Escola Politécnica da Universidade de São Paulo

THIAGO PEREIRA

# BOAS PRÁTICAS DE SEGURANÇA EM BIG DATA COM FOCO NA ÁREA DE DADOS

SÃO PAULO

2023

THIAGO PEREIRA

# BOAS PRÁTICAS DE SEGURANÇA EM BIG DATA COM FOCO NA ÁREA DE DADOS

**Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Engenharia de Dados e Big Data.**

**Área de concentração: Tecnologia da Informação – Engenharia/ Tecnologia/ Gestão**

**Orientador: Prof. WAGNER LUIZ ZUCCHI**

**SÃO PAULO**

**2023**

## FICHA CATALOGRÁFICA

Pereira, Thiago
BOAS PRÁTICAS DE SEGURANÇA EM BIG DATA COM FOCO NA ÁREA DE DADOS / T. Pereira -- São Paulo, 2024.
46 p.

Monografia (Especialização em ENGENHARIA DE DADOS E BIG DATA) - Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1.Big Data 2.Engenharia de Dados 3.Segurança em Big Data 4.Boas Práticas I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia II.t.

## AGRADECIMENTOS

Dedico este trabalho aos meus familiares, amigos e professores, que sempre me apoiaram e jamais me deixaram desistir.

## ESPECIALIZAÇÃO EM ENGENHARIA DE DADOS E BIG DATA

Coord.: Prof. Solange N. Alves de Souza
Vice-Coord.: Prof. Pedro Luiz Pizzigatti Corrêa

**Perspectivas profissionais alcançadas com o curso:**

[*À área de dados, está muito aquecida então recebo propostas em diversos níveis, porém o que mudou foi que passei a receber propostas em grande quantidade para o nível sênior e gestão.O reconhecimento que o mercado possui em relação ao nome da Universidade de São Paulo é incrível e ter no currículo em nível especialização, elevou o meu nível profissional.*]

## Resumo

O desenvolvimento de grandes plataformas de dados, permitiu que mais setores dentro das empresas pudessem utilizar os dados para algum tipo de análise. Olhando apenas para o negócio isso é ótimo, por aumentar a conscientização da tomada de decisão baseada em dados. Esse desenvolvimento trouxe muitos benefícios para as empresas, porém ao incluir um número maior de usuários com acesso aos dados, aumenta significativamente a exposição a riscos, principalmente ao incluir usuários com diferentes perfis.

Nesse contexto, é importante entender o quanto a segurança em big data evoluiu em relação ao tamanho crescente da população que demanda por consumo de dados. É necessário considerar quais são as boas práticas que o mercado propõe para acompanhar com segurança o crescimento operacional, levantar as principais estratégias de defesa em sistemas com grande volume de dados. As práticas aqui discutidas servem para elaborar um arcabouço de segurança, para sistemas de Big Data.

**Palavras-chave**: Hadoop, Segurança em Big Data e Riscos em Big Data.

## Abstract

The development of large data platforms has allowed more sectors within companies to use data for some type of analysis. Just looking at the business this is great, as it increases awareness of data-driven decision making. This development has brought many benefits to companies, but by including a greater number of users with access to data, it significantly increases exposure to risks, especially when including users with different profiles.

In this context, it is important to understand how much big data security has evolved in relation to the growing size of the population that demands data consumption. It is necessary to consider what good practices the market proposes to safely monitor operational growth and identify the main defense strategies in systems with a large volume of data. The practices discussed here serve to develop a security framework for Big Data systems.

**Keywords**: Hadoop, Big Data Security, Big Data Risks.

# 1 Introdução

A emergência da era Big Data tem reconfigurado o panorama da segurança da informação, introduzindo uma série complexa de desafios e oportunidades. Neste contexto, a segurança em Big Data não se limita apenas à proteção contra acessos não autorizados, mas abrange a preservação da confidencialidade, integridade, e disponibilidade de um volume crescente e diversificado de dados. Dados estes, que vão desde informações pessoais até dados corporativos críticos, cuja exposição ou comprometimento poderia ter consequências devastadoras tanto para indivíduos quanto para organizações.

Os desafios de segurança em Big Data são multifacetados, incluindo a dificuldade de gerenciar a privacidade de dados em vastos conjuntos de informações, a necessidade de implementar medidas de proteção robustas contra ataques sofisticados, e a urgência de assegurar a integridade dos dados em um ambiente dinâmico e em constante expansão. A problemática central reside na capacidade de desenvolver e implementar um arcabouço de segurança que não apenas responda aos desafios atuais, mas também seja adaptável às evoluções futuras da tecnologia e das ameaças cibernéticas.

Portanto, a adoção de boas práticas em segurança em Big Data é fundamental. Essas práticas englobam uma gama de estratégias e técnicas, bem como uma cultura de conscientização de segurança entre os usuários. Além disso, estas práticas devem estar em conformidade com regulamentações globais de proteção de dados, refletindo um compromisso ético com a privacidade e a confiança dos usuários.

Este trabalho tem como objetivo principal delinear um conjunto de boas práticas recomendadas para mitigar riscos de segurança. Espera-se que os resultados apresentados forneçam insights valiosos para profissionais da área, oferecendo diretrizes claras para a implementação de estratégias de segurança eficazes em sistemas de Big Data, contribuindo assim para a proteção de ativos de dados críticos em um ambiente cada vez mais digitalizado e vulnerável a ameaças.

O conjunto de práticas úteis em ambientes de Big Data adaptável e escalável de modo que as ações possam ser personalizadas de acordo com as necessidades específicas de cada organização e os tipos de dados que ela manuseia. Estas práticas incluem a classificação criteriosa de dados para identificar quais informações requerem níveis mais elevados de proteção, a segmentação de redes para limitar o acesso a dados sensíveis, e o emprego de soluções de segurança avançadas que empreguem inteligência artificial e aprendizado de máquina para detectar e responder a ameaças em tempo real.

A colaboração e o compartilhamento de informações sobre ameaças e vulnerabilidades entre organizações e dentro da comunidade global de segurança da informação são vitais para a criação de um ambiente de Big Data mais seguro. À medida que os adversários se tornam mais sofisticados em suas técnicas, a cooperação torna-se uma ferramenta poderosa na identificação de padrões de ataque emergentes e na rápida disseminação de estratégias de mitigação.

## 1.1 Objetivo

Os seguintes objetivos serão abordados:

- Identificar as principais ameaças à segurança em sistemas de Big Data.
- Relacionar os serviços e mecanismos de segurança com as ameaças típicas em um ambiente de Big Data
- Descrever as boas práticas de segurança recomendadas para sistemas de Big Data.

Espera-se que o estudo possa ser útil para os profissionais responsáveis por segurança e fornecer um direcionamento de como abordar a questão de Big Data, dentro de uma corporação.

## 1.2 Desafios em Segurança de Big Data

**a) Escalabilidade e Volume:** Um dos principais desafios em segurança em Big Data é a enorme escala em que os dados são gerados e processados. As soluções de segurança devem ser capazes de lidar com grandes volumes de informações sem comprometer a eficiência.

**b) Variedade de Dados:** Big Data abrange diversos tipos de dados, desde textos não estruturados a registros estruturados, aumentando a complexidade da segurança.

**c) Velocidade de Processamento:** A necessidade de processar dados em tempo real requer medidas de segurança que não prejudiquem o desempenho.

**d) Privacidade e Consentimento:** Com a coleta de informações pessoais em larga escala, a privacidade dos dados tornou-se um problema crítico. Garantir o consentimento apropriado para coleta e uso de dados torna-se um desafio.

**e) Ameaças Internas e Externas:** Tanto ameaças internas quanto externas precisam ser consideradas, incluindo ataques maliciosos e acesso não autorizado.

**f) Compliance Regulatório:** Muitos setores têm regulamentações rigorosas sobre a segurança de dados, o que adiciona complexidade à conformidade.

**g) Ameaças Avançadas:** A sofisticação das ameaças cibernéticas está em constante evolução. Isso requer soluções de segurança igualmente avançadas para identificar e combater ameaças em Big Data.

## 1.3 Algumas das ameaças mais comuns em ambientes de Big Data incluem

**a) Vazamento de Dados:** O vazamento de informações sensíveis pode ocorrer devido a falhas de segurança, configurações inadequadas ou ações maliciosas, comprometendo a privacidade e confidencialidade dos dados.

**b) Ataques de Injeção de Código:** Incluem ataques de injeção de SQL e NoSQL, nos quais os invasores inserem comandos maliciosos para acessar, modificar ou excluir dados do sistema de Big Data.

**c) Negligência Interna:** Erros de configuração e falhas humanas que comprometem a segurança.

**d) Ataques de Negação de Serviço (DDoS):** Ataques que visam sobrecarregar os recursos do sistema, tornando-o inacessível para usuários legítimos. Isso pode resultar na interrupção dos serviços de Big Data e na perda de disponibilidade dos dados.

**e) Ataques de Engenharia Social:** Manipulação de usuários para obter acesso não autorizado.

**f) Exploração de Vulnerabilidades de Software:** Vulnerabilidades em softwares utilizados em sistemas de Big Data podem ser exploradas por hackers para obter acesso não autorizado, comprometendo a segurança e integridade dos dados.

**g) Roubo de Identidade e Acesso Não Autorizado:** A obtenção de credenciais de acesso legítimas pode permitir que invasores acessem indevidamente os sistemas de Big Data, comprometendo a integridade e confidencialidade dos dados.

**h) Malware e Ransomware:** A infecção por malware ou ransomware pode resultar na criptografia ou exclusão de dados, causando interrupções nos

serviços e resultando em perda de dados importantes para as operações da empresa.

**i) Falhas de Segurança na Nuvem:** Como muitos sistemas de Big Data são implantados em ambientes de nuvem, falhas de segurança nesses ambientes podem expor os dados a riscos de acesso não autorizado, vazamento de informações e interrupção dos serviços.

**j) Falta de Controles de Acesso Adequados:** A falta de controles de acesso robustos pode permitir que usuários não autorizados obtenham acesso aos dados sensíveis armazenados em sistemas de Big Data, comprometendo a segurança e a privacidade das informações.

**l) Fraude e Manipulação de Dados:** A manipulação maliciosa de dados em sistemas de Big Data pode levar a análises incorretas e decisões comerciais equivocadas. Isso pode ocorrer tanto por parte de usuários internos quanto externos.

**m) Falta de Conformidade Regulatória:** A não conformidade com regulamentações de segurança e privacidade, como LGPD, GDPR e outras leis e regulamentações locais, pode resultar em multas substanciais e danos à reputação da empresa, além de expor os dados a riscos adicionais de segurança.

Essas ameaças ressaltam a importância de implementar medidas de segurança robustas e adotar uma abordagem proativa para proteger os sistemas de Big Data contra possíveis ataques e violações de segurança.

# 2 Serviços e Mecanismos de Segurança

Este capítulo discute os serviços e mecanismos de segurança que servem como pilares para as boas práticas em Big Data.

A confidencialidade de dados representa a primeira linha de defesa na proteção de dados, convertendo informações sensíveis em formatos codificados que só podem ser acessados por indivíduos autorizados. Ela se destaca pela sua capacidade de assegurar tanto os dados em trânsito quanto em repouso, mitigando os riscos de interceptação ou acesso indevido. A utilização de algoritmos reconhecidamente confiáveis e a gestão eficaz das chaves de criptografia são cruciais para a eficiência dessa técnica. A confidencialidade permite que, mesmo no caso de um acesso indevido aos dados, as informações comprometidas permaneçam ininteligíveis para os invasores.

A confidencialidade é um serviço de segurança geralmente realizado pela técnica de criptografia dos dados.

Os serviços de autenticação e de controle de acesso formam o cerne das estratégias de segurança em Big Data. A autenticação garante que apenas usuários verificados tenham acesso aos sistemas, enquanto o controle de acesso limita as ações que esses usuários podem executar, baseando-se no princípio do menor privilégio. A implementação de sistemas de autenticação multifatorial, que requerem mais de uma forma de verificação da identidade do usuário, aumenta significativamente a segurança, dificultando ataques de força bruta ou outras tentativas de acesso não autorizado.

O serviço de monitoramento e a detecção de ameaças emergem como vitais na identificação de atividades suspeitas dentro dos sistemas de Big Data. O uso de soluções de inteligência artificial e aprendizado de máquina para analisar padrões de acesso e comportamento dos usuários permite a identificação precoce de potenciais vulnerabilidades e ataques em andamento. Essas tecnologias habilitam as organizações a responderem rapidamente a ameaças, minimizando os danos potenciais.

A gestão de vulnerabilidades é outro serviço essencial, envolvendo a identificação, classificação e mitigação de vulnerabilidades dentro do ecossistema de Big Data. A realização de auditorias de segurança regulares e a aplicação de patches e atualizações de segurança de forma oportuna são práticas fundamentais para manter a robustez dos sistemas contra ameaças externas e internas.

Finalmente, a conformidade regulatória e a governança de dados são aspectos fundamentais que não só asseguram que as organizações sigam às leis e às normas aplicáveis à segurança e privacidade de dados, mas também estabelecem um quadro de responsabilidade e transparência. A aderência às regulamentações, como o GDPR na União Europeia e a LGPD no Brasil, não apenas protege as organizações de penalidades legais, mas também fortalece a confiança dos usuários nos sistemas de Big Data, garantindo que suas informações estejam seguras e sejam tratadas de forma ética.

A incorporação desses serviços de segurança em uma estratégia coesa e adaptável é vital para assegurar a proteção efetiva dos dados em Big Data, permitindo que as organizações explorem o potencial desses dados de forma segura e responsável.

## 2.1 Segurança da informação

De acordo com Spivey e Echeverria (2015), a segurança da informação é modelada pelo conceito CIA, que abrange confidencialidade, integridade e disponibilidade. Esses três componentes fundamentais podem ser aplicados a diversos sistemas de informação, plataformas de computação e também ao ambiente Hadoop.

## 2.2 Confidencialidade

A confidencialidade de dados desempenha um papel indispensável na segurança em Big Data. Com o aumento exponencial da quantidade de dados gerados e armazenados, é essencial garantir a confidencialidade e integridade dessas informações sensíveis. A confidencialidade permite proteger os dados por meio da transformação deles em um formato ilegível para qualquer pessoa que não possua a chave correta. Dessa forma, mesmo que ocorra um acesso indevido, os dados permanecem inacessíveis e ininteligíveis para os invasores (SANTOS, BRANCO, TEFFÉ, 2016).

Existem diversas técnicas utilizadas na proteção de dados em ambientes de Big Data. A técnica básica para obtenção da confidencialidade é a criptografia, que pode ser simétrica ou assimétrica. A criptografia é dita simétrica quando somente uma chave é usada para codificar e decodificar os dados. A criptografia simétrica é muito usada quando se conhecem as partes que irão acessar os dados antes do momento de acesso. A criptografia assimétrica utiliza diferentes chaves para codificação e decodificação. Ela é usada quando o público que irá utilizar os dados é indeterminado.

Há inúmeros algoritmos de criptografia em uso atual. Entre os principais, destacam-se o AES (Advanced Encryption Standard), o RSA (Rivest-Shamir-Adleman) e o DES (Data Encryption Standard). O AES é amplamente utilizado por sua eficiência e segurança comprovada, sendo adotado como padrão pelo governo dos Estados Unidos. O RSA é um algoritmo assimétrico baseado na dificuldade de fatorar números primos grandes, enquanto o DES e o 3-DES são algoritmos simétricos que utilizam uma chave compartilhada entre remetente e destinatário (OLIVEIRA, PRADO, 2023).

A criptografia simétrica e assimétrica apresenta vantagens e desvantagens distintas na proteção de dados em Big Data. A criptografia simétrica é mais rápida e eficiente do que a assimétrica, pois utiliza a mesma chave tanto para a criptografia quanto para a descriptografia dos dados. No entanto, a principal desvantagem da criptografia simétrica é a necessidade de compartilhar essa chave entre as partes

envolvidas, o que pode ser um desafio em ambientes de Big Data. Já a criptografia assimétrica utiliza um par de chaves, uma pública e outra privada. No entanto, seu desempenho é inferior ao da criptografia simétrica (AMARO, D. P.; DROZDA, F. O., 2019).

O gerenciamento de chaves é um serviço auxiliar essencial na criptografia de dados em Big Data. As chaves são responsáveis por garantir a segurança dos dados criptografados e devem ser armazenadas e protegidas adequadamente. Além disso, é necessário estabelecer políticas de rotação das chaves para evitar que elas se tornem obsoletas ou comprometidas ao longo do tempo. O gerenciamento eficiente das chaves também envolve a definição de políticas de acesso e controle sobre quem pode acessar e utilizar as chaves (DUTRA, MACEDO, 2016).

As técnicas de criptografia homomórficas desempenham um papel importante na segurança em Big Data. Essa técnica permite realizar operações matemáticas diretamente nos dados criptografados, sem a necessidade de descriptografá-los previamente. Isso possibilita o processamento seguro dos dados sem expor seu conteúdo sensível, o que é especialmente relevante em ambientes onde a privacidade e confidencialidade são essenciais (LIMA, 2017).

A implementação da criptografia de dados em ambientes de Big Data enfrenta diversos desafios. Um dos principais desafios é o desempenho, uma vez que a criptografia pode impactar negativamente o tempo necessário para processar grandes volumes de dados. Além disso, a escalabilidade também é um desafio, pois é preciso garantir que a criptografia seja aplicada de forma eficiente em ambientes com grande quantidade de dados e alta velocidade de processamento. Outro desafio é a compatibilidade entre diferentes sistemas e plataformas, uma vez que é necessário garantir a interoperabilidade dos algoritmos de criptografia utilizados (LINS, 2021).

## 2.3 Identificação e Autenticação de usuários

A identificação e autenticação de usuários em sistemas de Big Data desempenha um papel indispensável na garantia da segurança e integridade dos dados armazenados. Através desse serviço, é possível verificar a identidade dos usuários que acessam o sistema, bem como garantir que apenas usuários autorizados tenham acesso aos dados sensíveis. Além disso, a identificação e autenticação de usuários permite rastrear as atividades realizadas no sistema, facilitando a responsabilização em caso de incidentes de segurança (LUZ, 2018).

Existem diversas técnicas comumente utilizadas para a identificação e autenticação de usuários em sistemas de Big Data. Uma das técnicas mais utilizadas é a autenticação baseada em senhas. Nesse método, os usuários fornecem um nome de usuário e uma senha para acessar o sistema. Outra técnica bastante utilizada é a autenticação baseada em certificados digitais, onde os usuários possuem um certificado digital que ateste sua identidade. Além disso, existem também técnicas biométricas, como reconhecimento facial ou impressão digital, que podem ser utilizadas para autenticação (LINS, 2021).

Recentemente, têm ocorrido avanços significativos na área da identificação e autenticação de usuários em sistemas de Big Data. Uma das principais tendências é o uso de técnicas de aprendizado de máquina para melhorar a precisão e eficiência do processo de identificação. Algoritmos avançados podem ser treinados com grandes volumes de dados para reconhecer padrões e identificar possíveis ameaças ou comportamentos suspeitos. Além disso, estão sendo desenvolvidas soluções baseadas em blockchain, que permitem garantir a integridade e autenticidade dos dados de identificação dos usuários, aumentando ainda mais a segurança do sistema (OLIVEIRA, 2023).

Vamos aprofundar um pouco mais sobre essas novidades:

- **Aprendizado de Máquina na Identificação de Usuários**: O uso de técnicas de aprendizado de máquina representa um marco na segurança de Big Data. Com a capacidade de analisar grandes volumes de dados de forma rápida e

eficiente, os algoritmos de aprendizado de máquina podem identificar padrões sutis que indicam atividades maliciosas ou comportamentos suspeitos. Por exemplo, eles podem detectar anomalias nos padrões de acesso, como tentativas de acesso não autorizado ou atividades incomuns fora do horário comercial.

- **Algoritmos Avançados para Reconhecimento de Padrões**: Os algoritmos avançados empregados nesse contexto são capazes de reconhecer padrões complexos em conjuntos de dados massivos. Eles podem identificar correlações entre diversos pontos de dados que seriam difíceis de serem percebidas por sistemas tradicionais de segurança. Isso permite uma detecção mais precisa e eficaz de ameaças em tempo real, proporcionando uma resposta rápida a potenciais incidentes de segurança.
- **Soluções Baseadas em Blockchain**: O blockchain, famoso por sua aplicação em criptomoedas como o Bitcoin, também está sendo explorado para melhorar a segurança em Big Data. Ao usar o blockchain para armazenar e gerenciar informações de identificação de usuários, é possível garantir a integridade e autenticidade desses dados. Cada transação é registrada em blocos encadeados de forma imutável e distribuída, o que torna extremamente difícil para os invasores adulterarem ou comprometerem as informações de identificação

## 2.4 Controle de acesso ao dados

O serviço de controle de acesso aos dados em sistemas de Big Data é de extrema importância, considerando a quantidade e sensibilidade das informações armazenadas. Esses sistemas lidam com grandes volumes de dados, que podem incluir informações pessoais, financeiras e estratégicas das organizações. Portanto, garantir que apenas usuários autorizados tenham acesso a essas informações é indispensável para proteger a integridade e confidencialidade dos dados (FRANCO, 2020).

Dentre as principais técnicas utilizadas para o controle de acesso aos dados em ambientes de Big Data, destacam-se a autenticação, autorização e criptografia, que já foram descritas anteriormente. Um diferencial importante dos ambientes Big Data é a necessidade de definir quais dados cada usuário pode acessar por papéis. Uma abordagem do tipo “tudo ou nada” é praticamente inútil. Com o objetivo de refinar o processo de acesso foi criado o projeto Apache Sentry (CLOUDERA, 2021).

No futuro, espera-se que o controle de acesso aos dados em sistemas de Big Data seja impulsionado pelo uso crescente da inteligência artificial e da análise comportamental. Essas tecnologias podem ser utilizadas para identificar possíveis violações no acesso aos dados, analisando padrões de comportamento suspeitos ou atividades incomuns. Isso permitirá uma detecção mais rápida e precisa de ameaças, aumentando ainda mais a segurança dos sistemas de Big Data (MENDES, 2022).

Para fornecer contexto, se um usuário geralmente acessa apenas certos tipos de dados durante o horário comercial e de repente passa a acessar informações confidenciais durante a noite, isso pode disparar um alerta indicando uma possível violação de segurança. Além disso, a inteligência artificial tem a capacidade de analisar o histórico de comportamento dos usuários e detectar discrepâncias sutis que possam representar ameaças em potencial. Por exemplo, se um funcionário típico de repente inicia uma grande quantidade de consultas em um curto espaço de tempo, isso poderia levantar suspeitas de uma tentativa de extração massiva de dados para fins indevidos.

# 3 Boas Práticas

De acordo com (Spivey e Echeverria, 2015) os principais serviços de segurança que devem ser implementados em ambientes de Big Data são a autenticação por meio de nome de usuário e senha, tokens que são gerados em cada acesso, uso do Kerberos como sistema de controle de autenticação e camadas adicionais com restrições de acesso por setor e/ou hierarquia.

A implementação desses serviços deve ser realizada segundo boas práticas em segurança para proteger informações sensíveis e evitar violações de privacidade. A definição dessas boas práticas é um conhecimento gerado pelas rotinas operacionais desse tipo de ambiente em diversos tipos de empresas e que é compartilhado com a comunidade profissional de modo a tornar todo o ecossistema mais seguro.

Como se viu na seção 2 deste trabalho, as principais ameaças à segurança em Big Data incluem ataques cibernéticos, vazamento de dados e uso indevido de informações pessoais. Os ataques cibernéticos podem ocorrer por meio de diferentes técnicas, como phishing, malware e ataques de negação de serviço. O vazamento de dados pode ocorrer tanto por falhas internas quanto por ações maliciosas externas, resultando na exposição indevida das informações.

As boas práticas de segurança não dizem respeito apenas à escolha de serviços e técnicas que serão utilizados no ambiente. Não adianta implantar um sistema de segurança sem estabelecer políticas e procedimentos claros. Isso inclui definir responsabilidades específicas para cada área envolvida no processo, como TI, jurídico e compliance. Cada departamento deve ter conhecimento das suas obrigações em relação à segurança dos dados e seguir as diretrizes estabelecidas. Além disso, é essencial fornecer treinamento adequado aos funcionários, para que eles estejam cientes das melhores práticas de segurança e saibam como lidar com possíveis ameaças (OLIVEIRA, 2023).

Os profissionais de Tecnologia da Informação (TI) e Segurança, por exemplo, estão na linha de frente, encarregados de projetar e implementar as infraestruturas que suportam o Big Data. Para eles, a segurança começa na estruturação segura

dos sistemas, garantindo que as arquiteturas sejam desenvolvidas com uma mentalidade de segurança desde o início.

Para os profissionais da área de dados, que frequentemente trabalham diretamente com grandes volumes de informações, a conscientização sobre a classificação de dados e as práticas de minimização de dados torna-se crucial. Eles precisam entender quais dados são sensíveis e como esses dados devem ser manuseados, garantindo que as informações pessoais sejam protegidas e que apenas os dados necessários para análises específicas sejam acessados. Este grupo também beneficia do uso de técnicas de anonimização e pseudonimização para proteger a privacidade dos indivíduos cujos dados estão sendo analisados.

No lado gerencial e de negócios, os líderes precisam manter-se atualizados com as regulamentações globais de proteção de dados e garantir que a empresa esteja em conformidade com estas leis.

Os profissionais jurídicos e de conformidade desempenham um papel importante na interpretação dessas regulamentações, aconselhando sobre a implementação de políticas de dados que não apenas protejam a organização de potenciais sanções legais, mas também assegurem a confiança dos clientes e usuários. Eles precisam trabalhar em estreita colaboração com todos os departamentos para garantir que as práticas de coleta, armazenamento e processamento de dados estejam em conformidade com as leis pertinentes e que qualquer incidente de segurança seja prontamente e adequadamente endereçado.

As boas práticas em segurança em Big Data abrangem uma série de aspectos que visam garantir não só a proteção dos dados armazenados e processados nesse ambiente, mas toda a estrutura de negócio, envolvendo os múltiplos aspectos mencionados.

A realização de testes regulares nos sistemas de segurança em Big Data é essencial para garantir a eficácia das medidas adotadas. A simulação de ataques é uma prática recomendada nesse contexto, pois permite identificar possíveis falhas nos controles de segurança e corrigi-las antes que sejam exploradas por atacantes reais. Os testes também podem ser utilizados para avaliar a capacidade de resposta dos sistemas a incidentes, verificando se os processos de detecção, análise e resposta estão funcionando adequadamente

No tópico seguinte, apresentam-se um conjunto de boas práticas de implementação de serviços de segurança, destacando a necessidade de monitoramento contínuo e atualizações regulares para garantir a proteção adequada dos dados sensíveis.

## 3.1 Boas práticas de segurança recomendadas para sistemas de Big Data

Para referenciar as boas práticas, uma boa orientação vem da comunidade OWASP, amplamente reconhecida por suas contribuições para a segurança de aplicações web. A OWASP (The Open Web Application Security Project) é uma organização global sem fins lucrativos, dedicada a identificar e mitigar as causas de insegurança em aplicações web. Entre os diversos recursos disponibilizados pela OWASP, destaca-se o "Guia para Desenvolvimento Seguro de Aplicações Web e Web Services" (OWASP, TOP 10), o qual oferece diretrizes detalhadas para desenvolvedores e projetistas de software visando promover soluções seguras para o desenvolvimento de aplicações web

A OWASP (Open Web Application Security Project) é uma comunidade global dedicada a aprimorar a segurança de software. Sua missão é tornar o software mais seguro, auxiliando organizações no desenvolvimento e na manutenção de aplicativos web confiáveis. Composta por desenvolvedores, profissionais de segurança, pesquisadores e entusiastas da segurança de software, a comunidade OWASP é reconhecida pelo seu "Top 10 de Vulnerabilidades de Aplicações Web", uma lista das dez vulnerabilidades mais críticas em aplicações web. Essa lista é atualizada periodicamente e amplamente utilizada por desenvolvedores e profissionais de segurança para identificar e corrigir vulnerabilidades em seus aplicativos.

Com base nos trabalhos da OWASP e na experiência do autor, apresenta-se a seguir um rol de boas práticas úteis para a implementação de serviços de segurança. Embora esse trabalho não diga respeito aos procedimentos de instalação da infraestrutura de TI, algumas recomendações nesta área foram incluídas por sua especificidade e relevância.

As práticas são apresentadas dentro de uma visão abrangente e generalista. É natural que adaptações sejam feitas para cada tipo específico de empresa e de data center.

## 3.2 Checklist de Boas Práticas, aplicáveis para ambientes de Big Data

### 3.2.1 Validação de Entrada de Dados

A validação de entrada assegura a entrada correta no banco de dados, prevenindo falhas por dados incorretos. Evita que dados formatados incorretamente causem problemas no banco de dados.

**a)** Realizar todas as validações de dados em um sistema confiável;

**b)** Deve haver uma rotina de validação de entrada centralizada para o aplicativo;

**c)** Verificar se os valores do cabeçalho em solicitações e respostas contém apenas caracteres ASCII;

**c)** Validar dados provenientes de fontes não confiáveis, como bancos de dados e streams de arquivos.

**e)** Garantir que a aplicação rejeite dados que falhem na validação, evitando assim a persistência de dados incorretos.

**f)** Verificar se o sistema suporta conjuntos de caracteres estendidos UTF-8 e validar após a decodificação.

**g)** Validar todos os dados provenientes dos clientes antes do processamento, incluindo parâmetros, campos de formulário, URLs e cabeçalhos HTTP.

**h)** Implementar controles adicionais para caracteres 'perigosos' na entrada de dados, como codificação dos dados de saída e trilhas de auditoria.

**i)** Se qualquer caractere potencialmente 'perigoso' precisa ser permitido na entrada de dados da aplicação, certifique-se que foram implementados

controles adicionais como codificação dos dados de saída, APIs específicas que fornecem tarefas seguras e trilhas de auditoria no uso dos dados pela aplicação. Como exemplo de caracteres potencialmente 'perigosos', temos: <, >, ", ', %, (, ), &, +, \, \', \"

### 3.2.2 Saída de Dados

Quando os dados são inseridos, o sistema deve codificá-los antes de serem utilizados em operações de Big Data. A codificação de saída transforma caracteres especiais em um formato seguro para evitar possíveis riscos no processamento ou interpretação dos dados. O framework Flask oferece recursos integrados para essa codificação(Pallets, 2024) .

**a)** Centralizar o controle de codificação em um servidor confiável.

**b)** Utilizar rotinas testadas para diferentes tipos de codificação de saída.

**c)** Codificar dados de acordo com o contexto, especialmente aqueles provenientes de fontes não confiáveis devido à natureza dinâmica e heterogênea do ambiente de Big Data. Por exemplo, sistemas de coleta automática de dados podem capturar informações de fontes não verificadas, como feeds de redes sociais ou formulários online, onde a autenticidade dos dados não é garantida. Portanto, mesmo após o processamento inicial, é essencial manter medidas de segurança, como a verificação da fonte dos dados e a aplicação de técnicas de codificação adequadas.

**d)** Codificar todos os caracteres, a menos que sejam conhecidos por serem seguros.

**e)** Realizar sanitização de dados provenientes de fontes não confiáveis para construir consultas SQL, XML e LDAP, além de comandos de sistema operacional.

### 3.2.3 Autenticação e gerenciamento de senha

A autenticação e gestão de senhas são vitais em ambientes de Big Data para assegurar a segurança dos dados, embora as senhas sejam consideradas um método de autenticação menos robusto. Nesse contexto, o sistema de controle de autenticação Kerberos emerge como uma solução confiável, fornecendo uma camada adicional de segurança. Além disso, a implementação de camadas adicionais com restrições de acesso por setor e/ou hierarquia fortalece ainda mais a segurança do ambiente. Esses processos não apenas confirmam a identidade do usuário, mas também requerem práticas adequadas de armazenamento de senhas, garantindo a confidencialidade e disponibilidade dos dados de forma abrangente.

**a)** Estabelecer e utilizar serviços de autenticação padronizados e testados.

**b)** Exigir que os requisitos de complexidade e comprimento de senha sejam cumpridos.

**c)** Desativar a conta após um número pré-definido de tentativas inválidas de login.

**d)** Implementar autenticação de múltiplos fatores para contas altamente sensíveis.

**e)** Validar os dados de autenticação somente ao término de todas as entradas de dados.

**f)** Implementar monitoramento para identificar ataques contra várias contas de usuário, utilizando a mesma senha.

**g)** Implementar hash de senha em um sistema confiável.

**h)** As respostas de falha de autenticação não devem indicar qual parte dos dados de autenticação estava incorreta. Por exemplo, em vez de “Nome de usuário inválido” ou “Senha inválida”, apenas use “Nome de usuário e / ou senha inválidos” para ambos. As respostas de erro devem ser verdadeiramente idênticas na exibição e no código-fonte.

**i)** Aplicar os requisitos de comprimento de senha estabelecidos por política ou regulamento.

**j)** Exigir autenticação para todas as páginas e recursos, exceto aqueles especificamente destinados ao público.

**l)** Todas as funções administrativas e de gerenciamento de contas devem ser pelo menos tão seguras quanto o mecanismo de autenticação principal.

### 3.2.4 Controle de acesso

É essencial para a segurança de software em ambientes de Big Data, autorizar ou recusar solicitações específicas e gerenciar credenciais, sendo apoiado por diversos serviços de segurança.

**a)** Utilizar apenas objetos do sistema que sejam confiáveis para realizar a tomada de decisões de autorização de acesso.

**b)** Restringir o acesso aos dados da aplicação somente aos usuários autorizados.

**c)** Restringir o acesso às URLs protegidas somente aos usuários autorizados.

**d)** Restringir o acesso às funções protegidas somente aos usuários autorizados.

**e)** Limitar o número de transações que um único usuário ou dispositivo pode executar em um determinado período de tempo.

**f)** Implementar a auditoria das contas de usuário e assegurar a desativação de contas não utilizadas.

**g)** A aplicação deve dar suporte a desativação de contas e encerramento das sessões quando encerrar a autorização do usuário.

### 3.2.5 Práticas criptográficas

Práticas criptográficas garantem a confidencialidade e integridade dos dados em cenários de Big Data, utilizando abordagens como criptografia simétrica ou de chave pública para transferência e armazenamento seguros, adaptadas às necessidades de segurança e aos riscos envolvidos.

**a)** Todas as funções de criptografia utilizadas para proteger dados sensíveis dos usuários da aplicação devem ser implementados em um sistema confiável (neste caso o servidor).

**b)** Quando ocorrer alguma falha nos módulos de criptografia, é crucial permitir que as falhas ocorram de modo seguro. Isso significa que o sistema deve ser projetado para lidar com falhas de forma a minimizar os impactos na segurança e na integridade dos dados. Isso pode envolver a implementação de mecanismos de recuperação e redundância para garantir a continuidade das operações e a proteção dos dados, mesmo em situações adversas.

**c)** Os módulos de criptografia usados pela aplicação devem ser compatíveis com a FIPS 140-2 ou padrão equivalente.

### 3.2.6 Tratamento e registro de erros

No contexto de Big Data, o tratamento de erros aborda como lidar com resultados inesperados ao receber entradas incomuns. Falhas nesse processo podem revelar detalhes sensíveis e criar pontos de entrada para ataques. Por exemplo, erros em consultas SQL podem expor vulnerabilidades no sistema, que permitam acesso não autorizado a dados de usuários.

**a)** Implemente mensagens de erro genéricas e páginas de erro personalizadas.

**b)** Garantir que os logs armazenam eventos importantes.

**c)** Registrar em log todas as tentativas de autenticação, especialmente as falhas de autenticação.

**d)** Registrar em log todas as falhas de controle de acesso.

**e)** Registrar em log todas as exceções lançadas pelo sistema.

**f)** Registrar em log todas as funções administrativas, inclusive as mudanças realizadas nas configurações de segurança.

### 3.2.7 Proteção de dados

Em ambientes de Big Data, ataques de roubo de dados são uma ameaça significativa, podendo resultar na divulgação indevida de informações críticas ou protegidas. Assim, a proteção de dados é essencial para evitar a adulteração, comprometimento ou perda de informações importantes.

**a)** Desabilitar a funcionalidade de auto completar em formulários com dados sensíveis, como os de autenticação.

**b)** Implementar política de privilégio mínimo para restringir acesso apenas ao necessário.

**c)** Não armazenar senhas ou informações confidenciais em texto aberto no lado cliente.

**d)** Proteger todas as cópias temporárias ou em cache de dados sensíveis no servidor contra acesso não autorizado.

**e)** Remover aplicações e documentação desnecessárias que possam expor informações importantes.

**f)** Desabilitar o cache no lado cliente de páginas com dados sensíveis.

**g)** Dar suporte à remoção de dados sensíveis quando não forem mais necessários.

### 3.2.8 Segurança de comunicação

A segurança de comunicação evita que interceptadores obtenham dados compreensíveis, enquanto a criptografia protege os dados confidenciais através do

protocolo TLS, no caso de comunicações Web, ou usando IPSec para outros tipos de comunicações

**a)** Utilizar criptografia na transmissão de todas as informações sensíveis.

**b)** Os certificados TLS devem ser válidos, possuir o nome de domínio correto, não estarem expirados e serem instalados com certificados intermediários, quando necessário.

### 3.2.9 Configuração do sistema

A configuração do sistema em big data engloba hardware, procedimentos e elementos distintos. Atacantes visam pontos fracos não corrigidos ou expor contas padrão para obter acesso ilegal ou informações.

**a)** Garantir que os servidores, frameworks e componentes do sistema estejam atualizados.

**b)** Restringir privilégios, remover funcionalidades desnecessárias e desativar métodos HTTP não utilizados para fortalecer a segurança do sistema.

**c)** Isolar o ambiente de desenvolvimento da rede de produção e prover acesso somente para grupos de desenvolvimento e testes. Os ambientes de desenvolvimento comumente são configurados de modo menos seguro do que os ambientes de produção. Assim, os atacantes podem usar esse diferencial para descobrir vulnerabilidades compartilhadas ou encontrar caminhos para explorar as vulnerabilidades.

### 3.2.10 Segurança de banco de dados

No contexto de big data, a segurança do banco de dados abrange tecnologias, políticas e procedimentos para preservar a Confidencialidade, Integridade e Disponibilidade (CID) dos dados.

**a)** Desative todas as contas padrão que não são necessárias para oferecer suporte aos requisitos de negócios.

**b)** Desligue todas as funcionalidades desnecessárias do banco de dados.

**c)** Use somente procedimentos armazenados para abstrair o acesso aos dados e para permitir a remoção de permissões para as tabelas base no banco de dados.

**d)** Utilizar validação de entrada e codificação de saída e assegurar a abordagem de meta caracteres para evitar ataques de injeção de SQL.

**e)** Usar procedimentos armazenados para abstrair o acesso aos dados e permitir a remoção das permissões das tabelas no banco de dados.

### 3.2.11 Gerenciamento de arquivos

Envolve proteger dados confidenciais, estabelecendo políticas robustas de controle de acesso e autorização. Isso é feito pelos administradores do sistema através de um sistema de gerenciamento de arquivos, atribuindo funções e diferentes níveis de acesso aos usuários para aumentar a eficácia.

**a)** Solicitar autenticação antes de permitir o upload de um arquivo.

**b)** Limitar os tipos de arquivos aceitos para apenas os necessários aos objetivos do negócio.

**c)** Validar os arquivos enviados pelo tipo esperado, verificando os cabeçalhos.

**d)** Escanear arquivos submetidos por usuários em busca de vírus e malwares.

**e)** Impedir ou restringir o upload de qualquer arquivo que possa ser interpretado pelo servidor web.

**f)** Nunca enviar o caminho absoluto do arquivo para o cliente.

### 3.2.12 Gerenciamento de memória

No contexto de Big Data, técnicas de gerenciamento de memória são fundamentais para manter eficiência na alocação e monitoramento de recursos, garantindo a disponibilidade e otimização do sistema.

**a)** Verificar se o buffer é tão grande quanto o especificado.

**b)** Ao usar funções que aceitam um determinado número de bytes para realizar cópias, esteja ciente de que se o tamanho do buffer de destino for igual ao tamanho do buffer de origem, esse processo não pode encerrar a sequência de caracteres com valor nulo (null).

**c)** Verificar os limites do buffer caso as chamadas a função sejam realizadas em um loop e verificar se não há nenhum perigo de escrever além do espaço alocado.

**d)** Truncar todas as strings de entrada para um tamanho razoável antes de passá-las para as funções de cópia e concatenação.

**e)** Encerre os recursos de modo específico, sem contar com o garbage collector na liberação dos recursos alocados para objetos de conexão, identificadores de arquivo, etc.

**f)** Usar pilhas não-executáveis, quando disponíveis.

**g)** Evitar o uso de funções reconhecidas por serem vulneráveis, por exemplo: printf, strcat, strcpy, etc.

### 3.2.13 Práticas Gerais de Codificação

O desenvolvimento seguro de aplicativos é essencial para garantir a integridade e a confiabilidade das operações em ambientes de big data, onde a manipulação e a análise de grandes volumes de dados são frequentes.

**a)** Utilizar APIs que embutem tarefas específicas para realizar tarefas do sistema operacional. Não permitir que a aplicação execute comandos diretamente no sistema operacional, especialmente através da utilização de shells de comando iniciados pela aplicação.

**b)** Utilize mecanismo de verificação de integridade por checksum ou hash para verificar a integridade do código interpretado, bibliotecas, arquivos executáveis e arquivos de configuração.

**c)** Inicialize explicitamente todas as variáveis e outros dados persistidos, durante a declaração ou antes do primeiro uso da variável.

### 3.2.14 Backup e Recuperação

Os mecanismos de backup e recuperação garantem que os dados não se percam em caso de falha ou ataque. Manter cópias de segurança atualizadas e sistemas de recuperação eficazes pode minimizar a interrupção causada por incidentes de segurança.

**Formas de fazer backup:**

**a)** Robôs de fitas

**b)** Discos rígidos

**c)** Arranjos de discos

**d)** Discos removíveis

**e)** Em nuvem

**f)** Fitas magnéticas manuais (cada vez menos utilizadas)

Segundo (Lima, 2021), a maioria dos leitores da nova geração acredita que fazer backup em fitas ou em alguns desses modelos acima mencionados não supera o backup em nuvem, por ser mais fácil e prático. E isso seria verdade se não fosse o tipo mais caro e também demorado, dependendo da quantidade de dados e da limitação de banda na transmissão de dados das redes atuais. De qualquer forma, o backup em nuvem também utiliza outros métodos e o importante é, independente da forma, sempre manter uma cópia de segurança.

### 3.2.15 Governança de Dados

A governança de dados estabelece políticas e diretrizes claras para o gerenciamento de dados, incluindo segurança. Define responsabilidades, fluxos de

trabalho e processos para garantir que os dados sejam manuseados de maneira segura e em conformidade com as regulamentações.

O objetivo da segurança em aplicações é manter a confidencialidade, integridade e disponibilidade dos recursos de informação a fim de permitir que as operações de negócios sejam bem sucedidas. Esse objetivo é alcançado através da implementação de controles de segurança (OWASP, 2021).

**a) Políticas e Diretrizes Claras:** Desenvolver políticas e diretrizes claras que definam como os dados serão coletados, armazenados, processados e analisados. Essas políticas devem abordar questões como privacidade, segurança e conformidade regulatória.

**b) Responsabilidades Definidas:** Atribuir responsabilidades claras para garantir a conformidade com as políticas de governança de dados. Isso inclui designar proprietários de dados responsáveis pela integridade e qualidade dos dados em toda a organização.

**c) Catálogo de Dados:** Um catálogo de dados é uma ferramenta fundamental para organizar e documentar conjuntos de dados disponíveis em uma organização. Ele fornece metadados detalhados sobre cada conjunto de dados, incluindo sua origem, estrutura, significado e uso potencial. Isso facilita a descoberta, compreensão e uso eficiente dos dados por parte dos membros da equipe. Deve constar em um catálogo de dados:

**i) Metadados estruturais:** Fornecem informações sobre a organização e hierarquia dos dados, facilitando a visualização e navegação.

**ii) Metadados descritivos:** Descrevem o conteúdo, contexto e características físicas dos dados, auxiliando na busca e identificação.

**iii) Metadados de preservação:** Documentam o processo de preservação dos dados, incluindo informações sobre gerenciamento de direitos e ações de preservação.

**iv) Metadados administrativos:** Oferecem informações sobre governança, acesso, segurança e aspectos técnicos dos dados, auxiliando no gerenciamento e controle.

**v) Metadados de proveniência:** Indicam a origem e o histórico dos dados, permitindo rastrear seu ciclo de vida e consultar diferentes versões.

**vi) Metadados de definição:** Fornecem um vocabulário comum para entender o significado dos dados, incluindo definições, regras e lógica utilizada.

Atualmente, há uma variedade de opções de ferramentas para catalogar dados, incluindo soluções de código aberto e plataformas pagas. A partir desses repositórios, é possível estabelecer controle e aplicar políticas da empresa, sendo a catalogação de dados considerada essencial para a organização em uma boa governança.

**d) Compreensão e Conscientização:** É essencial que todos os funcionários compreendam as políticas e diretrizes de governança de dados e reconheçam a importância de proteger as informações. A conscientização e a educação contínuas são fundamentais para promover uma cultura de segurança de dados.

**e) Controles de Acesso e Segurança:** Implementar controles de acesso adequados para proteger os dados contra acesso não autorizado. Isso pode incluir autenticação de usuários, autorizações baseadas em funções e criptografia de dados sensíveis.

**f) Monitoramento e Auditoria:** Estabelecer mecanismos de monitoramento e auditoria para acompanhar o uso e o acesso aos dados. Isso ajuda a detectar atividades suspeitas e garantir a conformidade com as políticas de segurança de dados.

Auditoria e Monitoramento são processos fundamentais para garantir a segurança e integridade dos sistemas de Big Data. A Auditoria envolve a

análise detalhada das atividades e transações realizadas no ambiente, identificando potenciais violações de segurança ou comportamentos suspeitos. Por outro lado, o Monitoramento consiste na observação contínua das operações em tempo real, visando detectar e responder rapidamente a eventos de segurança. Juntos, esses processos proporcionam uma visão abrangente do ambiente de Big Data, ajudando a mitigar riscos e garantir a conformidade com as políticas de segurança estabelecidas. Algumas ferramentas sugeridas para auxiliar nesse processo são:

**i) Apache Sentry:** Pode desempenhar um papel importante na monitorização e gestão de erros. Ao integrar o Sentry com sistemas de Big Data, como Apache Hadoop e Apache Hive, as equipes podem monitorar e diagnosticar problemas de forma eficaz em ambientes distribuídos e de grande escala. Isso ajuda a garantir a integridade e o desempenho do sistema, permitindo uma detecção rápida e resolução de problemas para manter as operações de Big Data funcionando sem problemas (CLOUDERA, 2021).

**ii) Apache Ranger:** Desempenha um papel crucial no cenário de Big Data, oferecendo recursos robustos de controle de acesso e auditoria. Ele permite que os administradores definam políticas de segurança detalhadas para proteger os dados e recursos sensíveis em ambientes de Big Data, como o Apache Hadoop e o Apache Hive. Além disso, o Apache Ranger oferece capacidades avançadas de auditoria, permitindo que as organizações rastreiem e monitorem todas as atividades realizadas nos sistemas de Big Data (CLOUDERA, 2021).

**iii) Zabbix**: É uma plataforma de monitoramento de código aberto amplamente utilizada em ambientes de Big Data. Ele oferece recursos avançados de monitoramento e gerenciamento de infraestrutura. No contexto de Big Data, o Zabbix pode ser usado para monitorar servidores, clusters de dados, serviços e aplicativos, garantindo que tudo esteja funcionando conforme o esperado (ZABBIX, 2021).

**g) Gestão da Qualidade dos Dados:** Implementar processos e procedimentos para garantir a qualidade e a integridade dos dados. Isso inclui a validação de dados, a padronização de formatos e a correção de erros ou inconsistências.

**h) Conformidade Regulatória:** Garantir que as políticas de governança de dados estejam em conformidade com as regulamentações e leis de proteção de dados aplicáveis. Isso pode incluir regulamentações e leis de privacidade de dados locais.

**i) Avaliação e Melhoria Contínua:** Realizar avaliações regulares da eficácia das políticas e práticas de governança de dados e identificar áreas de melhoria. A governança de dados é um processo contínuo que requer adaptação às mudanças nas necessidades e regulamentações da organização.

### 3.2.16 Versionamento

O versionamento de código desempenha um papel crucial na garantia da integridade, rastreabilidade e colaboração no desenvolvimento e manutenção de pipelines de dados, algoritmos de processamento e análise, e outras soluções de software relacionadas ao tratamento de dados.

Principais vantagens do versionamento de código:

**a) Rastreabilidade:** Acompanha mudanças no código ao longo do tempo, facilitando a investigação de problemas e auditorias.

**b) Reprodutibilidade:** Permite reproduzir análises e processamentos anteriores, mantendo a consistência dos resultados.

**c) Colaboração:** Facilita o trabalho conjunto em código, sem conflitos entre membros da equipe.

**d) Controle de Versões:** Gerenciar diferentes versões do software, simplificando implantações e correções.

**e) Padronização:** Estabelece práticas para garantir consistência e qualidade no código produzido.

**f) Automação:** Integra o versionamento com pipelines de CI/CD para automatizar processos, aumentando eficiência e reduzindo erros.

Ao implementar uma governança de dados em Big Data, o versionamento de código deve ser considerado para garantir a qualidade, confiabilidade e colaboração no desenvolvimento e manutenção de soluções de software relacionadas ao tratamento de dados.

O GitHub e/ou GitLab são as principais ferramentas de versionamento de código adotadas pelo mercado e são totalmente integráveis a qualquer sistema de validação de acesso, VPN, autenticação e outras validações previstas nas políticas. No caso do GitHub, é necessário ter um cuidado adicional e sempre versionar os códigos no modo privado para evitar vazamentos de informações.

### 3.2.17 Hierarquias de papéis

A tecnologia de controle de acesso baseado em funções, conhecida como Role-Based Access Control (RBAC) em inglês, é um sistema que determina permissões de acesso com base nos papéis que os usuários desempenham em uma organização. As hierarquias de papéis nos permitem garantir que um gerente sempre terá acesso aos mesmos registros de seus subordinados. Se um funcionário desejar assumir o papel de outro gerente, ambos devem conversar para autorizar esse acesso. Isso garante um sistema de controle e segurança adequado.

### 3.2.18 Testes

Os testes são essenciais para validar os produtos desenvolvidos. Recomenda-se realizar testes durante o desenvolvimento e antes de realizar o deployment do código em ambiente produtivo. Algumas ferramentas são sugeridas para cada etapa de teste:

**a) Chaos Monkeys:** É utilizado para inserir falhas de forma deliberada no código. Seu uso no cenário de dados se justifica para validar o tratamento de dados e caso não estejam em conformidade, observar como o código se comporta.

**b) Biblioteca PaperMill:** É utilizada para validar as políticas de conformidade e segurança do código. Caso esteja dentro do previsto, é liberado para deployment. Caso não esteja em conformidade, o avanço da implantação do projeto é barrado.

### 3.2.19 Deployment/Produção

Recomenda-se adotar a arquitetura de microsserviços para projetos de Big Data, pois permite a divisão da estrutura em partes menores (MENDES, MARTINS, 2019). Isso possibilita que cada projeto ocupe uma parte independente, garantindo que, caso ocorra um erro em um projeto, os demais continuem operando normalmente. Uma proposta de estrutura de deployment é:

**a) Passo 1:** Unir os arquivos do projeto em um repositório GIT.

**b) Passo 2:** Realizar testes de validação nesse repositório com o Papermill.

**c) Passo 3:** Enviar os arquivos após validados para a estrutura do microsserviço.

Essa sequência é segura e oferece total controle sobre todas as etapas, sendo aplicável tanto em ambientes locais quanto na nuvem.

### 3.2.20 Cursos e atualizações

Expor os funcionários a treinamentos, trocas de ideias e experiências em relação à segurança é crucial para o desenvolvimento da maturidade corporativa. As políticas muitas vezes são percebidas como complexas e demoradas para serem compreendidas. Portanto, criar oportunidades em diversos cenários para os funcionários entenderem a importância desse processo tende a torná-lo mais fácil e

aceitável. Isso se torna ainda mais importante, uma vez que precisará ser repetido conforme as leis e regulamentações são atualizadas. Abaixo estão algumas sugestões e formas de disseminar as políticas dentro do ambiente corporativo:

**a)** Cursos/Treinamentos/Gameficação

**b)** Eventos Online ou Presencial

**c)** WorkShop com profissionais de mercado

**d)** Troca de conhecimento entre empresas do mesmo Segmento

**e)** Abordagem multidisciplinar envolvendo especialistas em Dados e Segurança.

**f)** LGPD - acionar a ANPD para conformidade e novas práticas a serem adotadas

Neste tópico, foi abordado um conjunto abrangente de boas práticas destinadas a otimizar a segurança dos ambientes de Big Data.

# 4 Conclusão

A segurança em Big Data é uma questão complexa em função da extensão da fronteira de ataque dessa categoria de sistema. Os desafios principais identificados, juntamente com as ameaças mais comuns, destacam a complexidade inerente à proteção desses ambientes. Por meio de uma lista abrangente de 101 boas práticas, este estudo oferece diretrizes fundamentais para a implementação de medidas eficazes de segurança em ambientes de Big Data, visando garantir a integridade, confidencialidade e disponibilidade dos dados.

No entanto, a implementação de boas práticas de segurança em Big Data enfrenta diversos desafios. Cada ambiente tem suas características próprias e as sugestões aqui apresentadas precisam ser cuidadosamente adaptadas a cada circunstância. Além disso, as boas práticas devem ser constantemente atualizadas para acompanhar as novas ameaças e vulnerabilidades que surgem no cenário digital.

As recomendações supõem a utilização de uma série de ferramentas e técnicas que precisam ser sempre atualizadas e reavaliadas, já que a segurança da informação se assemelha a uma corrida armamentista, onde cada lado procura sempre se adaptar e tirar vantagens dos pontos fracos do lado rival.

Nem todos os aspectos foram aqui abordados. Tópicos como segurança física, seleção de pessoal e ameaças a ambientes em nuvem exigem considerações especiais que exigem outras monografias específicas.

A adoção de boas práticas aumenta a confiança dos clientes, uma vez que eles se sentem mais seguros ao compartilhar suas informações com empresas que demonstram preocupação com a segurança.

A falta de segurança pode levar a prejuízos financeiros para as empresas, uma vez que vazamentos de informações podem resultar em multas e processos judiciais, além de perda de credibilidade

É importante ressaltar a importância da conscientização e treinamento dos profissionais envolvidos com Big Data. A adoção efetiva das boas práticas de segurança depende do conhecimento e da capacidade dos profissionais em implementar as medidas adequadas. Portanto, é necessário investir em programas

de treinamento e conscientização para garantir que todos os envolvidos estejam cientes da importância da segurança em Big Data e sejam capazes de implementar as medidas necessárias.

A adoção de boas práticas de segurança dos dados permite às empresas cumprirem regulamentações legais relacionadas à privacidade e proteção das informações.

Em relação a tendências futuras pode-se destacar o aumento do uso de técnicas avançadas de análise preditiva para identificar padrões suspeitos e prevenir ataques antes que ocorram. Além disso, espera-se um maior investimento em soluções de segurança baseadas em inteligência artificial e aprendizado de máquina, que são capazes de identificar ameaças complexas e se adaptar a novos tipos de ataques.

# 5 Referências Bibliográficas

SANTOS, CARLOS EDUARDO LESSA e CARVALHO, FELIPE FREIRE DE - PRIVACIDADE E PROTEÇÃO DE DADOS NA ERA DO BIG DATA

ALVIM, L. Perfil e competências do profissional da Informação para a gestão de dados em massa (Big data). 2018. Disponível em: <https://dspace.uevora.pt/rdpc/handle/10174/23392>. Acesso em: out. 2023.

LIMA, ADRIANO CARLOS DE. SEGURANÇA DE DADOS E BIG DATA, 2021. p. 98-110.

AMARO, D. P.; DROZDA, F. O. Aplicação de Big Data em inovações para a logística e gestão da cadeia de suprimentos: uma revisão sistemática da literatura. Revista Produção Industrial & Serviços, [S.l.], v. 1, n. 1, p. 1-10, 2019. Disponível em: <https://periodicos.uem.br/ojs/index.php/rev_prod/article/view/52424>. Acesso em: 10 out. 2022.

CABRERA-SÁNCHEZ, J. P. Fatores que afetam a adoção de análises de big data em empresas. Revista de Administração, 2020, v. 55, n. 3, p. 327-338. Disponível em: <https://www.scielo.br/j/rae/a/KxjtdnnkmCDL4FZs3q6psFQ/?lang=pt>. Acesso em: 08 Jul. 2023.

CARVALHO VICTORINO, M. de; SHIESSL, M.; OLIVEIRA, E. C. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais conectados. Informação & Sociedade: Estudos, v. 27, n. 2, p. 1-15, 2017. Disponível em: <https://www.brapci.inf.br/_repositorio/2017/05/pdf_bab240464a_0000023131.pdf>. Acesso em: 08 Jul. 2023.

CARVALHO, A. P. Proposta de um framework de compliance à Lei Geral de Proteção a Dados Pessoais (LGPD): um estudo de caso para prevenção a fraude no contexto de Big Data. 2021. Disponível em: <http://www.rlbea.unb.br/jspui/handle/10482/42510>. Acesso em: 20 Ago. 2023.

CAUMO, R. B. Indicadores socioeconômicos produzidos a partir de Big Data: um framework para avaliação da qualidade estatística aplicado ao turismo. 2021. Disponível em: <https://repositorio.ufsc.br/handle/123456789/227144>. Acesso em: 15 Set. 2022.

CAUMO, R. B.; SOUZA, J. A. A Qualidade de indicadores socioeconômicos produzidos a partir de Big Data. Repositório IPEA, 2021. Disponível em: <https://repositorio.ipea.gov.br/handle/11058/11757>. Acesso em: 15 Set. 2022.

COSTA, EAP da. Organização e Processamento de Dados em Big Data Warehouses baseados em Hive. 2017. Disponível em: <https://search.proquest.com/openview/cd55e3e4a25df66b0d1d062a0bdbc044/1?pq-origsite=gscholar&cbl=2026366&diss=y>. Acesso em: 05 Jun. 2022.

COSTA, MIP. Etiquetagem e rastreio de fontes de dados num Big Data Warehouse. 2019. Disponível em: <https://repositorium.sdum.uminho.pt/handle/1822/70190>. Acesso em: 05 Jun. 2022.

DUTRA, M. L.; MACEDO, D. D. J. Curadoria digital: proposta de um modelo para curadoria digital em ambientes big data baseado numa abordagem semi-automática para a seleção de objetos. Informação & Informação, [S.l.], v. 21, n. 2, p. 1-19, 2016. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27176>. Acesso em: 05 Jun. 2022.

FRANCO, JVM. Como o uso do big data pode influenciar na tomada de decisão. 2020. Disponível em: <http://ric.cps.sp.gov.br/handle/123456789/10426>. Acesso em: 05 Jun. 2022.
GOMES, RDP. Big Data: desafios à tutela da pessoa humana na sociedade da informação. 2017. Disponível em: <https://www.bdtd.uerj.br:8443/handle/1/9803>. Acesso em: 05 Jun. 2022.
LAIGNER, R. N. Desenvolvimento de sistemas big data: um mapeamento sistemático da literatura. 2017. Disponível em: <https://app.uff.br/riuff/handle/1/7607>. Acesso em: 25 Nov. 2022.
LIMA, A. C. Segurança de dados e Big Data. 2021. Disponível em: <https://books.google.com/books?hl=en&lr=&id=V_JQEAAAQBAJ&oi=fnd&pg=PT5&dq=Boas+Pr%C3%A1ticas+em+Seguran%C3%A7a+em+Big+Data+na+Engenharia+de+Dados+e+Big+Data&ots=PkN0EhSLaG&sig=JsZZH-SEISID_EyOIIB0XC5nAZs>. Acesso em: 05 Mar. 2023.
LIMA, C. M. O Big Data e a Ciência de Dados na produção bibliográfica brasileira da Biblioteconomia e da Ciência da Informação. 2022. Disponível em: <https://app.uff.br/riuff/handle/1/26746>. Acesso em: 05 Mar. 2023.
LIMA, F. L. G. V. Big Data Warehousing em tempo real: da recolha ao processamento de dados. 2017. Disponível em: <https://repositorium.sdum.uminho.pt/handle/1822/53679>. Acesso em: 05 Mar. 2023.
LINHARES, MVD. Uso de big data e criação de tecnologia (software e hardware), com prova de conceito e validação, para identificar, diagnosticar e prever os fatores de riscos no segurança alimentar; • Desenvolver tecnologia (software e hardware), com funcionalidade de Big Data… de falhas na execução das boas práticas apícolas em todo o processo produtivo. 2017. Disponível em: <https://repositorio.ufba.br/handle/ri/24108>. Acesso em: 2021.
SANTOS, CARVALHO. PRIVACIDADE E PROTEÇÃO DE DADOS NA ERA DO BIG DATA. 2019. Disponível em: <https://app.uff.br/riuff/bitstream/handle/1/13054/Carlos%20Eduardo_Felipe%20Freire.pdf?sequence=1&isAllowed=y>. Acesso em: 2022.
LINS, BFE. Big Data e Gestão. Relatório Técnico da Academia Brasileira da ..., 2021. Disponível em: <http://www.belins.eng.br/ac01/papers/ABQualidade_estudos_21_08.pdf>. Acesso em: 05 Mar. 2023.
LUZ, C. M. Análise da transição da tecnologia mediante as novas regras de privacidade e segurança de dados em empresas de Big Data no mercado de comunicação. Repositório Módulo, 2018. Disponível em: <https://repositorio.modulo.edu.br/jspui/handle/123456789/1922>. Acesso em: 05 Mar. 2023.
MACHADO, FNR. Big data o futuro dos dados e aplicações. 2018. Disponível em: <https://books.google.com/books?hl=en&lr=&id=2LdiDwAAQBAJ&oi=fnd&pg=PT5&dq=Boas+Pr%C3%A1ticas+em+Seguran%C3%A7a+em+Big+Data+na+Engenharia+de+Dados+e+Big+Data&ots=-5hJzgchM8&sig=FUZez_0mpAGbJAksMSIIRqQ_opc>. Acesso em: 11 Out. 2023.
MARQUESONE, R. Big Data: Técnicas e tecnologias para extração de valor dos dados. 2016. Disponível em: <https://books.google.com/books?hl=en&lr=&id=cbWlDQAAQBAJ&oi=fnd&pg=PT3&dq=Boas+Pr%C3%A1ticas+em+Seguran%C3%A7a+em+Big+Data+na+Engenharia+

de+Dados+e+Big+Data&ots=6nXko9LcC6&sig=ykCBmXbRR8Cmr7Z8z2YxTQzQSxM>. Acesso em: 11 Out. 2023.
MENDES, RJM. Interoperabilidade e normalização de plataforma de serviços de big data em ambientes federados. 2022. Dissertação/Projeto de Engenharia Informática (DPEI), Universidade de Lisboa. Disponível em: <https://repositorio.ul.pt/handle/10451/55595>. Acesso em: 11 Out. 2023.
O livro "Hadoop Security" de Ben Spivey e Joey Echeverria é um guia prático que fornece informações detalhadas sobre a segurança do Hadoop.
OLIVEIRA, D. F. Proteção de privacidade de dados em ambiente de big data analytics: um estudo da realidade brasileira. Disponível em: <https://www.teses.usp.br/teses/disponiveis/100/100131/tde-23012023-202448/en.php>. Acesso em: 23 jan. 2023.
OLIVEIRA, D. F.; PRADO, E. P. V. Causas e problemas de privacidade de dados em sistemas de big data analytics: uma revisão sistemática da literatura. 2022. Disponível em: <https://aisel.aisnet.org/amcis2022/lacais/lacais/2/>. Acesso em: 11 Out. 2023.
OLIVEIRA, D. F.; PRADO, E. P. V. Soluções para Privacidade de Dados em Sistemas de Big Data Analytics: Uma Aplicação da Técnica Delphi. 2023. Disponível em: <https://aisel.aisnet.org/amcis2023/lacais/lacais/10/>. Acesso em: 11 Out. 2023.
SANTOS, A.; BRANCO, S.; TEFFÉ, C. Privacidade em perspectivas. 2016. Disponível em: <https://itsrio.org/wp-content/uploads/2017/03/Andreia-Santos-V-revisado.pdf>. Acesso em: 11 Out. 2023.
SANTOS, I. S.; OLIVEIRA, P. A. M.; OLIVEIRA, V. T. et al. Big Data Fortaleza: Plataforma Inteligente para Políticas Públicas Baseadas em Evidências. Anais do XI Workshop ..., 2023. Disponível em: <https://sol.sbc.org.br/index.php/wcge/article/view/24877>. Acesso em: 11 Out. 2023.
SOUZA JÚNIOR, GL de; BARRETO, JMC; EPITAYA, E. Gestão de Saúde e Segurança do Trabalho na Era de Big Data. Disponível em: <https://portal.epitaya.com.br/index.php/ebooks/article/view/481>. Acesso em: 2022.
Spivey, Ben e Echeverria, Joey - Hadoop Security
TEIXEIRA, I. V. Inteligência artificial, big data e democracia: o caso Cambridge Analytica e a regulação de novas tecnologias no ordenamento jurídico brasileiro. Repositório PUC Goiás, 2023. Disponível em: <https://repositorio.pucgoias.edu.br/jspui/handle/123456789/6400>. Acesso em: 2023.
VAN LEIJDEN, E. M. L.; BUENO, C. F. S.; D'AMORIM, F. D. Iniciativas e desafios para prover um ambiente de compartilhamento e análise de dados corporativo: Big Data PE. Repositório ENAP, 2022. Disponível em: <https://repositorio.enap.gov.br/handle/1/7436>. Acesso em: 2023.
O que é o OWASP? Introdução às 10 principais vulnerabilidades e riscos do OWASP. 2023. Disponível em: <https://www.f5.com/pt_br/glossary/owasp> . Acesso em: 2024.
OWASP. O que é segurança de API. Disponível em: <https://owasp.org/www-project-api-security/>. Acesso em: 2024
MACVITTIE. O caso das estratégias integradas de segurança de aplicações e APIs. 2023. Disponível em: <https://www.f5.com/pt_br/company/blog/the-case-for-integrated-app-and-api-security-strategies>. Acesso em: 2024

SOUZA. DESENVOLVIMENTO SEGURO DE APLICAÇÕES WEB SEGUINDO A METODOLOGIA OWASP . Disponível em: <http://repositorio.ufla.br/jspui/bitstream/1/30670/1/MONOGRAFIA_Desenvolvimento_seguro_de_aplicacoes_web_seguindo_a_metodologia_owasp.pdf>. Acesso em: 2024

AVELINO. O que é OWASP. Disponível em: <https://www.dio.me/articles/o-que-e-owasp> . Acesso em: 2024

SEGINFO. Um guia para codificação segura do OWASP. Disponível em: <https://seginfo.com.br/2021/09/27/um-guia-para-codificacao-segura-do-owasp/>. Acesso em: 2024

OWASP. Melhores Práticas de Codificação Segura OWASP Guia de Referência Rápida. 2010. Disponível em: <https://owasp.org/www-pdf-archive/OWASP_SCP_Quick_Reference_PT-BR_v1.0.pdf >. Acesso em: 2024

Anonimização dos dados pessoais. Disponível em: <https://lgpd-brasil.info/capitulo_02/artigo_12>. Acesso em: 2024

CLOUDERA. PRODEMGE WORKSHOP. 2021. Disponível em: <https://ead-rh.prodemge.gov.br/pluginfile.php/24988/mod_resource/content/1/PRODEMGE_WORKSHOP.pptx.pdf>. Acesso em: 2024

Pallets Projects. Flask Documentation. Disponível em: <https://flask.palletsprojects.com/>. Acesso em: 2024

MENDES, MARTINS, 2019. Analisando o desempenho de microsserviços implementados através da decomposição parcial de sistemas monolíticos. Disponível em: <https://app.uff.br/riuff/bitstream/handle/1/8522/TCC%20-%20Carlos%20e%20Luiz%20-%20Microservic%CC%A7os.pdf?sequence=1&isAllowed=y>. Acesso em: 2024

Zabbix. Desenvolvido por Zabbix LLC. Versão 5.4. 2021. Disponível em: <https://www.zabbix.com/documentation/current/start>. Acesso em: 2024