodologia, impedindo-os de fornecer uma análise completa da eficácia da encriptação e das estratégias de anonimização de grandes dados. A identificação destas falhas pode ajudar na conceção de uma investigação mais rigorosa e exaustiva;
- Oportunidades: Uma pesquisa sobre criptografia e estratégias de anonimização de dados em massa dá oportunidades para investigar novas e inovadoras soluções de proteção de dados. Pode haver oportunidades, por exemplo, de criar métodos de

encriptação ou de anonimização que são construídos expressamente para *Big Data* ou poderem melhorar o equilíbrio entre a privacidade e o valor dos dados. Tais avanços podem dar vantagens substanciais às empresas que têm de assegurar dados sensíveis, ao mesmo tempo que os empregam para fins analíticos e de tomada de decisões;

- Ameaças: O desenvolvimento de um estudo académico sobre métodos de encriptação e a anonimização de *Big Data* exige também a análise de possíveis perigos. A possibilidade de violação de dados ou acesso não autorizado, que pode ameaçar a privacidade e segurança dos dados pessoais, é uma das preocupações mais graves. Outro risco é o potencial de reidentificação de dados anonimizados, o que pode comprometer a proteção da privacidade dos dados. Este trabalho deve avaliar estes perigos e fornecer formas de mitigar estes riscos;

Uma investigação sobre abordagens de encriptação e anonimização de *Big Data* proporciona várias oportunidades de análise, incluindo a descoberta de pontos fortes, fraquezas, oportunidades e ameaças. Ao empreender uma investigação exaustiva destes elementos, os investigadores serão capazes de conceber estratégias novas e mais eficazes para proteger os dados sensíveis na era de *Big Data*.

5.3 Valor da solução

Neste capítulo sobre o valor da solução, discutiremos as vantagens e o valor que uma solução sugerida pode proporcionar aos consumidores ou partes interessadas no contexto de métodos de encriptação e anonimização de *Big Data*.

5.3.1 Valor do cliente

A realização de um estudo de investigação sobre métodos de encriptação e a anonimização de *Big Data* pode gerar uma quantidade substancial de valor para o consumidor.

O artigo pode detetar e avaliar a eficácia de diferentes abordagens de encriptação e de anonimização na salvaguarda de dados sensíveis, aumentando assim a segurança dos dados. Isto pode dar aos clientes informações vitais sobre a segurança dos seus dados e ajudá-los a fazer escolhas instruídas relativamente às abordagens a empregar.

Muitos sectores estão sujeitos à legislação sobre privacidade de dados, tais como o RGPD. Ao oferecer uma revisão completa dos métodos de encriptação e anonimização, este documento pode ajudar o utilizador final a garantir o cumprimento destes requisitos e a evitar penalizações dispendiosas.

Ao demonstrar um compromisso com a privacidade e segurança dos dados, o documento pode ajudar o utilizador a ganhar a confiança dos seus interessados. Isto pode melhorar a reputação da organização e promover a lealdade dos clientes.

A segurança e a privacidade dos dados são cruciais para preservar uma vantagem competitiva na economia atual baseada em dados. Utilizando os mais recentes métodos de encriptação e anonimização, as empresas podem obter uma vantagem competitiva e distinguir-se no mercado.

Escrever um artigo científico sobre técnicas de encriptação e anonimização de *Big Data* pode fornecer um valor significativo ao utilizador final em termos de segurança de dados melhorada, maior conformidade com os regulamentos de privacidade de dados, maior confiança

nas práticas de tratamento de dados, e maior tomada de decisões e vantagem competitiva. Este documento pode ajudar o utilizador a navegar no entender melhor os conceitos existentes no mundo de *Big Data* e a garantir a segurança e a privacidade dos seus dados sensíveis, dando informações e sugestões úteis.

5.3.2 Valor Percebido

O valor percebido dos consumidores que leem um artigo académico sobre métodos de codificação e anonimização de *Big Data* baseia-se nas vantagens prospetivas e perceções que podem ser alcançadas:

- Em primeiro lugar, uma vez que a quantidade de dados recolhidos e guardados continua a aumentar drasticamente, os métodos de encriptação e a anonimização de *Big Data* estão a tornar-se mais vitais na era digital moderna. O utilizador final pode obter um melhor conhecimento dos mais recentes avanços na encriptação e anonimização, bem como dos perigos e obstáculos que os acompanham, lendo um artigo de investigação sobre estes temas.
- Em segundo lugar, o estudo pode dar uma visão sobre como implementar procedimentos eficientes de encriptação e de anonimização, que podem ser essenciais para as empresas que trabalham com dados sensíveis. Isto pode ajudar o utilizador a fazer escolhas mais instruídas sobre como salvaguardar os seus dados, e diminuir o risco de violações de segurança ou outras preocupações relacionadas com a segurança.

Além disso, o artigo pode incluir uma avaliação crítica do estado atual dos sistemas de encriptação e de anonimização, indicando áreas que necessitam de ser melhoradas ou mais estudadas. Isto pode ajudar os clientes a escolher as abordagens mais bem sucedidas e contribuir para o possível desenvolvimento futuro do campo. O estudo pode também dar aos leitores uma maior compreensão das ideias técnicas por detrás da encriptação e da anonimização, aumentando assim a sua competência no assunto.

Em conclusão, o valor percebido do utilizador final que leem um artigo científico sobre métodos de cifragem e anonimização de *Big Data* é elevado, uma vez que pode dar importantes perspetivas, aumentar a segurança dos dados, e alargar a sua compreensão deste tópico crucial.

5.4 Implementação de funções de qualidade

Quality Function Deployment (QFD) é um método sistemático de conceção de produtos e serviços que é amplamente utilizado em sectores como o fabrico, engenharia, e desenvolvimento de software. O método QFD implica a conversão das necessidades e desejos do cliente em requisitos técnicos, que são depois utilizados para impulsionar a conceção e desenvolvimento de um produto ou serviço. Neste procedimento, é utilizado um diagrama QFD para representar as ligações entre as necessidades do cliente e as características técnicas necessárias para alcançar esses requisitos.

Num diagrama QFD, as necessidades do utilizador final são apresentadas no lado esquerdo do gráfico, enquanto que as características técnicas são apresentadas em toda a parte superior. As células do gráfico mostram o significado de cada elemento tecnológico na resposta a cada necessidade do utilizador. Esta relevância é frequentemente denotada com uma ponderação numérica, com valores mais elevados a significar mais importância.

Num diagrama de Implementação da Função Qualidade (QFD), a ponderação é o ato de atribuir números quantitativos à relevância relativa de cada componente técnico na consecução de uma necessidade específica do cliente. Por exemplo, se a segurança for um requisito altamente importante do cliente, as características técnicas que contribuem para a segurança (tais como encriptação e controlo de acesso) podem receber maior ponderação no diagrama QFD do que as características técnicas que estão menos diretamente relacionadas com a segurança (tais como facilidade de utilização ou custo) (tais como facilidade de utilização ou custo).

Requisitos do Cliente	Características técnicas	Ponderação
Segurança	Criptografia totalmente homomórfica	9
	Criptografia Homomórfica	7
	Computação multipartidária segura (SMC)	6
Anonimização	Privacidade Diferencial	8
	K-Diversity	7
	L-Diversity	6
Desempenho	T-Closeness	5
	K-Computação multipartidária segura	8
	Máscara	6
Escalabilidade	Computação distribuída	9
Facilidade de utilização	Gestão de chaves	8
	Máscara	6

As ponderações são utilizadas para conduzir o processo de conceção e desenvolvimento, permitindo à equipa de conceção atribuir recursos e esforços com base nos aspectos técnicos mais essenciais para o cliente ou utilizador final. A QFD pode aumentar a satisfação do cliente e a qualidade geral do produto, fazendo corresponder os aspectos técnicos de um produto ou serviço com os requisitos e desejos do consumidor.

Com base neste diagrama QFD, as seguintes metodologias e técnicas podem ser utilizadas numa investigação abrangente das estratégias de encriptação e de anonimização de *Big Data*.

5.5 Resumo

Os métodos de encriptação e anonimização são utilizados para salvaguardar informação sensível e preservar a privacidade durante o processamento e análise de volumes de *Big Data*. Estas estratégias incluem a modificação de dados de uma forma que os torne ilegíveis para as partes não autorizadas, ao mesmo tempo que permite o seu acesso e avaliação pelas partes autorizadas.

Processos importantes, análise de valor e identificação de oportunidades ajudam as empresas a determinar o valor potencial da encriptação e abordagens de anonimização de grandes dados, bem como as vantagens e desvantagens da sua utilização. A análise de oportunidades implica a avaliação das vantagens, desvantagens e perigos ligados à encriptação e abordagens de anonimização para grandes conjuntos de dados.

Ao implementar estratégias de encriptação e anonimização para *Big Data*, o valor do cliente, o valor da solução e o valor percebido são todas considerações essenciais. O valor do cliente refere-se ao valor visto pelos clientes como resultado da implementação destes métodos, enquanto que o valor da solução se refere ao valor total das soluções de encriptação e de

anonimização. O valor percebido relaciona-se com as vantagens percebidas da utilização de procedimentos de encriptação e anonimização.

Ao implementar métodos de encriptação e de anonimização para grandes volumes de dados, é também crucial incorporar funções de qualidade. As funções de qualidade garantem que a implementação de métodos de encriptação e de anonimização cumpre os objetivos e resultados requeridos.

A aplicação de métodos de encriptação e de anonimização para conjuntos massivos de dados requer uma avaliação cuidadosa de uma variedade de aspetos, incluindo o valor potencial, possibilidades, obstáculos, e funções de qualidade ligadas a estas abordagens.

Capítulo 6

Implementação da solução

Neste capítulo, foi efetuado uma investigação exaustiva de uma coleção de metodologias de preservação da privacidade que podem ser utilizadas eficientemente utilizando a linguagem de programação Python. Neste estudo, analisamos e esclarecemos quatro estratégias básicas cruciais que servem como fundamentos significativos no campo da segurança e privacidade de dados. A encriptação homomórfica, a privacidade diferencial, a computação segura multipartidária (SMPC) e o *Masking* são os quatro métodos fundamentais abordados neste capítulo.

O principal objetivo deste estudo é fornecer aos leitores uma compreensão abrangente destas metodologias, permitindo-lhes fazer escolhas bem informadas relativamente à privacidade e segurança dos dados.

6.1 Homorphic Encryption

A utilização da encriptação homomórfica permite a execução de cálculos em dados encriptados sem necessidade de desencriptação. O resultado do cálculo é uma versão encriptada do resultado obtido pela aplicação da mesma operação ao texto original não encriptado. Neste estudo, foram utilizadas duas bibliotecas distintas.

O código python que está descrito na figura A.1 utiliza a biblioteca Paillier, por vezes conhecida como "phe", para ilustrar a noção de encriptação homomórfica. A encriptação homomórfica é uma metodologia criptográfica que facilita a execução de operações matemáticas em dados encriptados, resultando num resultado encriptado que, após a desencriptação, corresponde ao resultado desses mesmos processos realizados nos dados iniciais não encriptados. Esta caraterística permite o cálculo de informações sensíveis, mantendo a confidencialidade dos dados reais.

Após a instalação bem sucedida da biblioteca "phe", o código inicia a sua execução produzindo um par de chaves Paillier, que inclui uma chave pública e a sua chave privada correspondente. A chave pública é utilizada para efeitos de encriptação, enquanto a chave privada é designada apenas para desencriptação. O processo de geração de um par de chaves é da maior importância para permitir uma comunicação e computação seguras através da utilização do sistema Paillier.

Posteriormente, o código fornecido exemplifica um cenário de aplicação fundamental da encriptação homomórfica. A sequência começa com uma enumeração de valores numéricos ocultos, nomeadamente consistindo num único dígito "5". O processo de encriptação envolve a aplicação da chave pública a cada número individual, resultando numa lista de números encriptados designada por `encrypted_number_list`.

O `encrypted_number_list` contém versões encriptadas dos números em `secret_number_list` utilizando o esquema de encriptação de Paillier, tornando os números originais confidenciais e permitindo operações matemáticas sobre eles sem desencriptar.

É efetuada a soma dos números encriptados da lista dada, seguida da adição de um valor constante de '6' ao total encriptado resultante. Apesar de estas ações serem executadas em dados que foram encriptados, o resultado permanece num estado encriptado.

Por último, o valor encriptado é decifrado utilizando a chave privada, resultando em um valor desencriptado. O resultado desencriptado deve corresponder à soma dos números originais não encriptados, acrescidos da constante '6'. Isso evidencia a eficácia da criptografia homomórfica na preservação da privacidade dos dados ao mesmo tempo, em que possibilita cálculos.

Em resumo, o seguinte código oferece um exemplo demonstrativo da utilização da biblioteca Paillier com o objetivo de implementar a encriptação homomórfica. O sistema cria um par de chaves criptográficas, utiliza técnicas de encriptação para proteger dados numéricos, executa operações nos dados encriptados e verifica se o processo de desencriptação produz o resultado previsto. Este facto demonstra a aplicabilidade da abordagem para facilitar os cálculos que preservam a privacidade.

6.2 Seal

A biblioteca SEAL desenvolvida pela Microsoft [22], que pode ser acedida em Python através do módulo PySEAL. Esta biblioteca serve os princípios fundamentais subjacentes à encriptação homomórfica e a sua implementação prática com o objetivo de garantir uma computação segura.

A utilização da encriptação homomórfica permite a execução de operações matemáticas em dados encriptados, preservando assim a confidencialidade da informação subjacente. Esta abordagem mostra-se altamente benéfica em cenários em que a confidencialidade dos dados é essencial, assegurando que a informação subjacente permaneça protegida contra divulgações não autorizadas.

O código da imagem A.2 começa por importar módulos essenciais da biblioteca SEAL, incluindo funcionalidades de encriptação, desencriptação e avaliação de dados encriptados. O último passo consiste em estabelecer as definições de encriptação para o método BFV (Brakerski-Fan-Vercauteren), um sistema de encriptação homomórfica bem conhecido que facilita as operações aritméticas em dados encriptados.

O contexto SEAL é estabelecido utilizando os parâmetros especificados e é produzido um par de chaves através da utilização do KeyGenerator. A chave pública e a chave secreta são derivadas do par de chaves, a fim de facilitar as operações de encriptação e desencriptação, respetivamente.

Subsequentemente, dois valores de texto simples ('5' e '6') são instanciados como objetos Plaintext, denotando os valores numéricos destinados à encriptação. O Encryptor utiliza a chave pública para encriptar os valores de texto simples fornecidos, dando origem a dois textos cifrados.

O Avaliador é utilizado para executar uma operação de adição nos textos cifrados encriptados. A função `evaluator.add_inplace` executa uma operação que combina o primeiro

texto cifrado com o segundo texto cifrado, resultando na representação cifrada da soma de '6' e '5'.

Por fim, o descodificador utiliza a chave secreta para descodificar o texto cifrado, obtendo um resultado em texto simples. O resultado é apresentado, mostrando o total decifrado dos dois valores numéricos codificados, '11'.

Em resumo, este código exemplifica a utilização da biblioteca SEAL da Microsoft, que facilita a encriptação homomórfica, através da sua acessibilidade em Python através do PySEAL. O sistema estabelece definições de encriptação, cria pares de chaves criptográficas, executa uma operação de adição em dados encriptados e mostra a descodificação bem sucedida da saída resultante. Isto sublinha a eficácia da encriptação homomórfica para facilitar a computação segura, mantendo a confidencialidade dos dados.

6.3 Computação Segura Multipartidária

Esta abordagem é particularmente aplicável à Computação Multipartidária Segura (SMPC), uma metodologia que permite que várias entidades efetuem coletivamente cálculos nas suas respetivas entradas, mantendo a privacidade dessas entradas.

O código da imagem A.3 está estruturado da seguinte forma:

1. **Declarações de importação**: O código começa por importar os módulos necessários, como o módulo 'random', para gerar números aleatórios, e o módulo 'functools', que poderá eventualmente ser utilizado para utilitários de programação funcional, mas que não é utilizado no código apresentado.

2. **Definições de Funções**: A função `secret_share` é projetada para aceitar vários parâmetros, incluindo um valor secreto, o número desejado de ações a serem geradas (`n_shares`), o número mínimo de ações necessárias para reconstruir o segredo (`min_shares`), e um número primo (`prime`). Inicialmente, são gerados coeficientes aleatórios para formar um polinómio que simboliza a informação não revelada. O coeficiente principal engloba o próprio segredo. A função subsequentemente calcula a expressão polinomial para valores que variam de 1 a 'n_ações e gera uma lista de pares. Cada par da lista inclui um índice que representa o número da ação e o valor associado da expressão polinomial, que representa a ação. A função polinomial calcula a saída numérica de uma expressão polinomial, dado o valor de entrada 'x', os coeficientes do polinómio e um número inteiro primo. O programa utiliza para calcular a expressão polinomial e, em seguida, apresenta o resultado calculado.

O objetivo deste código é ilustrar o conceito subjacente ao sistema de partilha de segredos de Shamir. O processo de separar uma informação confidencial em várias partilhas e depois distribuí-las por diferentes entidades garante que o segredo original só pode ser reconstruído quando um determinado limiar, definido como o número mínimo de partilhas, é atingido e as partilhas são combinadas coletivamente pelas partes envolvidas. Este exemplo particular exemplifica um método utilizado na Computação Segura Multipartidária (SMPC), em que muitas partes se envolvem na computação colaborativa de funções utilizando os seus respectivos dados privados, assegurando simultaneamente a não divulgação desses dados umas às outras.

6.4 Privacidade diferencial

O princípio fundamental da privacidade diferencial é uma metodologia especificamente desenvolvida para proteger a privacidade dos dados individuais, permitindo simultaneamente a derivação de conhecimentos estatísticos agregados.

Numa primeira observação, o código da imagem A.4 parece seguir a seguinte sequência de acções:

O código começa por importar as bibliotecas necessárias. A biblioteca "pandas"é carregada e é-lhe atribuído um pseudónimo "pd"para permitir a manipulação dos dados. Além disso, o algoritmo "BoundedMean"é importado do módulo "pydp.algorithms.laplacian". O procedimento é de extrema importância para o cálculo de valores médios privados diferenciados.

O código procura obter o conjunto de dados "Adultos"a partir de um URL especificado, iniciando assim o procedimento de carregamento de dados.

Os parâmetros de privacidade diferencial incluem a inicialização de um valor 'epsilon', que é fixado em 1,0. O parâmetro 'epsilon' desempenha um papel crucial na determinação do equilíbrio entre a preservação da privacidade e a precisão dos resultados calculados. Valores reduzidos do parâmetro "epsilon"dão garantias mais rigorosas de privacidade, mas podem resultar em conclusões menos exatas.

A inicialização de um objeto BoundedMean envolve a instanciação de uma instância do algoritmo 'BoundedMean' com o valor previamente especificado de 'epsilon'. A técnica conhecida como 'BoundedMean' desempenha um papel crucial na produção de estimativas diferentemente privadas de valores médios derivados de um determinado conjunto de dados.

Para incluir novos dados e calcular a média, as idades das pessoas são retiradas da coluna "idade"do conjunto de dados e depois transformadas em valores inteiros através da utilização de uma compreensão de lista. As idades que foram transformadas são depois integradas no objeto `b.mean` utilizando a função `add_entries`.

O cálculo da idade média privada diferencial é efectuado invocando o método `result` no objeto `b.mean`.

O principal objetivo deste código é fornecer uma ilustração tangível da implementação da privacidade diferencial. Esta secção centra-se na utilização da biblioteca `python-dp` para calcular a idade média das pessoas num conjunto de dados, assegurando simultaneamente a proteção da privacidade individual. O `python-dp` facilita a criação de algoritmos diferencialmente privados para efetuar análises estatísticas em material sensível, preservando simultaneamente a privacidade individual.

6.5 Qual é o mecanismo subjacente ao seu funcionamento?

O Advanced Encryption Standard (AES) pode ser visto como semelhante a um sistema de chave e fechadura, concebido para proteger as comunicações contra o acesso não autorizado. Nas situações em que o emissor e o recetor de uma comunicação pretendem manter a sua confidencialidade, acordam em utilizar uma chave secreta, que pode ser comparada a um código exclusivo. Utilizando esta chave criptográfica, a mensagem que transmite é submetida a um processo sistemático de divisão e mistura precisa, por vezes designado por scrambling. O algoritmo AES repete este processo inúmeras vezes, estabelecendo assim vários níveis de segurança semelhantes à resolução de uma sequência de enigmas complexos.

A comunicação encriptada é então enviada ao destinatário pretendido, que utiliza uma chave de encriptação idêntica para inverter o processo de encriptação, revelando assim a informação original. Decifrar a mensagem sem a chave exata é uma tarefa muito difícil. A norma de encriptação avançada (AES) é amplamente utilizada em muitos contextos, como canais de comunicação seguros, transações em linha e proteção de dados sensíveis. O seu principal objetivo é assegurar a confidencialidade e a integridade da informação, impedindo assim o acesso ilegal e garantindo que esta permanece indisponível para partes não autorizadas.

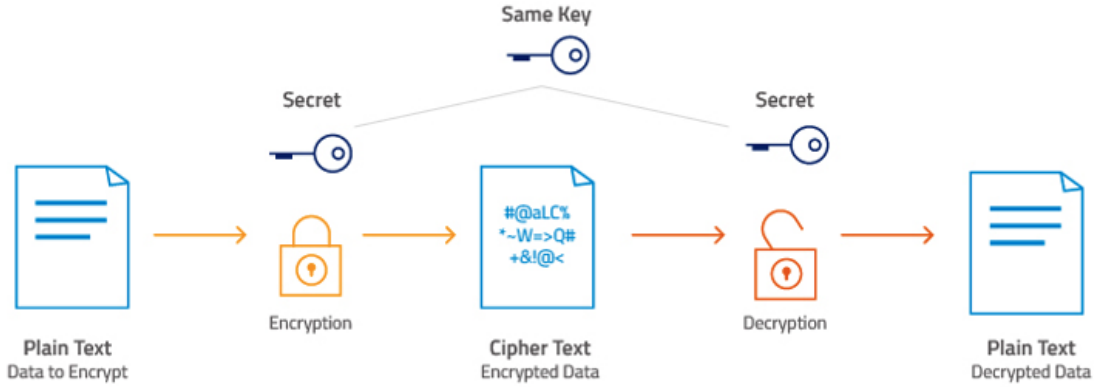

Figura 6.1: Advanced Encryption Standard [23]

O método também pode ser compreendido em seis fases simples.

1. O estabelecimento de um código secreto envolve um acordo mútuo entre os comunicadores, semelhante à utilização de uma chave especializada. A referida chave serve o objetivo de garantir e anular a segurança da comunicação.
2. O processo de codificação da mensagem envolve o rearranjo deliberado das suas partes constituintes, utilizando um sistema de codificação confidencial. O processo pode ser comparado à reorganização das palavras e letras de uma mensagem de forma deliberada.
3. O algoritmo AES emprega muitas iterações do processo de *jumbling* para melhorar as suas medidas de segurança. Em cada ronda sucessiva, a mensagem fica ainda mais distorcida.
4. Reforço das medidas de segurança: Para além do seu processo de encriptação, o AES incorpora mais dados confidenciais na mensagem encriptada, aumentando assim a sua resistência à desencriptação.
5. Transmissão da comunicação encriptada: A mensagem codificada pode agora ser enviada para a pessoa com quem está a comunicar. Em caso de interceção da comunicação, a sua compreensão torna-se impossível devido à natureza emaranhada do seu conteúdo.
6. Decifrar a comunicação: Após a receção da comunicação emaranhada, o destinatário utiliza um sistema de codificação oculto idêntico para a decifrar. O procedimento acima referido é executado de forma inversa, invertendo assim o efeito da confusão e revelando, em última análise, a mensagem original.

6.6 Base de dados

A utilização da encriptação de dados é crucial para garantir a proteção de informações sensíveis contra o acesso não autorizado. A encriptação é um mecanismo crucial que salvaguarda a informação sensível, incluindo dados pessoais, documentação financeira e dados proprietários, garantindo a sua confidencialidade e segurança. Um conjunto de dados exemplar que se adequa bem à encriptação dentro de um pipeline de dados é o *Adult Census Income Dataset*. As informações fornecidas foram cuidadosamente selecionadas e compiladas a partir do U.S. Census Bureau. Inclui uma vasta gama de características demográficas, tais como idade, educação, profissão e rendimento. O rendimento, no contexto em discussão, serve como um excelente exemplo de dados que necessitam de encriptação. O ato de ocultar este ponto de dados crucial do acesso público não só é uma abordagem sensata e cautelosa, como também serve de medida de proteção contra quaisquer violações de segurança. Ao utilizar métodos de encriptação especificamente concebidos para dados numéricos e categóricos, este conjunto de dados serve como um excelente exemplo que realça a necessidade de salvaguardar os dados. No contexto de uma sociedade orientada para os dados, a proteção de conjuntos de dados sensíveis, como o Conjunto de Dados sobre o Rendimento dos Adultos do Censo, é crucial para manter a privacidade e garantir a fiabilidade das condutas de dados.

Cada caraterística denota factos distintos relativos a uma pessoa. Segue-se uma análise exaustiva das características incluídas no conjunto de dados:

1. Idade: A variável idade é uma medida contínua que representa a idade cronológica da pessoa.
2. Classe de trabalho: É uma variável categórica que representa o tipo de classe de trabalho a que uma pessoa pertence.
3. Peso: Uma medida contínua que representa o peso final atribuído a cada participante no âmbito do inquérito.
4. Educação: É uma variável categórica que indica o grau mais elevado de habilitações literárias atingido por uma pessoa.
5. EducaçãuNum: É uma variável contínua que representa a codificação numérica das habilitações literárias.
6. Marital-status: Variável categórica que representa o estado civil de uma pessoa.
7. Ocupação: É uma variável categórica que indica a profissão específica de uma pessoa.
8. Relação: Uma variável categórica que representa o estado civil de uma pessoa.
9. Raça : É uma variável categórica que indica a origem racial de uma pessoa.
10. Sexo: Uma variável categórica que indica o género da pessoa.

Diversos elementos do conjunto de dados exibem atributos que os tornam apropriados para serem criptografados.

6.7 Resumo

Este capítulo apresenta uma análise aprofundada de um conjunto de técnicas de preservação da privacidade que podem ser utilizadas de forma eficiente através da utilização da linguagem de programação Python. A presente investigação centra-se em três técnicas

principais que têm uma importância considerável no domínio da segurança e privacidade dos dados: encriptação homomórfica, computação multipartidária segura (SMPC) e privacidade diferencial.

O principal objetivo desta investigação é fornecer uma compreensão completa destas abordagens, permitindo assim tomar decisões bem informadas em matéria de privacidade e segurança dos dados.

A secção 6.1 do documento aborda o conceito de encriptação homomórfica, com especial destaque para a biblioteca Paillier. A secção oferece exemplos práticos de implementação da encriptação homomórfica utilizando a linguagem de programação Python.

Nesta secção, é apresentado a biblioteca SEAL, versão 6.2, frequentemente conhecida como PySEAL. A biblioteca SEAL, criada pela Microsoft, oferece uma solução viável para a implementação da encriptação homomórfica.

No âmbito da secção 6.3, é examinada a noção de computação multipartidária segura (SMPC). A SMPC permite que muitas entidades participem em cálculos colaborativos, salvaguardando a confidencialidade dos respetivos dados de entrada. São realçadas as vantagens desta metodologia e as suas implementações pragmáticas.

Na secção 6.4 - Privacidade Diferencial centra-se no estudo da privacidade diferencial, uma abordagem que visa salvaguardar a privacidade dos dados individuais, permitindo simultaneamente a extração de conhecimentos estatísticos coletivos. Neste discurso, examinamos os conceitos subjacentes ao tema e exploramos as suas aplicações práticas.

Por fim na secção final, a base de dados utilizada nesta investigação é descrita em pormenor. Esta base de dados foi utilizada como ambiente para a apresentação de abordagens que garantem a preservação da privacidade. O presente discurso tem como objetivo elucidar as características inerentes e o significado do tema em questão, no que diz respeito à avaliação e implementação de metodologias de segurança e privacidade.

Este capítulo apresenta uma análise aprofundada dos procedimentos de preservação da privacidade discutidos nesta investigação, permitindo adquirir uma compreensão sólida destas abordagens e da sua implementação eficiente através da utilização de Python.

Neste capítulo, irá ser será focado à apresentação detalhada da solução proposta neste trabalho. Este é um passo crucial na compreensão do funcionamento e do valor do projeto. É abordado diversas dimensões essenciais, começando por uma análise aprofundada dos requisitos funcionais que são cruciais para o sucesso.

Para começar será feito um levantamento dos requisitos funcionais específicos para a solução. Cada um deles será minuciosamente descrito e contextualizado, com ênfase na maneira como eles se alinham com os objetivos globais do projeto. Esta análise fornecerá uma visão detalhada das capacidades e funcionalidades que a solução oferecerá aos utilizador final, estabelecendo as bases para uma compreensão sólida dos benefícios que trará.

6.8 Análise de Requisitos

Os requisitos funcionais desempenham um papel essencial na especificação de um sistema de software ou projeto. Os requisitos representam as funcionalidades or comportamentos específicos que um sistema deve oferecer, a fim de atender aos objetivos e necessidades dos utilizadores e das partes interessadas.

A relevância dos requisitos funcionais reside na medida em que eles estabelecem as especificações do que o sistema deve realizar.

A alinhamento de expectativas é um processo crucial do desenvolvimento de sistemas, os requisitos funcionais desempenham um papel fundamental nesse sentido. Eles permitem estabelecer uma compreensão compartilhada entre a equipa de desenvolvimento e os *stakeholders* sobre as funcionalidades e características que o sistema irá entregar.

A avaliação do sucesso de um sistema é facilitada pelos requisitos funcionais, que estabelecem as funcionalidades que o system deve possuir. Esses requisitos fornecem critérios objetivos para determinar se o sistema atende às expectativas e requisitos estabelecidos.

Em resumo, os requisitos funcionais são uma componente integral do processo de desenvolvimento de software, uma vez que desempenham um papel fundamental na definição do que um sistema deve realizar e na garantia de que satisfaz as necessidades e expectativas dos utilizadores finais.

6.8.1 Requisitos funcionais

Identificador	[RF001] Envio de Dados para o Kafka
Prioridade	Essencial
Descrição	O sistema deve ser capaz de enviar os dados do arquivo (linhas) para um tópico específico no Kafka.
Motivação	O sistema deverá proporcionar o envio de dados encriptados para um tópico específico.

Identificador	[RF002] Encriptação dos dados
Prioridade	Essencial
Descrição	Os dados devem ser encriptados antes de serem enviados para o Kafka.
Motivação	A criação deste requisito é de extrema importância, faz com que os nossos dados sensiveis consigam ser encriptados.

Identificador	[RF003] Desencriptar os dados
Prioridade	Essencial
Descrição	O sistema deve ser capaz de desencriptar os dados após serem recebidos no Kafka.
Motivação	O sistema deverá ser capaz de desencriptar os dados de forma a verificar o resultado do que tem o tópico de kafka.

Identificador	[RF004] Chave de encriptação
Prioridade	Essencial
Descrição	Uma chave de criptografia aleatória deve ser gerada para cada conjunto de dados a ser criptografado.
Motivação	Uma chave deverá ser gerada de forma a podermos encriptar e desencriptar os dados

Identificador	[RF005] Base de dados
Prioridade	Essencial
Descrição	O sistema deverá ser capaz de obter informação sensivel de forma a testar o funcionamento de criptografia.
Motivação	De forma a fazer o teste de criptografia é necessario obter essa informação de uma base de dados pública

6.8.2 Requisitos não funcionais

Os requisitos não funcionais incluem os critérios que definem as características ou atributos gerais que um sistema de software deve possuir, em conjunto com as funções específicas que se espera que forneça. Os pontos focais incluem várias dimensões, nomeadamente o desempenho, a segurança, a facilidade de utilização, a fiabilidade e a eficiência. Os requisitos funcionais incluem as especificações que definem o comportamento pretendido do sistema, enquanto os requisitos não funcionais dizem respeito ao modo e ao grau em que o sistema realiza a funcionalidade pretendida.

Identificador	[RNF001] Segurança
Prioridade	Essencial
Descrição	A principal preocupação deste projeto é a segurança dos dados. Os requisitos não funcionais devem garantir que os dados sejam criptografados de forma segura e que a chave de criptografia seja mantida em sigilo
Motivação	O sistema deverá proporcionar o envio de dados encriptados para um tópico específico.

Identificador	[RNF002] Disponibilidade
Prioridade	Essencial
Descrição	O sistema deve estar disponível para enviar dados para o Kafka durante o tempo especificado no código, mesmo em situações de carga.
Motivação	O sistema deverá proporcionar o envio de dados encriptados para um tópico específico.

6.8.3 Atores

Este subcapítulo elucidará as principais partes interessadas envolvidas no sistema proposto. Serão identificadas as partes interessadas importantes envolvidas, desenvolvimento e utilização deste projeto. Cada interveniente desempenha um papel único e contribui para a eficácia global da solução, desde os utilizadores finais que interagem diretamente com o sistema até aos programadores que são responsáveis pela sua execução. Para compreender plenamente a dinâmica do ecossistema de segurança dos dados, é importante aprofundar os deveres e as interações de cada participante.

- Utilizador final referem-se às pessoas ou organizações que se envolvem diretamente com o sistema e utilizam as suas funções. Esta categoria engloba aqueles que têm acesso ou utilizam dados ou informações encriptadas, incluindo, mas não se limitando a analistas de dados, investigadores e administradores.

- Os programadores referem-se ao grupo de indivíduos que são responsáveis pela criação e manutenção contínua do sistema. Os indivíduos são responsáveis pela conceção, implementação e manutenção das funcionalidades do sistema, que incluem a encriptação e a comunicação com o Kafka.

6.8.4 Casos de uso

O diagrama de casos de uso apresentado fornece uma representação visual das interações entre os atores principais e o sistema, bem como as capacidades específicas que cada ator é capaz de executar. Os casos de utilização fornecem uma descrição exaustiva das atividades exatas que cada interveniente é capaz de executar no sistema. Isto ajuda a compreender o envolvimento das várias partes e a forma como o sistema satisfaz os seus requisitos.

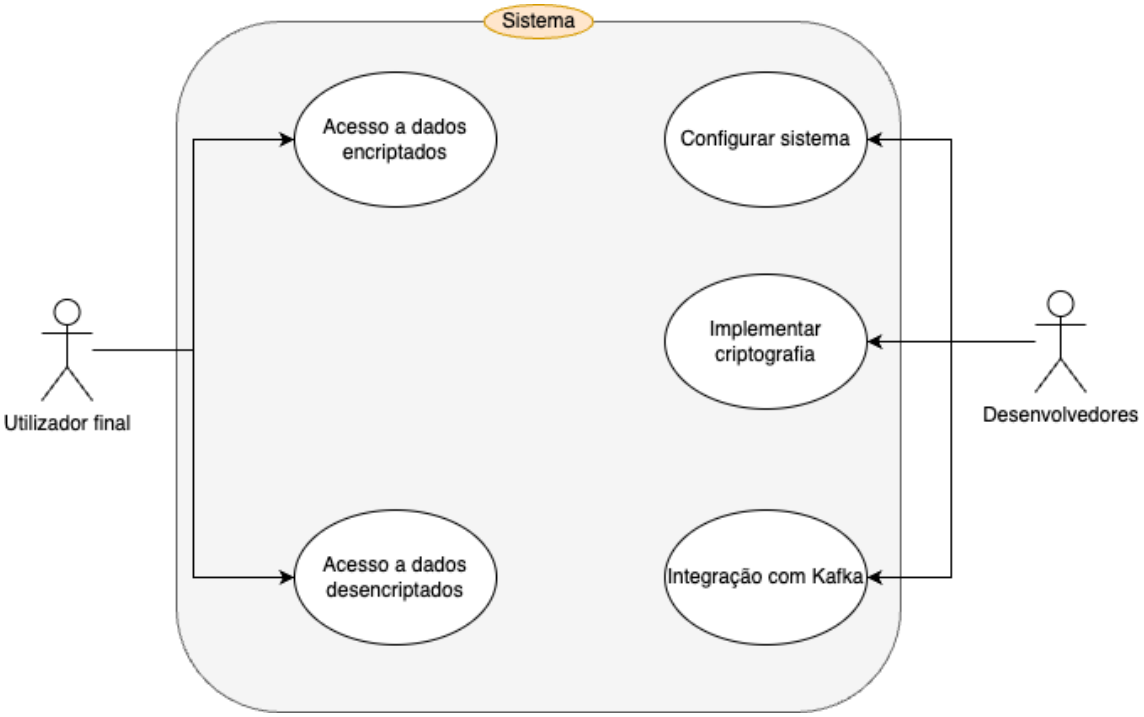

Figura 6.2: Diagrama de Caso de uso

6.9 Conjunto de tecnologias utilizadas

As tecnologias utilizadas neste projeto giram em torno da utilização de Python e Kafka. Python, uma linguagem de programação conhecida pela sua versatilidade e utilização generalizada, constitui a base da nossa solução. A presença de um conjunto diversificado de bibliotecas e frameworks no seu ecossistema permite a criação de código forte e eficiente.

Para além de utilizarmos a linguagem de programação Python, utilizamos as capacidades do Kafka, uma plataforma de streaming distribuída e bem reconhecida. A força do Kafka reside na sua capacidade de gerir eficazmente fluxos de dados em tempo real, tornando-o uma escolha altamente adequada para coordenar o fluxo de dados e as interações dentro do nosso projeto. Ao utilizar as capacidades do Python e do Kafka, desenvolvemos uma estrutura versátil e eficiente para o projeto.

6.10 Implementação do projeto e fluxo de trabalho

O principal objetivo do projeto é construir um fluxo de dados que seja seguro e eficiente, utilizando as tecnologias Python e Kafka. Ao aderir a uma metodologia metódica, garantimos a preservação da integridade dos dados, a privacidade e a capacidade de analisar informações em tempo real.

O processo de recolha de dados é da maior importância, uma vez que envolve a identificação e a extração de fontes de dados. O estudo inicia-se com o apuramento da fonte de dados, nomeadamente a Base de Dados de Rendimentos dos Censos de Adultos. O conjunto de dados utilizado neste estudo serve de base fundamental para a nossa demonstração. Uma componente essencial do processo é o desenvolvimento de um módulo Python que tem a capacidade de extrair eficazmente dados de diferentes fontes. A principal função deste módulo é recuperar dados de fontes externas e depois integrá-los no pipeline de dados para facilitar o processamento posterior.

A transformação de dados é uma etapa crucial no pré-processamento de dados, que envolve a conversão de dados brutos num formato adequado para análise e modelação. Este processo inclui várias técnicas, como a limpeza de dados, a normalização e a encriptação. A encriptação, em particular, é utilizada para proteger informações sensíveis. Esta fase engloba as atividades de limpeza, pré-processamento e normalização dos dados, que são realizadas para melhorar a qualidade e a fiabilidade dos dados. O tratamento de valores em falta, a remoção de duplicados e a conversão de tipos de dados são efetuados conforme necessário. O ponto fulcral desta fase é a implementação de técnicas de encriptação. O procedimento acima mencionado garante a salvaguarda dos dados sensíveis incluídos no conjunto de dados, impedindo assim que indivíduos não autorizados tenham acesso aos mesmos. São examinadas diferentes técnicas de encriptação para obter diferentes graus de segurança, de acordo com as necessidades específicas de um projeto.

Armazenamento de dados

A seleção de uma solução de armazenamento adequada é crucial para garantir a preservação da integridade dos dados que foram convertidos e encriptados. Várias opções, incluindo bases de dados relacionais, bases de dados NoSQL, sistemas de ficheiros distribuídos e lagos de dados, são avaliadas de acordo com os requisitos específicos do projeto. Um esquema ou estrutura de dados meticulosamente especificado é estrategicamente criado e executado para garantir a compatibilidade com o sistema de armazenamento escolhido.

O processamento de dados é um aspeto crítico dos sistemas informáticos modernos. Uma abordagem popular para a análise de dados em tempo real é a utilização do Apache Kafka. O Kafka é uma plataforma de fluxo distribuído que permite o processamento de grandes volumes de dados em tempo real. Ao utilizar as capacidades do Kafka, as organizações podem analisar e processar dados de forma eficiente. O núcleo do projeto está centrado na utilização do Apache Kafka para o processamento de dados. O Kafka funciona como uma plataforma de fluxo distribuído que facilita o processamento em tempo real de fluxos de dados. Os produtores Kafka são concebidos para transmitir dados encriptados da camada de armazenamento para tópicos Kafka. Em seguida, os consumidores Kafka subscrevem os tópicos, desencriptam os dados recebidos e efetuam o processamento posterior. Esta etapa engloba vários processos, como o enriquecimento de dados, a agregação e a utilização de modelos de aprendizagem automática, o que realça a multifuncionalidade do Kafka no tratamento de fluxos de dados em tempo real.

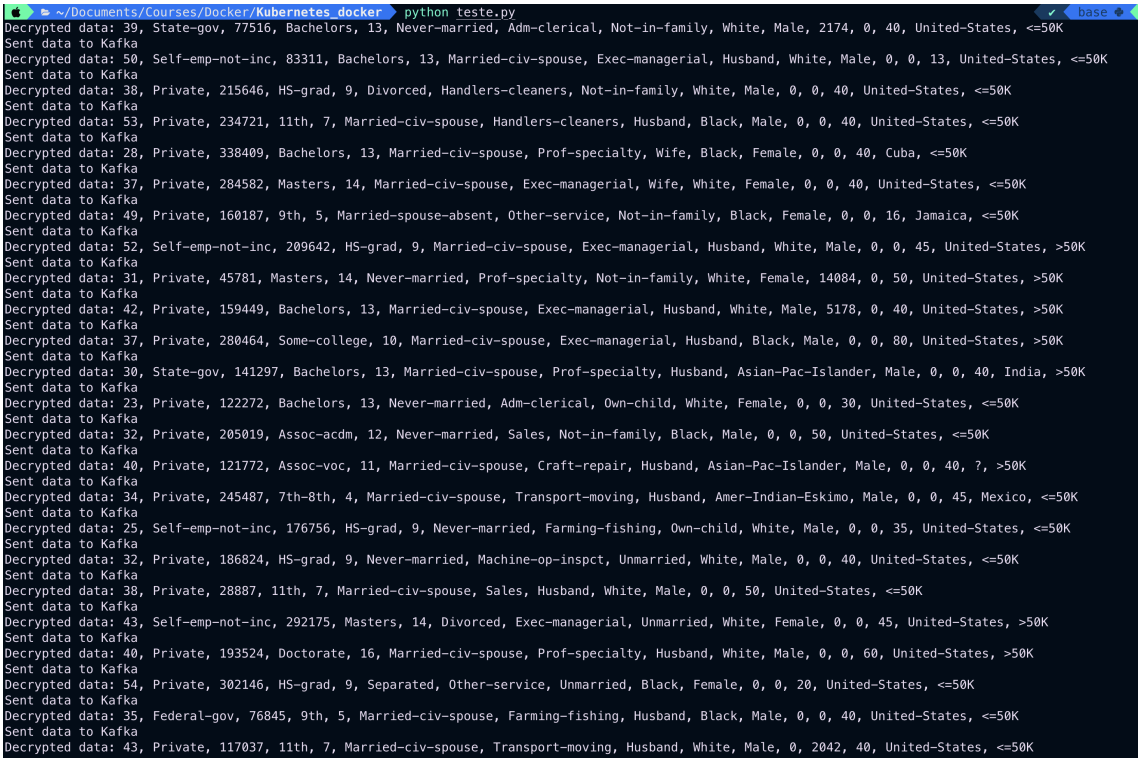

Figura 6.3: Pipeline segura com encriptação AES e integração Kafka

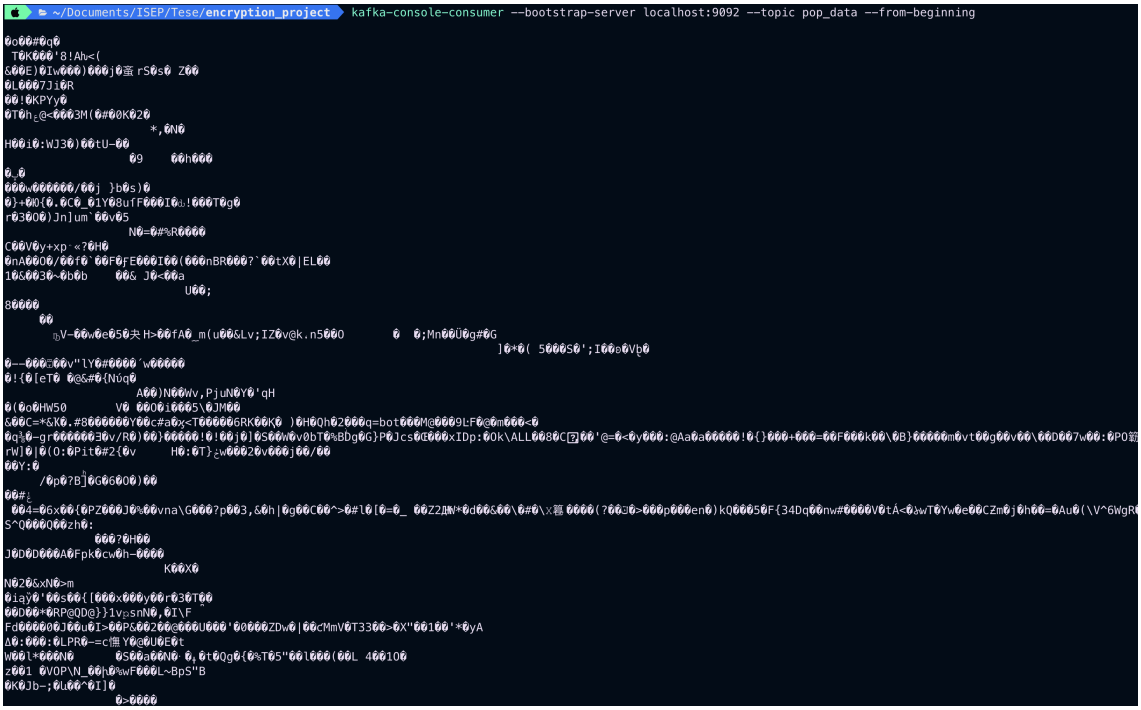

Figura 6.4: Encriptação AES e integração Kafka

A análise de dados é um campo que envolve o exame, a interpretação e a extração de conhecimentos e padrões significativos de grandes conjuntos de dados. As ferramentas e estruturas de análise, como o Apache Spark ou o Apache Flink, são incluídas no pipeline de processamento de dados para extrair informações significativas dos dados processados. Estas tecnologias facilitam a execução do processamento em lote ou do processamento em fluxo em tempo real, permitindo a adaptação às necessidades específicas do projeto. Através da utilização destas tecnologias, as capacidades inerentes permitem a identificação e revelação de padrões, tendências e correlações presentes na informação.

A visualização de dados é um aspeto crucial para apresentar eficazmente os conhecimentos derivados da análise de dados. A visualização é um componente crucial da análise de dados. A comunicação bem sucedida dos conhecimentos sobre os dados é facilitada pela utilização de bibliotecas ou plataformas de visualização de dados adequadas. A utilização de técnicas de visualização facilita a apresentação eficaz de informações complexas de uma forma facilmente compreensível, permitindo assim que os intervenientes compreendam prontamente e extraiam conhecimentos essenciais.

A fase final do procedimento é a transmissão dos dados processados aos destinatários ou sistemas designados. Isto pode incluir a geração de relatórios, a iniciação de alertas ou a integração com outros sistemas para permitir outras ações baseadas nos conhecimentos gerados a partir dos dados.

6.11 Problemas enfrentados neste projeto

Esta parte tem uma importância significativa, uma vez que o processo de construção de uma explicação concisa desta noção se revelou algo complexo devido a vários fatores.

Um dos principais desafios enfrentados ao longo do projeto foi a implementação da encriptação homomórfica. A encriptação homomórfica é uma metodologia muito potente que permite a execução de cálculos em dados encriptados sem a necessidade de desencriptação. Essa capacidade específica foi procurada para aumentar os níveis de segurança e privacidade no projeto. No entanto, foram encontradas algumas dificuldades na utilização dos módulos Python existentes para a encriptação homomórfica. Foi feito um esforço para instalar bibliotecas como a Tenseal e a PySEAL, que fornecem capacidades para a encriptação homomórfica. Infelizmente, essas bibliotecas não ofereceram um suporte adequado e apresentaram obstáculos durante o processo de instalação. Ao longo do procedimento de instalação, muitas complicações surgiram, incluindo problemas de compatibilidade e dependências, mesmo ao tentar instalá-las usando o pip. A existência desse fator impediu significativamente a implementação eficaz da encriptação homomórfica no projeto, como originalmente planeado. O processo de instalação apresentou desafios significativos e, mesmo após a conclusão, apenas o código de exemplo fornecido se mostrou funcional.

Outro desafio que foi encontrado estava relacionado com a própria biblioteca Kafka-Python, que foi utilizado para interagir com o Apache Kafka. Houve alguns problemas com a biblioteca, como comportamentos inesperados e erros. Para ultrapassar estes problemas, foi criado um novo ambiente virtual Python e reinstalar todas as bibliotecas necessárias. Esta abordagem ajudou a resolver os problemas e garantiu o bom funcionamento da integração do Kafka no meu projeto. Apesar destes desafios, foram ultrapassadas as dificuldades explorando soluções alternativas e otimizando a implementação. Embora não tenha sido possível incorporar a encriptação homomórfica devido às limitações das bibliotecas disponíveis. Estas dificuldades levaram a importância da pesquisa exaustiva de bibliotecas, das técnicas de

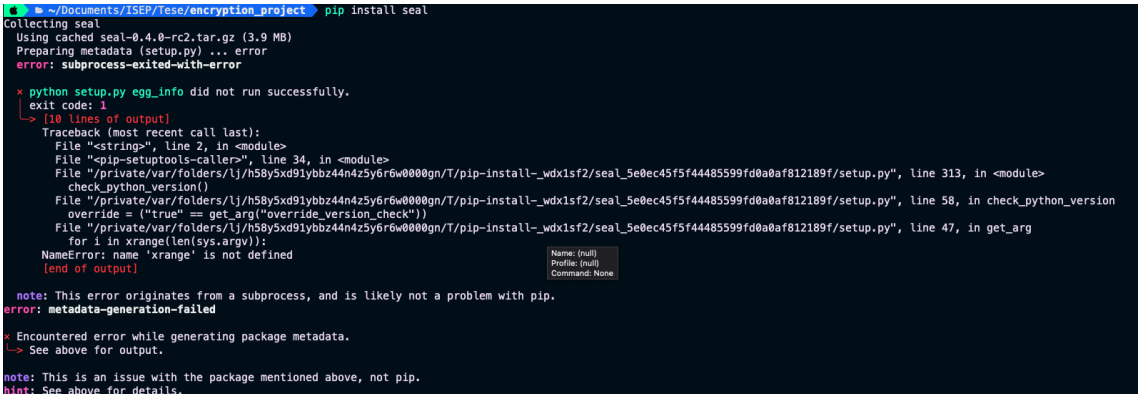

Figura 6.5: Instalação biblioteca

resolução de problemas e da necessidade de adaptação a obstáculos imprevistos no processo de desenvolvimento. Como alternativa viável, decidi utilizar a encriptação AES da biblioteca Cryptography Python, que provou ser uma solução de encriptação fiável e eficaz para proteger dados sensíveis no meu projeto.

6.12 Resumo

Em resumo, o estudo conduziu a uma investigação exaustiva das técnicas de encriptação associadas às condutas de dados encriptados. A aplicação eficaz do algoritmo de encriptação AES resultou no estabelecimento de um quadro seguro para a transmissão e o armazenamento de informações sensíveis, reforçando assim o seu carácter secreto e a sua integridade. Apesar do meu objetivo inicial de utilizar métodos de encriptação homomórfica para melhorar a qualidade dos cálculos que preservam a privacidade, deparei-me com várias limitações nas bibliotecas atualmente disponíveis. Esta experiência serviu-me de catalisador para desenvolver um pensamento inovador e adotar novas abordagens, o que acabou por resultar numa maior viabilidade do projeto.

Durante esta experiência, adquiri conhecimentos significativos sobre as características complexas das técnicas de encriptação. A expedição realçou a necessidade de efetuar uma pesquisa exaustiva sobre bibliotecas acessíveis, de aperfeiçoar as capacidades de resolução de problemas e de ser adaptável em resposta a problemas imprevistos ao longo do processo de produção. Os desafios enfrentados, que vão desde as restrições das bibliotecas de encriptação homomórfica até às complexidades da interface Kafka-Python, sublinham a necessidade de realizar uma avaliação exaustiva e um processo de seleção da tecnologia.

Esta dissertação funciona efetivamente como uma investigação demonstrativa; no entanto, é crucial reconhecer que, para uma solução mais completa e escalável, é altamente aconselhável utilizar recursos como o MongoDB Atlas e o Confluent Cloud. O MongoDB Atlas posiciona-se como uma opção muito vantajosa, fornecendo um serviço de base de dados gerido que é melhorado com mecanismos de segurança integrados. O Confluent Cloud oferece uma solução Kafka gerida abrangente que garante um fluxo de dados fiável. Ao utilizar estas tecnologias, o alcance do projeto pode ser alargado para tratar eficientemente grandes quantidades de dados, garantir um tempo de funcionamento consistente e estabelecer facilmente uma interface com outros sistemas atuais.

Em conclusão, este projeto proporcionou uma oportunidade valiosa para explorar os conceitos fundamentais da encriptação, as complexidades das condutas de dados e os obstáculos significativos envolvidos na integração de diversas tecnologias. A experiência acima mencionada funcionou como uma oportunidade valiosa e impactante, destacando a importância significativa de uma seleção cuidadosa da tecnologia, a fim de otimizar a criação de condutas de processamento de dados eficazes e seguras.

Capítulo 7

Conclusão

Em resumo, este trabalho de investigação levou a cabo uma análise abrangente das metodologias de encriptação e anonimização no que diz respeito às estruturas de Big Data. Esta investigação realizou uma avaliação exaustiva de vários algoritmos de encriptação, incluindo técnicas simétricas e assimétricas, a fim de elucidar os seus atributos, benefícios e limitações distintos. Através da análise dos tempos de encriptação e da implementação de investigação empírica, foi adquirido um conhecimento significativo sobre as ramificações práticas de cada metodologia.

Além disso, a utilização empírica de determinados algoritmos de encriptação e anonimização num conjunto substancial de dados, incluindo informações confidenciais, demonstrou a relevância prática destas metodologias em cenários do mundo real. A avaliação da segurança dos dados, da preservação da privacidade e da usabilidade dos dados demonstrou a eficácia das metodologias selecionadas para garantir a proteção da informação, mantendo a sua praticabilidade.

Os resultados obtidos com este estudo proporcionam avanços substanciais no domínio da segurança e da privacidade dos dados no contexto das plataformas de grandes volumes de dados. A avaliação exaustiva, as avaliações empíricas e a aplicação pragmática fizeram avançar conjuntamente a compreensão da encriptação e da anonimização, facilitando assim a tomada de decisões bem informadas sobre métodos de proteção de dados.

7.1 Síntese

Este tema de tese teve como objetivo investigar as várias estratégias de encriptação e anonimização utilizadas em sistemas de Big Data. Os principais objetivos do projeto foram o exame, a avaliação e a utilização prática de técnicas de encriptação para melhorar a segurança, a confidencialidade e a facilidade de utilização dos dados. A metodologia de investigação utilizou uma abordagem metódica, incluindo análise teórica, investigações empíricas e aplicação prática. A expedição acima mencionada forneceu revelações significativas relativas ao complexo domínio da segurança de dados, elucidando os méritos e as desvantagens de várias metodologias e apresentando sugestões para investigações e implementações futuras.

7.2 Objetivos Realizados

O objetivo deste secção é estabelecer uma ligação entre os objectivos delineados no Capítulo 1 e os resultados alcançados ao longo desta dissertação de mestrado. Conforme delineado

na introdução, os principais objectivos foram:

1. Comparar e avaliar várias técnicas de encriptação e anonimização utilizadas em sistemas de *Big Data*, incluindo algoritmos de encriptação simétrica e assimétrica: Neste contexto, foram exploradas diversas técnicas de encriptação e anonimização, incluindo encriptação homomórfica, privacidade diferencial e outras. As avaliações detalhadas dessas técnicas foram realizadas, destacando as diferenças entre algoritmos simétricos e assimétricos.
2. Realizar um estudo empírico dos algoritmos de encriptação mais populares, avaliando as suas características, benefícios e desvantagens: Foi realizado um estudo empírico abrangente dos algoritmos de encriptação mais populares, analisando suas características, vantagens e desvantagens.
3. Identificar os benefícios e desvantagens de cada técnica de encriptação e anonimização em termos de segurança dos dados, proteção da privacidade e usabilidade dos dados: Foi identificado cuidadosamente os prós e contras de cada técnica de encriptação e anonimização em relação à segurança dos dados, proteção da privacidade e usabilidade. Isso envolveu uma análise detalhada das implicações de cada técnica.
4. Desenvolver um cenário prático que envolva um elevado conjunto de dados contendo informações sensíveis, onde serão implementadas as técnicas de encriptação e anonimização selecionadas: Foi criado um cenário prático que envolveu um grande volume de dados contendo informações sensíveis. Nesse cenário, foi implementado as técnicas de encriptação e anonimização selecionadas para avaliar sua eficácia na prática.
5. Propor conclusões e recomendações com base nos resultados obtidos, incluindo possíveis melhorias nas técnicas de encriptação e anonimização de *Big Data*, e áreas de investigação futura: Com base nos resultados obtidos, foi proposto conclusões abrangentes e recomendações. Além disso, destacado em áreas de pesquisa futura e possíveis melhorias nas técnicas de encriptação e anonimização de *Big Data*.

7.3 Resultados Alcançados

Os resultados do projeto abrangem:

Esta dissertação tem como objetivo fornecer uma análise abrangente dos sistemas de encriptação e anonimização, elucidando os seus méritos e desvantagens individuais. Este estudo fornece provas empíricas sobre o desempenho dos algoritmos de encriptação, permitindo que os indivíduos façam julgamentos informados ao selecionar uma estratégia. Este documento pretende explorar a implementação prática de métodos de encriptação e anonimização, destacando a sua relevância e eficácia em contextos do mundo real. Avaliação da segurança dos dados, preservação da privacidade e usabilidade em relação às metodologias utilizadas.

7.4 Áreas potenciais para investigação futura

Dado o potencial do projeto para expandir as fronteiras da segurança e privacidade dos dados, é imperativo ir mais longe em vários domínios que necessitam de mais investigação.

Este estudo pretende investigar métodos de encriptação híbridos que incluam as vantagens dos sistemas de encriptação simétricos e assimétricos. Este estudo pretende examinar novos algoritmos de encriptação especificamente concebidos para sistemas de Big Data, com foco

na avaliação do seu desempenho e características de segurança. O cenário prático é alargado para incluir casos de utilização mais complexos e uma gama mais vasta de conjuntos de dados, oferecendo assim uma avaliação mais abrangente da eficácia da abordagem. A vigilância e o ajustamento contínuos das técnicas de encriptação e anonimização em resposta ao panorama em constante mudança das preocupações com a segurança dos dados.

Bibliografia

[1] Isitor Emmanuel e Dr Clare Stanier. «Defining Big Data». Em: *Review of Scientific Instruments* 72.12 (dez. de 2016), pp. 4477–4479. url: https://dl.acm.org/doi/pdf/10.1145/3010089.3010090.

[2] Rashbir Singh Avi Bhardwaj Vikas Deep. «A Technique for Big Data Testing considering 3V's». Em: *Review of Scientific Instruments* 69.3 (mar. de 2018), pp. 222–225. url: https://ieeexplore.ieee.org/document/8752996.

[3] Nils Gruschka et al. «Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR». Em: *Materials Today: Proceedings* 46.1 (mar. de 2019), pp. 5027–5033. url: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8622621.

[4] Avi Bhardwaj, Vikas Deep e Rashbir Singh. «A Technique for Big Data Testing considering 3V's». Em: *Review of Scientific Instruments* 69.3 (mar. de 2018), pp. 222–225. url: https://ieeexplore.ieee.org/document/8752996.

[5] Avi Bhardwaj, Vikas Deep e Rashbir Singh. «A Technique for Big Data Testing considering 3V's». Em: *Review of Scientific Instruments* 69.3 (mar. de 2018), pp. 222–225. url: https://ieeexplore.ieee.org/document/8752996.

[6] Deepika Sharma Priya Matta Minit Arora. «A comparative survey on data encryption Techniques: Big data perspective». Em: *Materials Today: Proceedings* 46.1 (mar. de 2021), pp. 11035–11039. url: https://doi.org/10.1016/j.matpr.2021.02.153.

[7] Abdul Majeed e Sungchang Lee. «Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey». Em: *Review of Scientific Instruments* 4.3 (mar. de 2021), pp. 4236–4243. url: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9298747.

[8] Nils Gruschka et al. «Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR». Em: *Materials Today: Proceedings* 46.1 (mar. de 2019), pp. 5027–5033. url: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8622621.

[9] K S Ananda Kumar et al. «BIG DATA CHARACTERISTICS, CLASSIFICATION AND CHALLENGES-A REVIEW». Em: *Review of Scientific Instruments* 69.3 (mar. de 2021), pp. 4236–4243. url: https://www.turcomat.org/index.php/turkbilmat/article/download/8316/6490/14910.

[10] Wang Licheng, Wang Licheng e Wang Licheng. «Privacy-preserving and efficient public key encryption with keyword search based on cp-abe in cloud». Em: *Review of Scientific Instruments* 4.3 (mar. de 2020), pp. 4236–4243. url: https://www.mdpi.com/2410-387X/4/4/28.

[11] Balachandran Santhi e Raj Helen Wilfred. «A survey of data anonymization techniques for privacy-preserving mining in bigdata». Em: *Review of Scientific Instruments* 4.3 (mar. de 2018), pp. 4236–4243. url: https://thescipub.com/pdf/jcssp.2020.194.201.pdf.

[12] Ren Kui, Jiang Jinghua e Zheng Yifeng. «Toward Privacy-Preserving Personalized Recommendation Services». Em: *Review of Scientific Instruments* 4.3 (mar. de 2018), pp. 4236–4243. url: https://thescipub.com/pdf/jcssp.2020.194.201.pdf.

[13] Tianqing Zhu et al. «Differentially Private Data Publishing and Analysis: A Survey». Em: *Review of Scientific Instruments* 4.3 (mar. de 2017), pp. 4236–4243. url: https://ieeexplore.ieee.org/ielaam/69/7970215/7911185-aam.pdf.

[14] Azman Samsudin. «A survey of homomorphic encryption for outsourced big data». Em: *Review of Scientific Instruments* 4.3 (mar. de 2016), pp. 4236–4243. url: https://www.researchgate.net/publication/307637971_A_Survey_of_Homomorphic_Encryption_for_Outsourced_Big_Data_Computation.

[15] T Dumitrescu e D Evans. «SMPAI: Secure Multi-Party Computation for Federated Learning». Em: *Review of Scientific Instruments* 4.3 (mar. de 2020), pp. 4236–4243. url: https://www.jpmorgan.com/content/dam/jpm/cib/complex/content/technology/ai-research-publications/pdf-9.pdf.

[16] Ashwin Machanavajjhala, Johannes Gehrke e Daniel Kifer. «L-Diversity: Privacy Beyond k-Anonymity». Em: *Review of Scientific Instruments* 4.3 (mar. de 2016), pp. 4236–4243. url: https://personal.utdallas.edu/~mxk055100/courses/privacy08f_files/ldiversity.pdf.

[17] Chaobin Liu et al. «A novel privacy preserving method for data publication». Em: *Review of Scientific Instruments* 4.3 (mar. de 2019), pp. 4236–4243. url: https://www.sciencedirect.com/science/article/pii/S0020025519305614.

[18] Hilder Vitor Lima Pereira. «Bootstrapping fully homomorphic encryption over the integers in less than one second». Em: *Review of Scientific Instruments* 4.3 (mar. de 2016), pp. 4236–4243. url: https://eprint.iacr.org/2020/995.pdf.

[19] ILACHANDRAKAR e VISHWANATH R HULIPALLED. «EFFICIENT SCHEME FOR PRIVACY PRESERVING REAL TIME BIG DATA MINING». Em: *Review of Scientific Instruments* 4.3 (mar. de 2021), pp. 4236–4243. url: https://tianjindaxuexuebao.com/dashboard/uploads/40.GPQBE.pdf.

[20] ILACHANDRAKAR e VISHWANATH R HULIPALLED. «StreamSketch: Exploring Multi-Modal Interactions in Creative Live Streams». Em: *Review of Scientific Instruments* 4.3 (mar. de 2020), pp. 4236–4243. url: https://eprints.whiterose.ac.uk/151333/7/main.pdf.

[21] Min Zhao e Yang Geng. «Homomorphic Encryption Technology for Cloud Computing». Em: *Review of Scientific Instruments* 4.3 (mar. de 2020), pp. 4236–4243. url: https://www.sciencedirect.com/science/article/pii/S1877050919307811.

[22] Shereen Mohamed Fawaz Nahla Belal Adel ElRefaey e Mohamed Waleed Fakh. «A Comparative Study of Homomorphic Encryption Schemes Using Microsoft SEAL». Em: *Review of Scientific Instruments* 4.3 (mar. de 2021), pp. 4236–4243. url: https://iopscience.iop.org/article/10.1088/1742-6596/2128/1/012021/meta.

[23] Image. «Secure your data with AES-256 encryption». Em: *Review of Scientific Instruments* 4.3 (mar. de 2029), pp. 4236–4243. url: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.atpinc.com%2Fblog%2Fwhat-is-aes-256-encryption&psig=AOvVaw3p5quzqY4th6WomMd7z1U7&ust=1696863943367000&source=images&cd=vfe&opi=89978449&ved=0CBEQjRxqFwoTCOjrtvbc5oEDFQAAAAAdAAAAABAE.

Apêndice A

Imagens

```
[1]: pip install phe

Collecting phe
  Downloading phe-1.5.0-py2.py3-none-any.whl (53 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 53.7/53.7 kB 4.8 MB/s eta 0:00:00
Installing collected packages: phe
Successfully installed phe-1.5.0
Note: you may need to restart the kernel to use updated packages.
```

```
[5]: from phe import paillier

public_key, private_key = paillier.generate_paillier_keypair()
#secret_number_list = [3.141592653, 300, -4.6e-12]
secret_number_list = [5]
encrypted_number_list = [public_key.encrypt(x) for x in secret_number_list]
# Encrypted numbers can be added together or added to a constant
encrypted_sum = sum(encrypted_number_list) + public_key.encrypt(6)
# Decryption happens as expected.

#print ("Encrypted :", encrypted_sum)
print("Decrypted result :", private_key.decrypt(encrypted_sum))

Decrypted result : 11
```

Figura A.1: Encriptação homomórfica com Paillier em Python

In [15]:
```
# Note: The PySEAL library provides Homomorphic encryption.
# https://github.com/Huelse/SEAL-Python
# You will need to install the library to use this code.

import seal
from seal import EncryptionParameters, SEALContext, KeyGenerator, Encryptor, Decryptor, \
    Plaintext, Ciphertext, Evaluator

parms = EncryptionParameters(seal.SCHEME_TYPE.BFV)
parms.set_poly_modulus_degree(4096)
parms.set_coeff_modulus(seal.CoeffModulus.BFVDefault(4096))
parms.set_plain_modulus(256)

context = SEALContext.Create(parms)
keygen = KeyGenerator(context)
public_key = keygen.public_key()
secret_key = keygen.secret_key()

encryptor = Encryptor(context, public_key)
evaluator = Evaluator(context)
decryptor = Decryptor(context, secret_key)

plain1 = Plaintext("5")
plain2 = Plaintext("6")
encrypted1 = Ciphertext()
encrypted2 = Ciphertext()

encryptor.encrypt(plain1, encrypted1)
encryptor.encrypt(plain2, encrypted2)

evaluator.add_inplace(encrypted1, encrypted2)
result = Plaintext()
decryptor.decrypt(encrypted1, result)

print("Decrypted result: ", result.to_string())  # Should print 11 (5 + 6)
```

```
Decrypted result:  11
```

Figura A.2: Adição homomórfica segura usando o Microsoft SEAL

[18]:
```
import random
import functools

def secret_share(secret, n_shares=3, min_shares=2, prime=101):
       coefficients = [secret] + [random.randrange(prime) for _ in range(min_shares - 1)]
       return [(i, polynomial(i, coefficients, prime)) for i in range(1, n_shares +1 )]

def polynomial(x, coefficients, prime):
       return sum ((coef * x **i) % prime for i, coef in enumerate(coefficients))
shares= secret_share(1234, n_shares=5, min_shares=3, prime=5000)
print(shares)
```

```
[(1, 6037), (2, 5364), (3, 4215), (4, 2590), (5, 5489)]
```

Figura A.3: Implementação da Partilha de Segredos para Computação Multipartidária Segura

In [21]:
```
import pandas as pd
from pydp.algorithms.laplacian import BoundedMean

# Load the Adult dataset.
data = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data', header=None)

# Rename columns
data.columns = ['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occupation',
                'relationship', 'race', 'sex', 'capital_gain', 'capital_loss', 'hours_per_week', 'native_country', 'income']

# Initialize the differential privacy parameters.
epsilon = 1.0

# Instantiate a BoundedMean object.
bmean = BoundedMean(epsilon)

# Add the age of each person in the dataset, converted to integers.
ages = [int(age) for age in data['age']]
bmean.add_entries(ages)

# Calculate the differentially private mean.
dp_mean = bmean.result()

print("Differentially private mean age: ", dp_mean)
```

```
Differentially private mean age:  38.606762134443585
```

Figura A.4: Privacidade diferencial

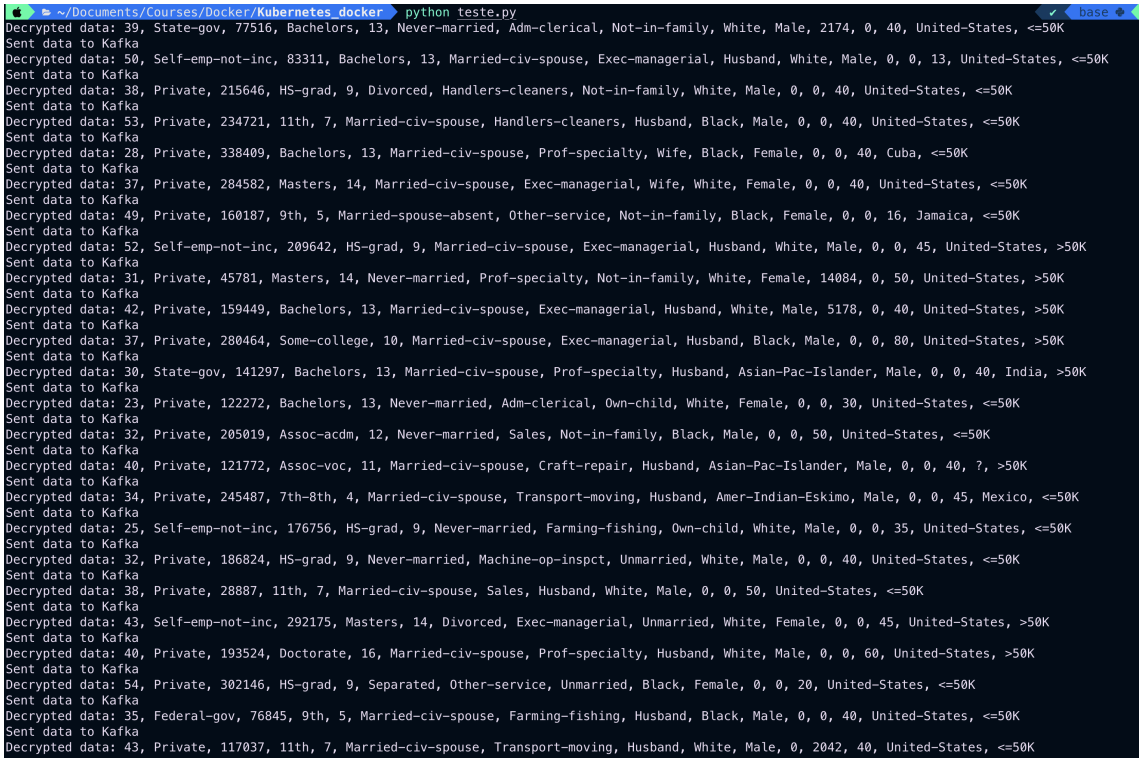

Figura A.5: Pipeline segura com encriptação AES e integração Kafka

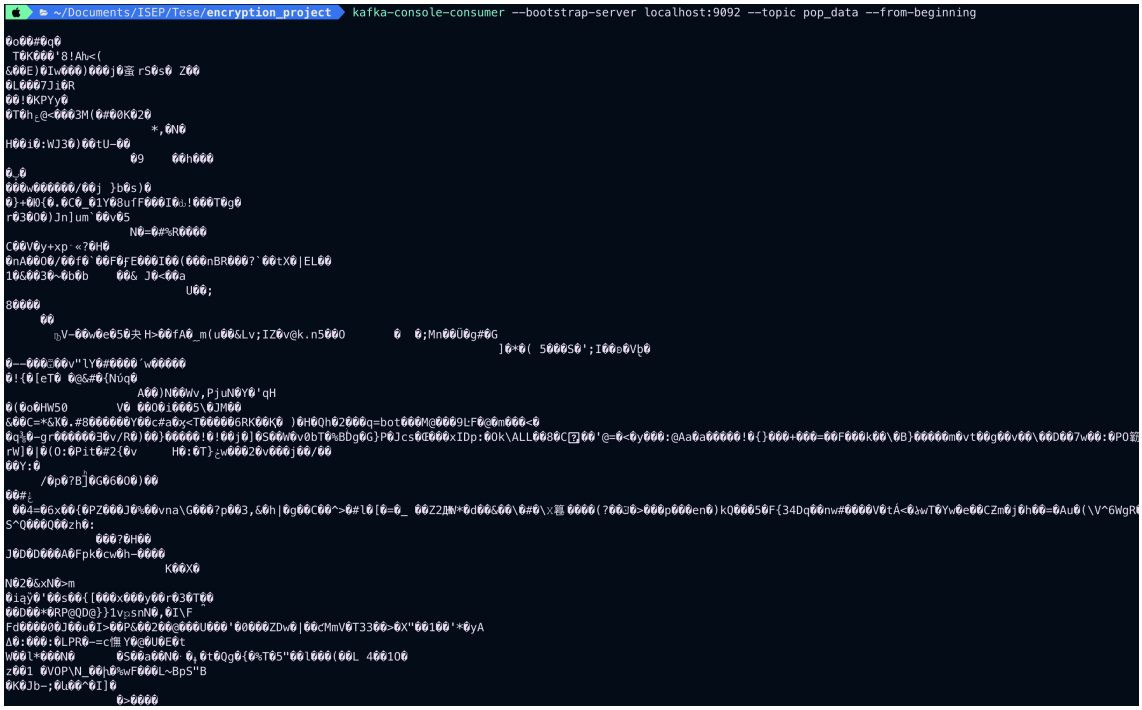

Figura A.6: Encriptação AES e integração Kafka

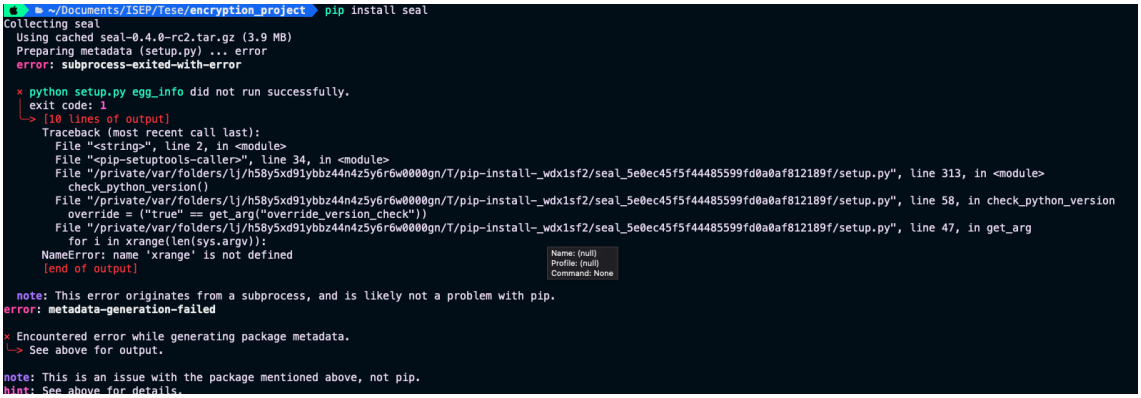

Figura A.7: Instalação biblioteca