REINALDO GOMES DA SILVA

SEGURANÇA E PRIVACIDADE DE DADOS NO USO DE BIG DATA

São Paulo
2016

REINALDO GOMES DA SILVA

SEGURANÇA E PRIVACIDADE DE DADOS NO USO DE BIG DATA

Monografia apresentada ao PECE – Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para conclusão do curso de MBA em Tecnologia de Software.

São Paulo
2016

Catalogação-na-publicação

SILVA, REINALDO GOMES DA
SEGURANÇA E PRIVACIDADE DE DADOS NO USO DE BIG DATA / R. G. D. SILVA -- São Paulo, 2016.
69 p.

Monografia (MBA em Tecnologia de Software) - Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1.BIG DATA 2.SEGURANÇA 3.PRIVACIDADE 4.DADOS 5.ENGENHARIA DE SOFTWARE I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia II.t.

REINALDO GOMES DA SILVA

SEGURANÇA E PRIVACIDADE DE DADOS NO USO DE BIG DATA

Monografia apresentada ao PECE – Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para a conclusão do curso de MBA em Tecnologia de Software.

Área de Concentração: Tecnologia de Software

Orientador: Prof. Dr. Jorge Rady de Almeida Junior

São Paulo
2016

Dedico este trabalho a Deus,
pelo dom da vida.

AGRADECIMENTOS

À Universidade de São Paulo que durante esses 16 anos de vínculo tem oferecido oportunidades de crescimento intelectual.

À Escola Politécnica da Universidade de São Paulo que desde a graduação tem proporcionado espaço para pesquisa em seus laboratórios e bibliotecas.

Ao PECE – Programa de Educação Continuada em Engenharia que com muita competência soube organizar e oferecer um curso de qualidade.

A Bibliotecária Ana Maria Badiali pelo profissionalismo exemplar e revisão técnica.

Aos meus pais, Abel e Maria, irmãos Marcelo e Lucas, e família pela compreensão e incentivo, para a realização e concretização do curso.

A minha noiva Renata, pelo incentivo, apoio na arte final e revisão.

A meu orientador Prof. Dr. Jorge Rady pelos ensinamentos, paciência e compreensão.

"Educação não é o quanto você tem guardado na memória, nem o quanto você sabe. É ser capaz de diferenciar entre o que você sabe o que você não sabe. É saber aonde ir para encontrar o que você precisa saber; e é saber como usar a informação que você recebe."

(William Feather)

RESUMO

Neste trabalho é realizado um estudo de diversos mecanismos e técnicas que visam impedir o uso indevido e não autorizado dos sistemas de armazenamento em Big Data, assegurando a preservação da privacidade e da segurança de dados, além de proporcionar uma melhora no gerenciamento. Com o advento e popularização dos serviços de rede, e o rápido crescimento, expansão dos sistemas de banco de dados, vários sinais apontam para uma nova fase de captura, análise e utilização de informações. Em todo o mundo, muitas empresas e órgãos do governo já fazem uso crescente de tecnologias de manipulação do grande volume e variedade de dados que são produzidos constantemente. No entanto, essa imensa quantidade de dados também vem servindo de fonte para extração de informações privativas e sigilosas, tanto de empresas ou de indivíduos. Finalmente, são apresentadas soluções tecnológicas que têm como prioridade manter segurança no acesso aos dados para cada tipo de ocorrência sugerida nos sistemas de armazenamento, conforme as características que representam *Big Data*, tendo como prioridade mante-los protegidos e disponíveis.

Palavras-chave: *Big Data.* Criptografia. técnicas de segurança e privacidade. métricas. autenticação.

ABSTRACT

This work is a study of various mechanisms and techniques that aimns to prevent misuse and unauthorized storage systems in Big Data, ensuring the preservation of privacy and data security, in addition to providing an improved management. With the advent and popularization of network services and rapid growth, expansion of database systems, several signs point to a new capture phase, analysis and use of information. Worldwide, many companies and government agencies are already making increasing use of large volume handling technologies and variety of data that are produced constantly. However, this vast amount of data is also serving as a source for extraction of private and confidential information, whether companies or individuals. Finally, technological solutions are presented that have a priority to maintain secure access to data for each type of occurrence suggested the storage systems according to the characteristics that represent Big Data, with priority to keep them protected and available.

Keywords: Big Data. encryption. security and privacy techniques. metrics. authentication.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3DES .....*Triple Data Encryption Standard*
AES.....*Advanced Encryption Standard*
API.....*Application Program Interface*
BDWG.....*Big Data Working Group*
CSA.....*Cloud Security Alliance*
DNS.....*Domain Name System*
DoS.....*Denial of Service*
DW.....*Data Warehouse*
GPS.....*Global Positioning System*
GSI.....*Grid Security Infrastructure*
HADOOP.....*Highly Archived Distributed Object Oriented Programming*
HDFS.....Hadoop *Distributed File System*
HTTP.....*Hypertext Transfer Protocol*
IaaS.....*Infrastructure as a Service*
IDS.....*Intrusion detection System*
IoT.....*Internet of Things*
IP.....*Internet Protocol*
iRODS.....*Integrated Rule-Oriented Data System*
KP-ABE.....*Key-policy Attribute-based Encryption*
NIST.....*National Institute of Standards and Technology*
NoSQL.....*Not only* SQL
NSA.....*National Security Agency*
NTFS.....*New Technology File System*
PaaS.....*Platform as a Service*
PLN.....Processamento de Linguagem Natural
SaaS.....*Software as a service*
SGBD.....Sistema Gerenciador de Bases de Dados
SIEM.....*Security Information and Event Management*
SNS.....*Security Management System*
SQL.....*Structured Query Language*
TI.....Tecnologia da informação
XML.....*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Há alguns anos, o *World Economic Forum 2013* – Fórum Econômico Mundial lançou um projeto chamado *Rethinking Personal Data* (Kearney, 2014), que identificava possibilidades de crescimento econômico e seus benefícios sociais. Como parte dessa iniciativa, o Fórum definiu dados pessoais, no sentido informativo, como um novo ativo econômico.

Com o advento da Internet e o rápido crescimento e expansão dos serviços de rede, vários sinais apontam para uma nova fase de captura, armazenamento, e reutilização de dados pessoais, pois a popularização no acesso à rede tornou-se facilitada, diminuiu-se o custo de acesso, e com isso, o rastreamento de informações, tanto de governo, como de empresas passou a ser um procedimento de grande utilidade. Isso porque a difusão das redes e das telecomunicações facilita a criação de novos modelos de negócio. Como exemplo, é muito comum obter informações de hábitos de compra de clientes ou dados demográficos de surgimento de doenças. Os dados não estruturados, como e-mails, fotos e vídeos, possuem informações valiosas para uma organização.

Economias desenvolvidas e emergentes já fazem o uso crescente de tecnologias intensivas de processamento de dados. Estatísticas apontam para um número cada vez maior de assinaturas de uso de telefonia móvel com acesso de dados via Internet, aumentando ainda mais a quantidade de tráfego, produção de conteúdo, ampliação de fontes de dados e consequentemente, um aumento da informação.

A Tecnologia da Informação já faz uso das ferramentas de análise de grandes conjuntos de dados para melhorar a segurança da informação, como exemplo, na verificação de operações financeiras, arquivos de *logs* (históricos de registros de sistemas), em tráfego de rede para identificar anomalias e atividades suspeitas, correlacionando múltiplas fontes de informação. Com isso, existe a melhora e o aperfeiçoamento de desempenho com novas visualizações e resultados coerentes de estratégica de negócios.

Porém, o aumento da produção de dados, por si só, produz preocupações mais amplas quanto à segurança e privacidade, uma vez que as organizações sejam públicas ou privadas, estão acumulando e analisando imensos volumes de dados, que antes eram apenas fragmentados, desconexos ou não organizados.

Através do rastreamento e captura de dados na *Internet* o anonimato deixa de existir, pois é cada vez maior a capacidade de se associar uma identidade da vida real do indivíduo com seus hábitos, fazendo uma mudança no quesito privacidade, desfazendo a distância entre o público e o privado.

Este acúmulo na produção de dados e o procedimento tecnológico de extração e armazenamento formam o chamado *Big Data.*

Sobre os aspectos éticos associados ao uso de *Big Data*, (Davis, 2012) cita quatro elementos relacionados com a discussão na captura e uso das informações:

- Identidade do proprietário dos dados: refere-se à relação existente entre a identidade real e a identidade virtual;
- Privacidade: é possível fazer correlações sobre indivíduos gerando novos dados;
- Propriedade dos dados: este elemento faz refletir sobre uma série de questões relacionadas à posse dos dados;
- Reputação individual: como são percebidos e avaliados pelas informações pós-análise, da qual está relacionada essencialmente com o que se conhece do indivíduo e os seus pensamentos.

A privacidade também implica nas aderências a aspectos regulatórios e legislativos. Além do próprio setor financeiro, muitos outros setores, como os da saúde, já se preocupam com o sigilo de informações, como por exemplo, os cuidados na preservação de registros eletrônicos de pacientes. Abrir e integrar dados, antes fechados para a sociedade, gera preocupações e receios, principalmente devido às características específicas de determinadas informações, inclusive na administração pública. O gerenciamento da segurança e da privacidade de dados é efetivamente um desafio social, mas também técnico, e tais aspectos podem ser tratados em conjunto.

A segurança também requer uma inteligência analítica de prevenção. Em vez de se limitar a reagir à violação de segurança, a TI precisa ser capaz de prever onde as ameaças à segurança surgirão e quem será responsável por elas. É necessário identificar ou prever possíveis indícios de tentativas de invasão a um sistema, ou obter dados estatísticos das áreas mais vulneráveis no passado. As empresas que conseguem se adiantar as futuras ameaças de segurança e impedi-las inevitavelmente terão mais sucesso do que as que se limitam em resolver problemas quando eles apenas se evidenciam.

A preservação do anonimato da informação, tendo como alvo de pesquisa a **segurança** dos dados, possui o mesmo foco e importância da garantia da **privacidade** e a confidencialidade, com objetivo primordial da preservação do anonimato, dentre os outros já citados. Por isso, essa correlação entre esses dois termos, conforme a literatura pesquisada é tratada conjuntamente neste trabalho – **segurança e privacidade** em *Big Data.*

1.1 Objetivo

Este trabalho tem como objetivo analisar diversos mecanismos e técnicas contra a invasão da privacidade dos dados (acesso confidencial) e da segurança (impedimento do uso indevido e não autorizado) de informações pessoais armazenadas digitalmente. São propostas soluções tecnológicas que auxiliam em projeto de segurança em *Big Data*, onde se deseja como prioridade manter os dados seguros e disponíveis, além de garantir que estão sendo utilizados de forma adequada por usuários autorizados, e com nível de privilégio pré-determinado. Os mecanismos possibilitam antecipar eventuais tentativas de violação de segurança dos sistemas e privacidade de dados, analisando registros de eventos, comunicações por correio eletrônico ou comentários em mídias sociais, utilizando mecanismos de análises de dados.

1.2 Motivações

Existe um consenso de que o desafio do uso de *Big Data* nos próximos anos será a questão da segurança e privacidade dos dados. Exemplo disso é o risco de uma imensa quantidade de dados confidenciais de órgãos públicos e privados serem extraviados e divulgados e/ou reutilizados para fins ilícitos.

O desafio está na aplicação de mecanismos mais adequados na prevenção de ataques, assim como na utilização de algoritmos cada vez mais eficientes na preservação do anonimato da informação, bem como o desenvolvimento de protocolos mais rígidos

A conscientização dos profissionais de TI sobre a importância da segurança e privacidade não é suficiente. Tem sido cada vez mais comum a análise de dados de informações vindas exclusivamente de terminais de consulta. Com isso, novas técnicas para garantir o sigilo dos dados, possivelmente utilizando técnicas de criptografia, serão cada vez mais incorporadas, como por exemplo, em transações financeiras ou em pesquisas científicas.

A maior disponibilidade de recursos de TI nos servidores virtuais de arquivos (computação da nuvem ou *cloud computing*) é mais um motivo de importância, já que os fornecedores de serviços em nuvem precisam documentar e apresentar confiabilidade além da preocupação em evitar vazamento de dados sigilosos, seja por descuido, ou por invasões propositais.

Como o volume de dados, a velocidade de processamento, a complexidade do tipo de dados e as especificações de segurança e privacidade continuam a crescer além das expectativas, as empresas são forçadas a buscar novos recursos, mecanismos e maneiras de atender a necessidades operacionais, comerciais e jurídicas.

1.3 Justificativas

Há muitos trabalhos sobre segurança e privacidade produzidos na comunidade científica. Segundo (Saraladevi *et al.*, 2015), praticamente todos os dispositivos de TI (computadores, equipamentos de rede, dispositivos de armazenamento) descartam dados de alguma maneira. Com isso, deixam-se rastros digitais a todo o momento, seja utilizando *Internet Banking* para compras *on-line*, acessando um buscador eletrônico, usando um *smartphone*, ativando serviços de localização por GPS, etc. Comentários de fóruns de discussão e/ou de redes sociais também podem oferecer diversas informações de comportamento e consumo, auxiliando empresas e governo na tomada de decisões com objetivos estratégicos nas áreas de políticas governamentais, seleção de investimentos, gestão de empresas e negócios, dentre outros.

Segundo (Adluru *et al.*, 2015) e (Davenport, 2014), uma grande quantidade de dados vem servindo de fonte para extração de informações dos usuários, sendo utilizada e analisada por governos e empresas de tecnologia em redes sociais, como a *Google e Facebook,* que conseguem armazenar grandes volumes através dos serviços de *Cloud Computing*, e com isso, podem correlacionar informações que permitam obter uma visão comportamental abrangente das pessoas, seus hábitos e costumes. A contrapartida negativa na utilização destes dados é o uso indevido, ou não autorizado, de informações privativas e sigilosas.

Entretanto, dados coletados também podem ser analisados para um estudo de **previsão** de risco de segurança, e/ou na **prevenção** de ataques, isso quando realizado com ferramentas adequadas.

Existem outras linhas de pesquisa que tratam de política de uso das informações considerando os aspectos de privacidade e segurança de dados ou na investigação de questões éticas e/ou legais. Outros trabalhos também sugerem uma necessidade de se criar padrões de acesso, e regras bem definidas de privacidade e segurança de acesso, relacionadas ao *Big Data*.

1.4 Estrutura do Trabalho

O Capítulo 1 apresenta as motivações, objetivo, justificativas, metodologia utilizada e a estrutura do trabalho.

O Capítulo 2 apresenta o conceito de *Big Data*, exemplos em números estatísticos e as diferenças com bancos de dados tradicionais. São explicadas as características de *Big Data* (Volume, Variedade, Velocidade, etc.), os componentes representativos e a classificação tecnológica, quanto à maneira que são tratados os dados. Também é feita uma breve explicação de NoSQL, e dada uma maior ênfase na explicação do conceito de Hadoop.

O Capítulo 3 apresenta oito mecanismos em segurança e privacidade (Segurança em Hadoop, Segurança *em Cloud Computing*, Monitoramento e Auditoria, Gerenciamento de Chaves, Mascaramento, Segurança em *DataGrid* e Gerenciamento de Incidentes), mostrando seus funcionamentos e aplicações.

O Capítulo 4 descreve os atributos e a utilização dos mecanismos em segurança e privacidade, conforme a característica, ocorrência e possível solução no uso.

Finalmente, o Capítulo 5 descreve recomendações mais significativas na utilização de bancos de dados não relacionais, *DataGrid*, *Cloud Computing* e em processamentos em massa com Hadoop e as melhores práticas em segurança e privacidade em *Big Data*.

1.5 Metodologia

Para a realização deste trabalho foram realizadas pesquisas bibliográficas nas bases de dados científicas, tais como do *Institute of Electrical and Electronics Engineers* (IEEE), da *Springer Science* e *Cloud Security Alliance* (CSA), e outras fontes que constam nas referências. Além das *tags* (segurança, privacidade, Big Data), os termos, palavras chaves e assuntos utilizados nas consultas às bases foram sobre os mecanismos e técnicas na prevenção de ataques e invasão a banco de dados, utilização de algoritmos de criptografia mais eficientes na preservação do anonimato da informação e aplicações de protocolos mais rígidos e que exigem maiores restrições de acesso aos sistemas.

2 REVISÃO BIBLIOGRÁFICA

Numa análise preliminar, o assunto Big Data, no que se refere ao armazenamento e uso de dados é também um objeto de estudo para a Ciência da Informação, tornando-se necessário uma apresentação conceitual neste trabalho.

2.1 Conceito de Big Data

Big Data pode ser conceituado como uma quantidade de dados suficientemente diferenciada, que leve a uma mudança nas formas mais tradicionais de análise. Segundo (Davenport, 2014), *Big Data* é um termo genérico para a manipulação de dados que não podem ser contidos nos repositórios tradicionais, cujas características de tamanho destacam-se por ser volumosos demais para serem armazenados em um único servidor; não estruturados para se adequar a um banco de dados organizado em linhas e colunas; o fluxo de informação é demasiadamente dinâmico e muito fluído para ser armazenado em um *Data Warehouse* (DW) estático.

Historicamente, o conceito *Big Data* surgiu em 2001, através de um relatório de pesquisa do analista *Doug Laney* da *META Group* (atualmente chamado *Gartner Group*), onde sua definição é feita da seguinte forma: "*Big Data* é [um conjunto de dados] volumoso, de alta velocidade e/ou de variedade; são ativos de informação que exigem formas rentáveis e inovadoras de processamento para permitir uma aprimorada tomada de decisão, a descoberta de conhecimento e a otimização de processos". (Gartner, 2016)

Uma melhor definição sugerida seria um repositório ou acervo de bases de dados tão heterogêneo e volumoso que torna complexa sua utilização se fizer uso de ferramentas tradicionais de processamento e/ou nas aplicações mais tradicionais de gerenciamento.

A análise e estudo de *Big Data* podem revelar padrões, tendências ou associações, especialmente em relação ao comportamento humano e suas interações.

Num ambiente de *Big Data* é quase impossível processar, capturar, armazenar, filtrar, compartilhar, analisar e visualizar dados sem tratamento fazendo uso de tecnologias de bancos de dados relacionais e/ou na utilização de interfaces comuns

de visualização, pois nesse cenário exige-se *software* com poder de processamento quantitativo, distribuído e em paralelo, sendo executado em dezenas, centenas ou milhares de servidores. Na Tabela 1 são apresentas as diferenças entre os Sistemas Gerenciadores de Bancos de Dados (SGBD) tradicionais e *Big Data.*

Tabela 1 Diferença entre sistemas de bancos de dados tradicionais e *Big Data*

Parâmetro comparativo	*Big Data*	Sistemas Tradicionais
Consulta	Ausência de SQL	Presença de SQL
Arquitetura	Distribuída	Centralizado
Tipos de Dados	Estruturados, semi-estruturados e não estruturados.	Estruturados
Modelo de dados	Ausência de esquema	Presença de esquema
Troca de Dados	Desconhecida ou complexa	Troca conhecida
Volume dos Dados	Petabytes ou Exabytes	Terabytes
Tráfego dos Dados	Maior	Menor
Integridade dos Dados	Menor	Maior

Fonte: (Saraladevi *et al.*, 2015)

A Tabela 1 menciona o termo esquema ou *Schema que* se refere a um mecanismo de manipulação de dados, que consiste na aplicação de um procedimento ou plano que determina a maneira como os dados são retirados do armazenamento. A sua ausência em *Big Data* é devido ao constante fluxo de dados, tornando o tráfego mais fluído, se comparado aos bancos relacionais. Também é mencionada a linguagem de pesquisa padrão para base de dados relacional SQL (*Structured Query Language)* ou Linguagem de Consulta Estruturada. Faz-se necessário exemplificar os tipos dos dados que são mencionados no decorrer do trabalho: estruturados (comércio eletrônico, mercado financeiro, setor da saúde); semi-estruturados (*web logs*, *e-mails*, documentos) e não estruturados (imagens, vídeo, áudios, *web sites*).

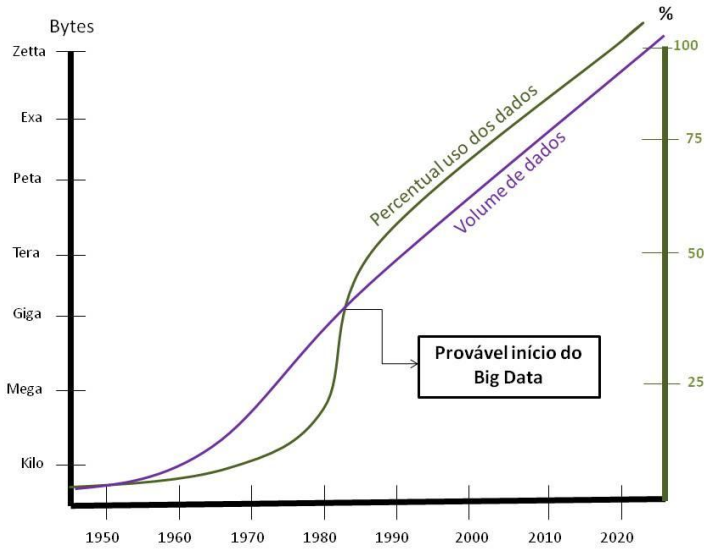
Segundo (Vijayakumar *et al.*, 2015), a cada dois anos praticamente dobra-se a quantidade de dados produzidos pela economia mundial. Todos os dias são gerados aproximadamente 2,5 exabytes de dados (Ibm, 2015), e existe uma previsão de que esse volume irá aumentar 300 vezes, de 130 exabytes em 2005 para 40.000 exabytes em 2020. Pesquisas realizadas em 2013 mostram que 90% dos dados no

mundo foram produzidos nos dois anos anteriores (Taurion, 2013) e, como consequência dessa revolução tecnológica, *Big Data* está se tornando cada vez mais uma fonte de preocupação nas ciências, governos e empresas.

Segundo a IDC *International Data Corporation (Chau et al., 2015)*, existem mais de um bilhão de pessoas por mês usando um telefone móvel para transferir informações, e constantemente esses dados estão sendo armazenados e monitorados por empresas e organismos de telecomunicações. Com isso, as análises de *Big Data* hoje permitem identificar rapidamente riscos de segurança, devido ao aumentando das capacidades de análises preditivas, ou como oportunidades de negócios, ao se extraírem idéias, criando novas formas de valor aos dados.

Segundo *Diego Kuonen*, professor de Ciência de Dados da Universidade de Genebra, conforme o decorrer dos anos, o percentual de dados utilizados pela sociedade continuará tão elevado, quanto a quantidade de dados produzidos, conforme Figura 1.

Figura 1 - Esboço gráfico representativo da utilização de *Big Data*



Fonte: (Irizarry *et al.*, 2014), (Kuonen, 2014)

Este crescimento do *Big Data* deve-se também ao constante aumento na quantidade dos usuários de dados. Na década de 1980 com o advento da popularização dos microcomputadores pessoais, documentos em papel passaram a ser digitalizados e informatizados. Em seguida, na década de 1990, com a democratização da Internet, os dados passaram a ser armazenados nos servidores, e compartilhados na rede. E consequentemente, as inúmeras possibilidades e recursos dessa informatização favoreceram para o crescimento no uso e acesso dos usuários aos meios eletrônicos.

2.2 Características do *Big Data*

Em 2001, *Doug Laney* (Taurion, 2013) propôs um modelo tridimensional composto pelas iniciais em 3vs, como sendo um novo desafio mundial de crescimento de dados, destacando um aumento do Volume, Variedade e Velocidade, descritos a seguir:

a) Volume: Esta característica descreve a **quantidade** de dados. O volume de dados cresce, em parte, porque estão sendo produzidas cada vez mais informações através de dispositivos móveis, registros de software, equipamentos de vigilância, leitores de identificação de radiofrequência, redes de sensores sem fio, etc., aumentando-se exponencialmente dia após dia.

b) Variedade: Esta dimensão descreve os diferentes **formatos** de dados armazenados no *Big Data*, como textos de redes sociais, metadados de imagens, áudios, vídeos, documentos de texto, dados matemáticos, e-mails, dados de sensores, etc, alem dos diferentes setores de negócios que fazem intercâmbio de informações, como os de área de saúde e ciências da vida, centros de pesquisas, centros geológicos, centros astronômicos e meteorológicos, redes sociais, agências de notícias, e assim por diante.

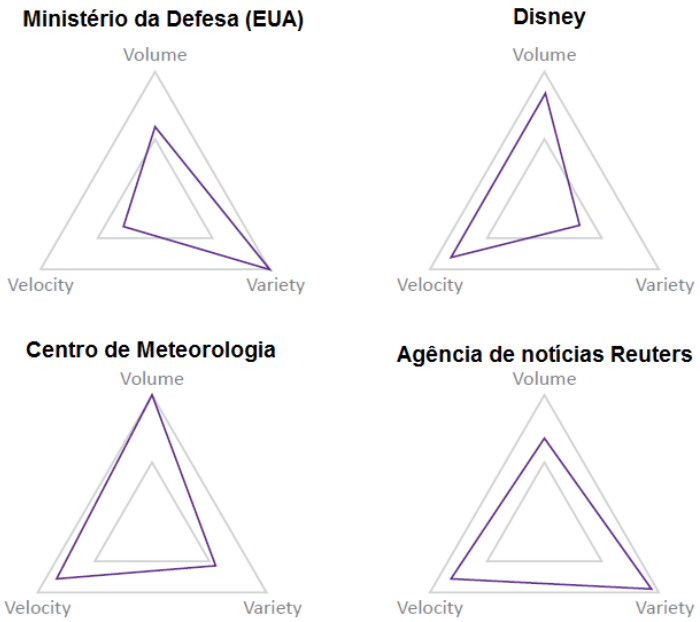
A Variedade é considerada atualmente uma das características mais críticas para a segurança e privacidade em *Big Data*, devido à grande **diversidade** das fontes, tipos e formatos dos dados, deixando mais complexo o tratamento da informação.

c) Velocidade: Esta dimensão descreve a **velocidade da produção** de dados. Com os avanços tecnológicos, têm-se cada vez mais redes com maiores largura

de banda e, todos os dias, usuários de internet em todo o mundo estão produzindo dados.

A Figura 2 mostra uma comparação relativa das características dominantes de dados armazenadas por várias empresas, públicas e privadas. A dimensão Volume do Ministério da Defesa Americano pode até ser considerada como alta, porém quando comparada a outras organizações, não se tem tanta relevância, quanto a sua Variedade, e assim por diante.

Figura 2 – Forma ilustrativa das dimensões Volume, Velocidade e Variedade.



Fonte: (Network Technical Authority, 2013)

As características de Variabilidade e Veracidade surgiram posteriormente devido aos desafios do processamento dos dados relacionados às áreas das ciências da vida, com necessidade de se combinar diferentes tipos fontes de dados, conforme descritas a seguir:

a) Variabilidade: (Adluru *et al.*, 2015) esta dimensão descreve a **inconsistência** no fluxo de dados em um determinado momento, isto é, a irregularidade da quantidade de dados produzidos num determinado momento. Por vezes, a produção de dados pode ser elevada e outras vezes muito baixa.

b) Veracidade: Esta dimensão descreve a **qualidade** dos dados armazenados no ambiente de *Big Data*. Os dados produzidos globalmente têm qualidades diferentes. A tomada de decisões por usuários de *Big Data* depende diretamente a autenticidade da informação.

Pesquisadores (Terzi *et al.*, 2015) e empresas (IBM, 2015) também fazem do uso de outras dimensões para descrever importantes características em *Big Data*, como Volatilidade, Validade, Visualização, Valor e Complexidade, descritas a seguir:

a) Volatilidade: esta dimensão descreve o **período** em que um determinado dado é relevante e por quanto tempo esses dados deveram ser armazenados.

b) Validade: esta dimensão descreve a **exatidão** dos dados, ou seja, o armazenamento incorreto de dados, levando-se ao desperdício de espaço. Assim, esta característica é utilizada para descrever um maior rigor ao se tratar com grandes volumes de dados.

c) Visualização: esta dimensão descreve a **apresentação** dos dados. É a maneira de representar os dados de forma legível e acessível, através de *Visual Analytics* contendo dezenas de variáveis e parâmetros da informação.

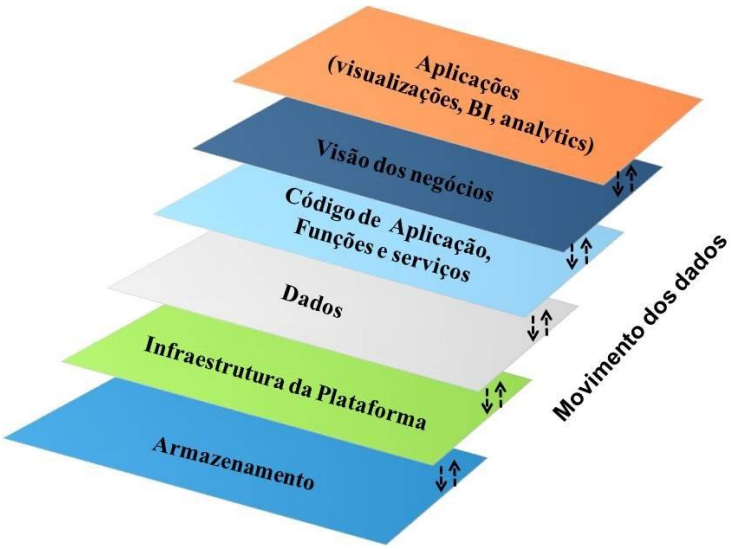
d) Valor: esta dimensão descreve o **custo** dos dados. Esta característica trata da análise para implantação de modelos de negócio e tomadas de decisão. Em *Big Data*, dados e análises são totalmente independentes, porém ao combiná-los, pode-se produzir um poderoso mecanismo de fonte informações úteis e praticamente ilimitados.

e) Complexidade: esta dimensão descreve a **dificuldade** no gerenciamento de *Big Data* devido à grande quantidade de dados a serem armazenados e analisados.

2.3 Componentes representativos de camadas do *Big Data*

Assim como acontece com outras tendências tecnológicas, são apresentados na Figura 3 uma ilustração representativa, típica de *Big Data*. Esses degraus podem ser considerados em camadas, trabalham em conjunto, e se auto ajustam, em armazenamento e em processamento especializado de alto desempenho.

Figura 3 - Representação de um típico *Stack* de camadas em *Big Data*



Fonte: (Davenport, 2014)

A seguir é feita uma breve descrição dos termos mencionados nas camadas da Figura 3:

Armazenamento: grandes e diversificadas quantidades de dados em *cluster* de discos em ambientes Hadoop (Programação Orientada a Objetos Distribuídos Extremamente Arquivados).

Infraestrutura da Plataforma: é o conjunto de funções que levam a alto desempenho do processamento de *Big Data*, utilizando o Hadoop como plataforma, pois possui a capacidade de integrar e administrar dados, além da própria aplicação de um sofisticado processamento computacional.

Dados: essa camada representa a informações e metadados, com suas respectivas propriedades. Metadados são, resumidamente, informações que descrevem os próprios dados, ou seja, servem para identificar, localizar, compreender e gerencia-los.

Código de aplicação, funções e serviços: nesta camada, o código de aplicação é utilizado para manipular e processar os dados. O Hadoop utiliza um *framework* de processamento, chamado *MapReduce*, não somente para distribuir os dados entre os discos, mas também para aplicar instruções computacionais complexas a esses dados.

Visão dos negócios*:* dependendo do modelo de negócio, é aplicado um processamento adicional na construção de uma estrutura de dados para análises ou consultas com ferramenta baseada em SQL, por meio do *MapReduce.*

Aplicações (*Visual Analytics, Business Intelligence* (BI) e *Data analytics*): nessa camada, os resultados do processamento do *Big Data* são analisados e visualizados para posteriores tomadas de decisão de negócio.

2.4 Classificações em *Big Data*

Big Data pode ser classificada tecnologicamente por diferentes **categorias** para uma melhor compreensão das suas características (3Vs) e seus atributos. A importância da classificação deve-se à alta demanda de dados na **nuvem** e suas particularidades. Na Tabela 2 segue uma classificação tecnológica, quanto ao formato dos dados, fonte, finalidade, tipo de análise, tipo e método de processamento, e armazenamento.

Cada categoria tem suas próprias características e complexidades, paradigmas ou atributos, conforme descrito na Tabela 2.

Tabela 2 - Classificação tecnológica de *Big Data*

Formato	Fonte de dados	Consumo dos dados	Uso dos dados	Frequência
Estruturado	Redes Sociais	Humana	Industrial	*On-demand* (*Batch*)
Semiestruturado	*Internet of Things*	Processo de negócio	Acadêmico	
Não estruturado	Intranet Corporativas	Aplicações corporativas	Governamental	tempo-real
	Provedores de Serviços	Repositórios de dados	Centro de Pesquisas	

Tipo de Análise	Tipo de Processamento	Método de Processamento	Armazenamento	Teste de dados
Interativo	Preditivo	Alto Desempenho Computacional	relacional e não relacional	Normalização
tempo-real	Analítico			Modificação
	Modelamento	*Distribuído*	*Baseado em Grafos*	Manutenção
	Reporting	*Paralelo*	*chave-valor*	
		Cluster	*Orientado a colunas*	
		Grid	*Orientado a mídias*	

Fonte: adaptado de (Terzi *et al.*, 2015) e (Hashem *et al.*, 2015)

Alguns termos que merecem comentários da Tabela 2:

a) As fontes de dados incluem as redes sociais e a Internet das Coisas, onde os dados se alteram entre estruturados para altamente não estruturados, sendo armazenados em vários formatos. Em consequência da ampla variedade de fontes de dados capturados nas redes, a diferenciação fica a critérios de atributos como tamanho, redundância, consistência, etc.

b) Sobre os tópicos frequências e análises em *Big Data*, estes podem ser divididos em dois grupos: processamento em lote (*batch*) onde são focadas as análises nos dados já armazenados e o processamento em tempo-real (*stream*), com foco nas análises sobre dados em movimento. (Murthy *et al.*, 2014)

A Tabela 3 apresenta um resumo dos tipos de tecnologias e ferramentas de processamento e análise de dados em *Big Data*.

Tabela 3 - Tipos de ferramentas de aplicações em *Big Data*

Hadoop	*Framework* operacional de código aberto para o processamento de *Big Data* numa série de servidores, acessando e armazenando dados de forma paralela.
MapReduce	*Framework* arquitetônico de código aberto no qual o Hadoop se baseia - oferece um modelo de processamento, da qual, consultas são divididas e distribuídas entre os nós da rede.
Linguagem de *Script*	Linguagens de programação interativas, adequadas ao *Big Data* (exemplo, *Python, Pig, Hive*)
***Cloud Computing* (computação em nuvem)**	Virtualização de sistemas e arquivos, utilizando memória, processamento e armazenamento em outros sistemas remotos de hospedagem.
DataGrid	Recurso para processamento distribuído e em paralelo, fazendo uso do *MapReduce*.
Aprendizado de máquina	*Software* para identificar rapidamente um modelo mais adequado ao conjunto de dados.
***Visual analytics* (*Data Analytics*)**	Apresentação dos resultados analíticos em formatos visuais ou gráficos.
Processamento de Linguagem Natural (PLN)	*Software* para análise de texto - frequência, sentido etc.
In-memory analytics	Recursos para processamento de *Big Data* na memória do computador para obter maior velocidade (*Spark*)

Fonte: (Davenport, 2014)

Os termos que aparecem na Tabela 3 e que merecem uma descrição são:

Processamento de linguagem natural (PLN) é uma ramificação da área de inteligência artificial, que estuda os problemas da produção e compreensão automática de línguas humanas naturais. Convertem informação e ocorrências de bancos de dados de computadores em linguagem compreensível ao ser humano.

Visual *analytics* é a ciência do raciocínio analítica apoiada por interfaces visuais interativas. Atualmente, o armazenamento de dados tem aumentado num ritmo mais

rápido do que a capacidade de analisá-lo. Os *programas de Data Analytics* (análise de dados) utilizam uma abordagem orientada a dados.

2.5 NoSQL

Num ambiente de *Big Data*, torna-se complexo fazer operações consideradas simples como ordenação, sumarização ou remoção de dados de forma eficiente utilizando Sistemas Gerenciadores de Bases de Dados (SGBD) tradicionais (Brito, 2010).

Esses SGBD são uma categoria de banco de dados do tipo relacional, ou seja, possuem uma estrutura fixa e bem definida. Porém, aplicativos que fazem uso da Web 2.0, chamada de colaborativa, como as redes sociais, requerem um gerenciamento de grandes volumes de dados não estruturados, produzidos por milhões de usuários diariamente, além da busca do compartilhamento de informações, conhecimentos e interesses diversos.

Neste contexto, surgiram soluções de bancos de dados NoSQL (Bernadette, 2011), atendendo uma alta demanda de gerenciamento de dados não estruturados, e contribuindo com as necessidades do domínio de *Big Data,* como:

a) Execução de consultas e tratamento com baixa latência;
b) Escalabilidade elástica de manipulação, armazenamento, replicação e distribuição dos dados;
c) Fornecimento de mecanismos de inserção de novos dados de forma incremental e eficiente;
d) Necessidade de persistência (armazenamento não volátil) dos dados em aplicações de *Cloud Computing*;

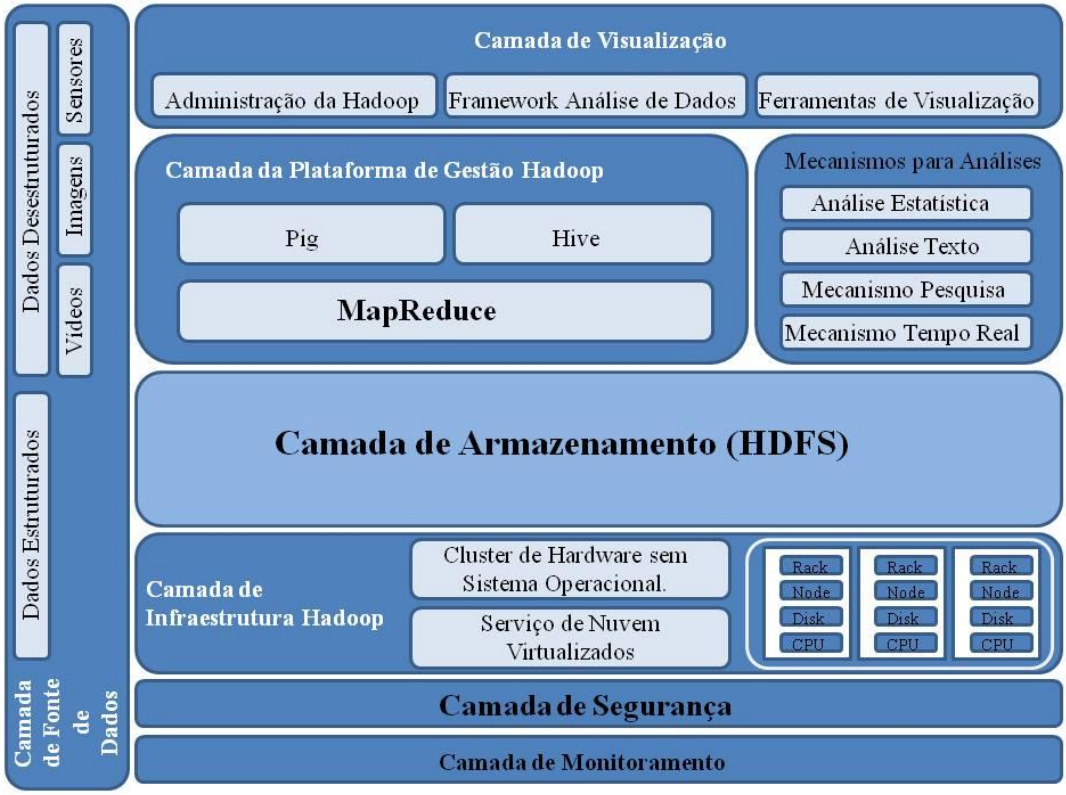
Com isso, houve um crescimento na adoção e difusão de tecnologias NoSQL nos mais diversos domínios de aplicação no contexto de *Big Data (J Han, 2011)*. No conjunto de produtos e ferramentas mais representativos de NoSQL está a plataforma Hadoop, juntamente com a implementação do algoritmo e paradigma *MapReduce*.

2.6 Hadoop

A plataforma Hadoop é um *framework* voltado para computação distribuída em *clusters* e processamento em *Big Data (Saraladevi et al., 2015)*. Desenvolvido na linguagem de programação *Java,* apresenta um alto grau de tolerância a falhas e grande escalabilidade de aplicações em diferentes cenários. Pode lidar com grande quantidade de dados estruturados e não estruturados (imagens, vídeos, áudios, etc.) em qualquer formato, e armazenados em *cluster* Hadoop, apresentando uma baixa latência de funcionamento.

Na Figura 4 é apresentado um modelo simplificado da arquitetura Hadoop, destacando a camada de visualização, plataforma de gerenciamento, fonte de dados, infraestrutura, segurança, monitoramento e os mecanismos de análises.

Figura 4 - Arquitetura Hadoop



Fonte: Adaptado de (Saraladevi *et al.*, 2015)

Há alguns termos que aparecem na Figura 4 que merecem uma breve descrição:

HDFS – acrônimo de Hadoop *Distributed File System* (Sistema de Arquivos Distribuídos) e escalonável, escrito em linguagem *Java.*

Hive é uma implementação da Data *Warehouse* (DW) para Hadoop*,* servindo para a compactação, consulta e análise de dados.

Pig é uma linguagem de consulta para Hadoop e gerencia as camadas inferiores. Possuem semelhanças com a linguagem para bancos de dados relacionais SQL.

A camada de segurança, foco deste trabalho, permite a criptografia de arquivos por camadas e autenticação utilizando protocolos seguros como o *Kerberos*.

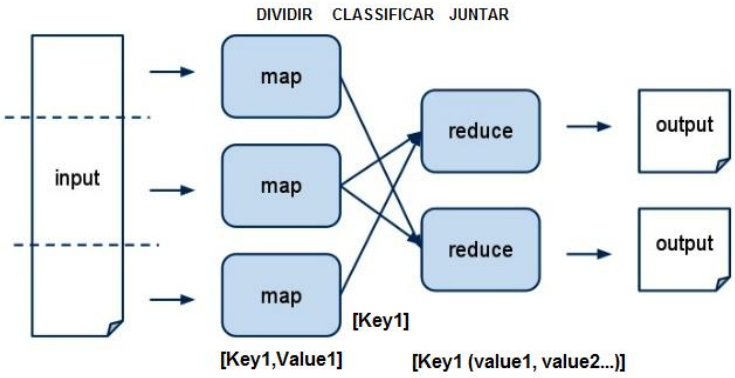
A camada de monitoramento deve ser capaz de lidar com uma distribuição de servidores e cluster, suportar diferentes sistemas operacionais, trabalhar com diferentes tipos de hardware, capaz de se comunicar com os protocolos de alto nível como o XML.

Originalmente, Hadoop não possuía recursos de prevenção de dados e segurança. Com sua popularização no mundo de pesquisa acadêmico, começaram a ser desenvolvidas muitas implementações. Conforme a Figura 5, seus principais componentes são:

***a) MapReduce*:** é um modelo/paradigma de programação paralela utilizado pelo Hadoop ou seja, um *framework* usado no processamento e análise de dados. Possui um conjunto de pares de entrada *key/value* (chave/valor)*,* produzindo na saída outro conjunto de pares *key/value*. O mecanismo de geração de pares de entrada e saída é dividido em três etapas: *Map* (mapear), *Shuffle* (embaralhar), e *Reduce* (reduzir). A etapa *Map* processa os dados de *input* (entrada/recebimento), a etapa Shuffle realoca dados com base na chave de saída, de tal modo que todos os dados pertencem a uma chave, e a etapa *Reduce* produz um *output* (saída ou a entrega).

Key/value (chave/valor) é um paradigma de armazenamento de dados, projetado para gerenciamento de matrizes associativas, e estrutura de dados (dicionário ou *hash*).

Figura 5 - Forma representativa do *MapReduce*



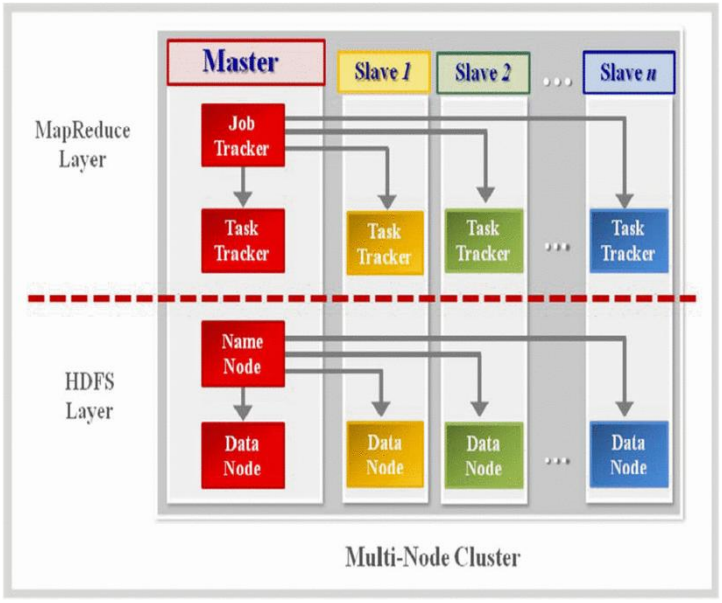
Fonte: (Adluru et al., 2015)

b) **HDFS -** Hadoop *Distributed File System*: sistema de armazenamento de dados Todos os arquivos são divididos e armazenados em blocos de tamanho fixo. O HDFS se encarrega de controlar a distribuição dos dados para cada máquina no *cluster* (*Datanode)*. HDFS também é responsável por adicionar e remover os *node* do *cluster* e na recuperação de *Datanode.* O *NameNode* atua como ponto central, pois é onde se encontram armazenados os metadados do sistema, além de ser responsável pela gestão dos blocos de *Datanodes*.

2.6.1 Funcionamento do Hadoop

Em um *cluster* de menor dimensão, a plataforma Hadoop possui um único *Master* e múltiplos *worker node* (*slaves*), conforme Figura 6. O *node master* principal contêm um *JobTracker*, *TaskTracker*, *NameNode* e *DataNode*. Um *slave* ou *worker node* atua igualmente ao *DataNode* e *TaskTracker*. Em um *cluster* de maior dimensão, o HDFS é gerenciado através de um servidor *NameNode* dedicado para alocação do sistema de índice de arquivos e um *NameNode* secundário, que pode criar um *snapshot* do *NameNode* (estruturas de memória), prevenindo o corrompimento do sistema de arquivos e reduzindo perda de dados.

Figura 6 - Funcionamento de *cluster* de uma arquitetura Hadoop



Fonte: (Jeong *et al.*, 2012)

Cada *node* em uma instância Hadoop normalmente tem um único *DataNode*. Um conjunto de *DataNodes* formam o *cluster HDFS* que tem um *NameNode* com a função de servidor de metadados. A *NameNode* executa várias atividades no *namespace* do sistema de arquivos e diretórios (abrir, fechar, renomear).

Um arquivo é dividido em mais de um bloco que é armazenado em um *DataNode* e o HDFS determina mapeamento entre o *DataNodes* e os blocos. Cada *DataNodes* executa operações de leitura e escrita da requisção do sistema de arquivo-cliente.

O *framework MapReduce* processa volumes de dados em paralelo em *clusters* de grandes dimensões (milhares de *nodes*) de forma escalonável, e confiável (tolerante a falhas).

Conforme Figura 6, na parte superior da representação encontra-se o *MapReduce* no sistema de arquivo, que consiste num *JobTracker*, onde as aplicações clientes enviam *Jobs MapReduce.* No cluster, o *JobTracker* envia material para o *node* do *TaskTracker*, mantendo o material mais próximo dos dados.

Com um sistema de arquivos do tipo *rack-aware*, o *JobTracker* conhece em quais *nodes* se localizam os dados, e quais as máquinas estão mais próximas. Se o

material não pode ser armazenado num *node* real, é dada a prioridade a um *node* no mesmo *rack*, reduzindo assim o tráfego de rede no backbone principal.

Em cada *node* o *TaskTracker* produz um processo em Máquina Virtual *Java* evitando, caso venha a falhar o *TaskTracker*, que o *Job* em execução cause um travamento no sistema.

Para verificação da situação atual do sistema, um sinal é enviado num intervalo de poucos minutos, do *TaskTracker* para o *JobTracker* e as informações podem ser visualizadas por um servidor HTTP, a partir de um navegador *web.*

A Figura 7 apresenta o funcionamento do sistema de armazenamento HDFS numa arquitetura Hadoop*.*

1) Uma máquina cliente carrega um arquivo para o *cluster;*

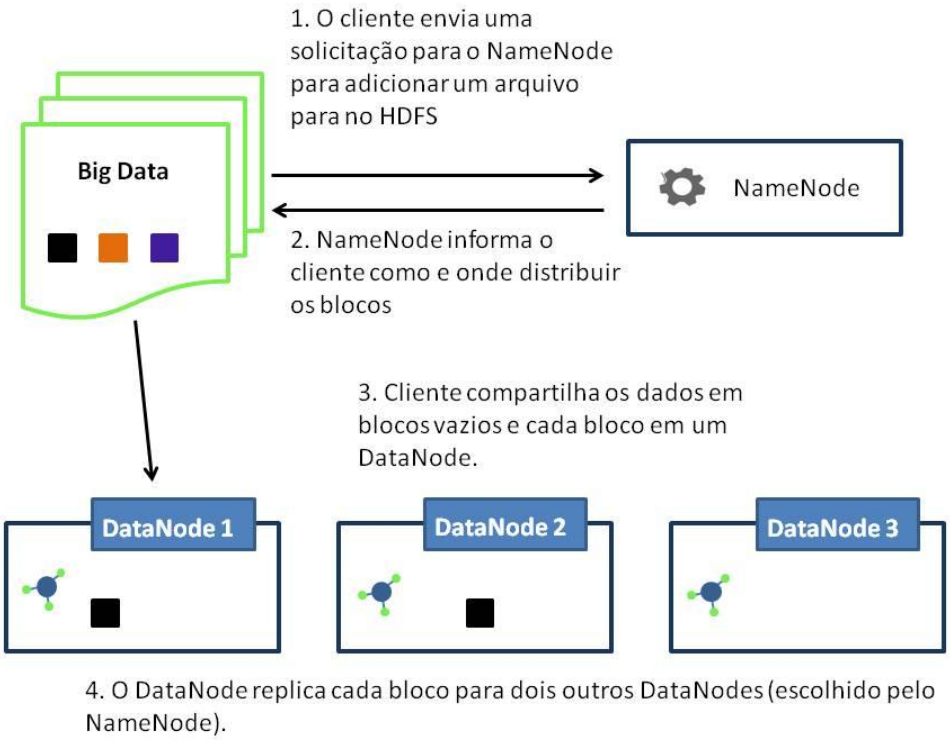
2) O *Datanode* quebra o arquivo em blocos fixos e armazena-os em todo o *cluster*;

3) O *BlockReport*, que contém a informação do bloco, decide qual *Datanode* deverá ser alocado;

4) Por padrão, são criadas três réplicas de *Datanode,* que são alocadas em setores diferentes;

5) O computador do cliente escreve num bloco de dados para um *Datanode* e automaticamente é feita uma numa replicação de blocos pelo sistema.

Figura 7 - Forma representativa do cluster de uma arquitetura Hadoop



Fonte: (Adluru *et al.*, 2015)

A vantagem do Hadoop esta na sua alta escalabilidade, além do custo-benefício. Ao comparar com o SGBD relacional, devido à grande quantidade de dados existentes, os SGBD mantêm de forma isolada os dados brutos *(raw*), enquanto o Hadoop considera de forma igual, os dados brutos e os tratados.

3 MECANISMOS E TÉCNICAS EM SEGURANÇA E PRIVACIDADE

Segundo (Terzi *et al.*, 2015) as soluções mais tradicionais de segurança e privacidade são insuficientes quando se trata de *Big Data*, pois algoritmos de criptografia e controles de permissões de acesso podem ser quebrados (Perlroth *et al.*, 2013), *firewalls* e mecanismos de segurança de rede na camada de transporte podem ser violados. Além disso, a origem dos dados muitas vezes é desconhecida e os dados considerados anônimos podem ser identificados. Por estes motivos, foram desenvolvidas técnicas para proteger, monitorar e auditar processos de *Big Data* em termos de infraestrutura, aplicativos e dados, descritos nas próximas seções.

São estudados neste trabalho os seguintes mecanismos e técnicas em segurança e privacidade de dados: Segurança em Hadoop, *Cloud Computing* e *DataGrid,* Monitoramento e Auditoria, Gerenciamento de Chaves e Incidentes e Mascaramento de rede. Estes foram separados e selecionadas neste trabalho pois o método de funcionamento envolve algoritmos de criptografia, autenticação e métricas.

Sobre criptografia, se num eventual tráfego de rede os dados ainda estiverem num formato de texto simples, a informação poderá ser furtada. Para evitar isso, faz-se necessário o uso dos algoritmos de criptografia com chaves públicas e privadas, simétricas e assimétricas, para restringir ou controlar o acesso de usuários a Sistemas Operacionais ou em máquinas virtualizadas, fazendo uso de **políticas de acesso**, ou também de uma autoridade certificadora de chaves.

3.1 Segurança em Hadoop

Para que a implantação de um sistema Hadoop garanta um nível aceitável de segurança e privacidade de dados na nuvem, segundo (Adluru *et al.*, 2015) são apresentadas técnicas que impedem, por exemplo, uma invasão num sistema para obter dados sigilosos ou privados:

a) É implementado um mecanismo de **autenticidade** entre o usuário e o *NameNode*, que é o componente do HDFS que produz *Datanode*. O usuário deve se autenticar para acessar o *NameNode*. Primeiramente, o usuário envia uma função *hash*, então

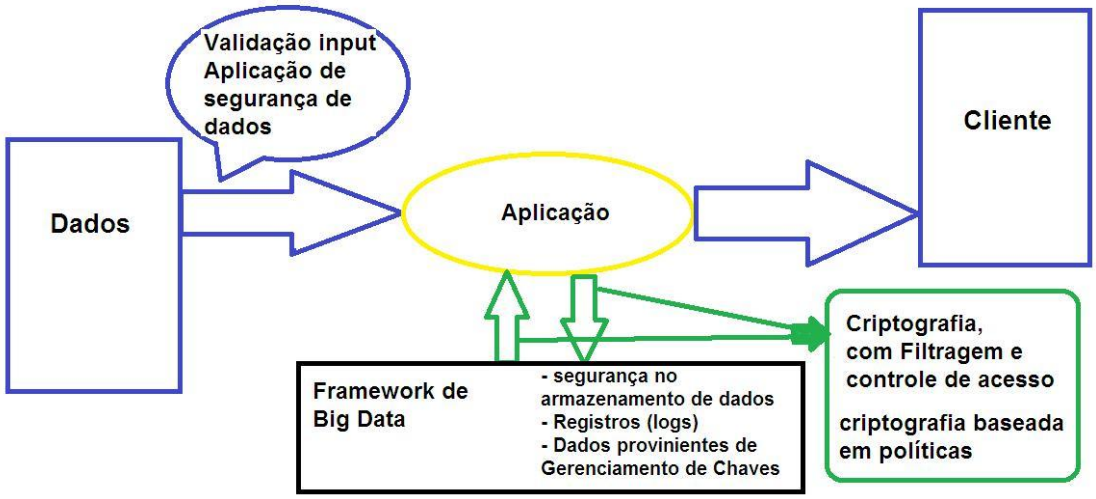
o *NameNode* cria função, e compara as duas. Se o resultado estiver correto, será autorizado o acesso ao sistema. Nesta etapa é utilizada uma técnica de autenticação *hashing* chamada de SHA-256 (*Secure Hash Algorithm)* - algoritmo de *hash* seguro, projetado pela NSA (Agência de Segurança Nacional Americana).

b) Também são utilizadas técnicas de criptografia assimétrica, tais como RSA, AES e RC6, com o objetivo de limitar todo o acesso de *hackers*. Nesta abordagem, o *MapReduce* é executado num processo de criptografia e seu inverso, conforme Figura 8.

c) A camada de HDFS também apresenta falhas de segurança. A fim de melhorar o problema de autenticação, o método utilizado é o protocolo *Kerberos* (Saraladevi *et al.*, 2015), cujo funcionamento consiste no serviço orientado a *ticket*. O segundo método está baseado no monitoramento de todas as informações utilizando algoritmo *Bullseye,* que gerencia as relações entre os dados originais e os replicados.

As duas técnicas foram testadas por (Adluru *et al.*, 2015) utilizando um fluxo de dados intencional como forma de indicativo de incidência de segurança.

Figura 8 - Representação do fluxo dos dados em Hadoop



Fonte:(Adluru *et al.*, 2015)

3.2 Segurança em *Cloud Computing* (Computação na nuvem)

O termo nuvem torna-se cada vez mais popular, numa referência ao uso de armazenamento de conteúdo digital e de virtualização de aplicações de Serviços(SaaS), Plataformas (PaaS) e Infraestruturas (IaaS) (Cheng *et al.*, 2015). A nuvem é um ambiente de recursos de hardware e software em *datacenters* que prestam diversos serviços de rede e/ou *Internet*. Segundo o NIST (Technology e NIST, September 2011), Instituto Nacional de Padrões e Tecnologia, *Cloud Computing*, ou Computação em Nuvem permite um *pool* compartilhado de recursos de computação ajustáveis (redes, servidores, armazenamento, aplicações e serviços) que podem ser rapidamente fornecidos e liberados com esforço mínimo de gerenciamento ou interação do provedor de serviços. Dentre inúmeras vantagens, destaca-se a sua rápida implantação de dimensionamento elástico de hospedagem, memória e processamento.

Existe uma correlaçao entre *Cloud Computing* e *DataGrid* , sendo que este último concentra e integra diversos recursos unificados de sistemas, fornecendo serviços de computação de alto desempenho. Além disso, *cloud computing* combina os vários recursos **distribuídos** de processamento e armazenamento de dados em larga escala, de forma dinâmica e elástica.

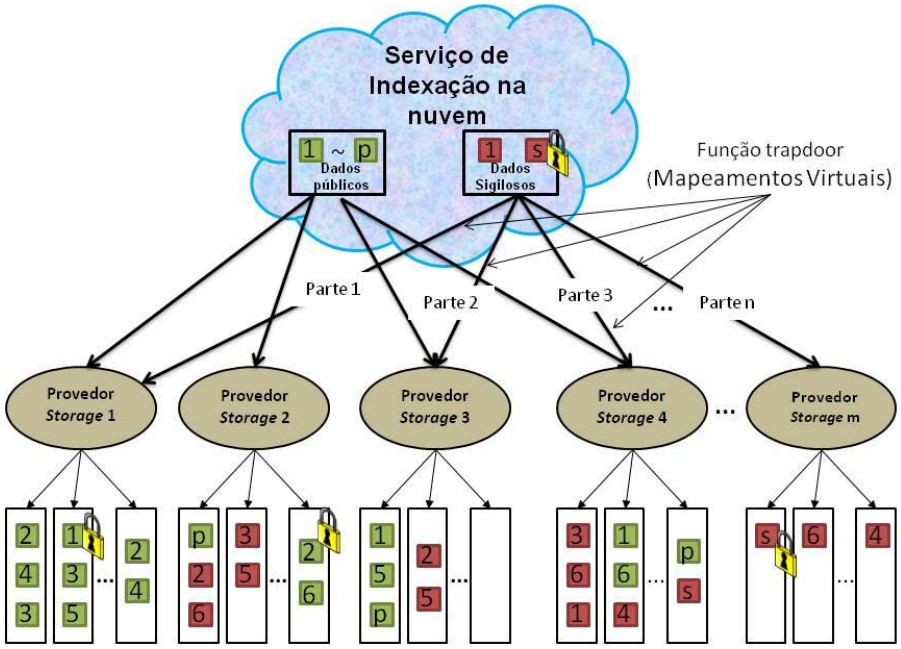
Um dos maiores desafios está no armazenamento de dados de forma segura e eficiente, pois não há uma garantia total de proteção, uma vez que a nuvem também utilizada para alocação de conteúdo malicioso e ameaças para novos ataques de rede, como o *Denial of Service* (DoS) - tentativas de tornar os recursos do sistema indisponíveis, *Spoofing* (mascaramento de pacotes IP utilizando endereços de remetentes falsificados), e ataques de *Brute Force* (tentativas de descoberta de senhas/*logins* através de processos manuais ou automatizados).

Outra preocupação dos usuários é em relação à confidencialidade e no sigilo de informações pessoais em provedores de serviços em nuvem, pois as soluções tradicionais de segurança de criptografia podem não ser tão eficazes ou nem adequadas para a proteção de informações em *Big Data*.

3.2.1 Segurança no procedimento de armazenagem na nuvem

A proposta desta técnica é dividir os dados em partes sequenciadas e armazená-las em vários discos na nuvem, aplicando um método de proteção no mapeamento virtual de fluxo de dados (Cheng *et al.*, 2015). É utilizado um método de criptografia em um mapeamento virtual para criar caminhos de dados para cada *Storage* de armazenamento, usando uma função chamada *trapdoor*. Como os dados estão separados em muitas partes sequenciadas e cada parte está localizada em diferentes *DataGrids*, ao invés de criptografar todos os dados do armazenamento em *Big Data*, devido ao grande volume, é criptografado apenas o mapeamento virtual. Para conseguir disponibilidade de dados, o sistema mantém várias cópias de cada parte e cada uma tem acesso ao índice. Assim, se qualquer parte dos dados for perdida, por alguma razão, é mantida com eficiência a disponibilidade das informações. Conforme Figura 9, os dados ficam armazenados em diferentes discos distribuídos de armazenamento. Esses discos pertencem a diferentes *DataGrids.*

Figura 9 - Representação do sistema de armazenamento em nuvem



Fonte:(Cheng *et al.*, 2015)

Segundo (Cheng *et al.*, 2015), análises comparativas e simulação demonstram que a solução proposta é eficiente e segura para *Big Data*, num ambiente de Computação na Nuvem.

3.2.2 Segurança nos dados armazenados na nuvem

Esta técnica oferece uma plataforma segura e confiável que permite que apenas os proprietários tenham acesso aos dados armazenados na nuvem, utilizando procedimentos de autenticação, criptografia e compressão, num ambiente de computação em nuvem.

Nesse método é realizada a forma tradicional com autenticação de senha, através da uma conta de e-mail para autorização de acesso. Então os dados são armazenados num formato criptografado e depois comprimidos. Por precaução, utilizam-se três servidores de backup de redundância. Se houver um interrompimento no funcionamento com algum destes servidores, os dados são descriptografados com uma chave privada.

Com objetivo de se obter um controle seguro de acesso, escalonável e eficiente, utilizam-se, de forma única e combinada, três tipos de criptografia: KP-ABE, PRE e recriptografia.

Durante o processo, é realizada uma divulgação de chaves de decodificação somente para usuários autorizados. Os usuários não autorizados, incluindo servidores de nuvem não conseguem decifrá-los, uma vez que eles não possuem as chaves de descriptografia.

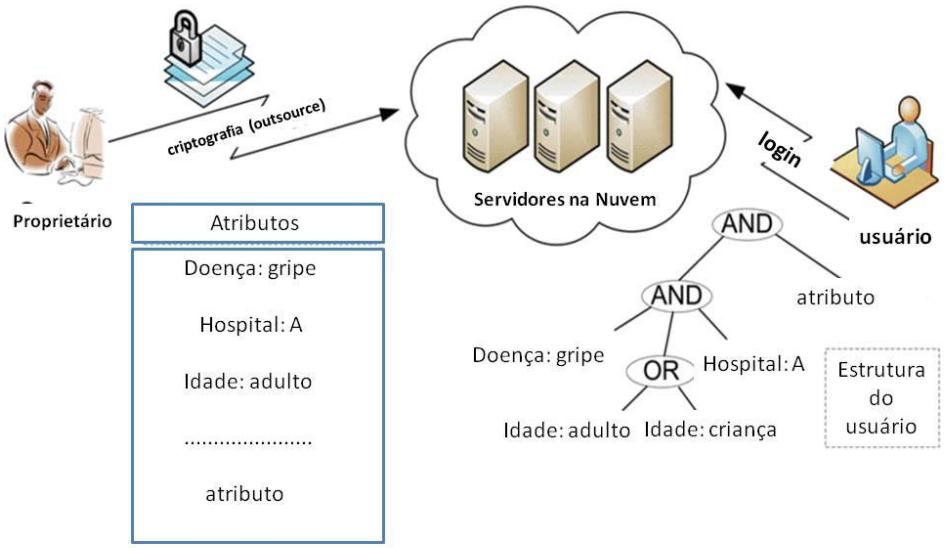
Especificamente, associa-se cada arquivo de dados a um conjunto de atributos, e a partir destes, define-se, a cada usuário, uma estrutura de acesso, sendo que cada um destes atributos possui um componente de chave pública; desta forma os arquivos são criptografados.

São definidas chaves privadas do usuário conforme a permissão de acesso de suas estruturas de acesso. O usuário apenas conseguirá decifrar um texto, se os atributos

de arquivo de dados satisfizerem a sua estrutura de acesso. Para impor esse tipo de controle de acesso, utiliza-se o tipo de criptografia KP-ABE para guardar as chaves de criptografia dos arquivos. Este cenário permite que o proprietário dos dados controle o acesso de seus arquivos com uma sobrecarga mínima em termos de esforço de computação e tempo *on-line* e, portanto, se torna favorável num ambiente de nuvem

Na Figura 10 é exemplificado um cenário de um centro hospitalar, onde o proprietário dos dados armazena milhões de registros médicos na nuvem. Isso permitiria que os usuários, tais como médicos, pacientes, pesquisadores tivessem acesso a vários conteúdos confidenciais de registros. Para isso, é necessária uma política de acesso, pois para os proprietários dos dados, se existe a vantagem de aproveitar os recursos que a nuvem fornece como eficiência e economia, por outro lado, deve-se preservar a confidencialidade do conteúdo

Figura 10 - Cenário de registros médicos hospedados na nuvem.



Fonte: (Yu *et al.*, 2010)

As principais vantagens dessa técnica são:

a) a complexidade da criptografia está relacionada apenas ao número de atributos associados ao arquivo de dados, e é independente do número de usuários do sistema;

b) as operações de criação de arquivo de dados e nova concessão ao usuário só afetam o atual sem envolver atualização do arquivo de dados de todo o sistema ou redigitação.

3.3 Monitoramento e Auditoria

A grande vantagem na utilização de tráfego de rede para monitoramento e auditoria deve-se à enorme quantidade disponível de fontes de informação (Marchal *et al.*, 2014). Isso porque com o crescente volume de dados, torna-se mais complexo fazer uma análise holística do tráfego. No entanto, para aplicações que requerem níveis aceitáveis de análise de segurança em tempo real, monitoramento e auditoria continuam sendo imprescindíveis para a prevenção, identificação e proteção de tentativas de **invasões** de rede.

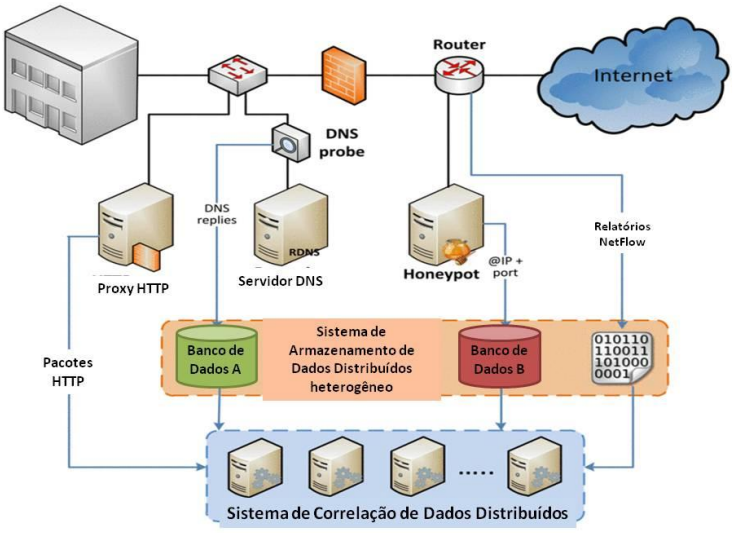
O monitoramento tem o objetivo extrair e analisar incidentes de tráfego e a auditoria executa **políticas** sistêmicas, que envolvem diversos métodos definidos no sistema.

3.3.1 Monitoramento de rede

Um Sistema de Identificação de Invasão (IDS) tem como objetivo fortalecer a segurança dos Sistemas de Informação e Comunicação, monitorando as atividades de rede e identificando invasões, fornecendo relatórios, a partir do levantamento destes incidentes.

Na Figura 11 é apresentada uma arquitetura aplicada numa empresa de comércio eletrônico, numa atividade de pagamento de cartão de crédito. O objetivo do sistema é, principalmente, a prevenção e identificação de invasões, podendo também gerar relatórios para serem utilizados, por exemplo em análise forense.

Figura 11 - Arquitetura aplicada numa empresa de comércio eletrônico.



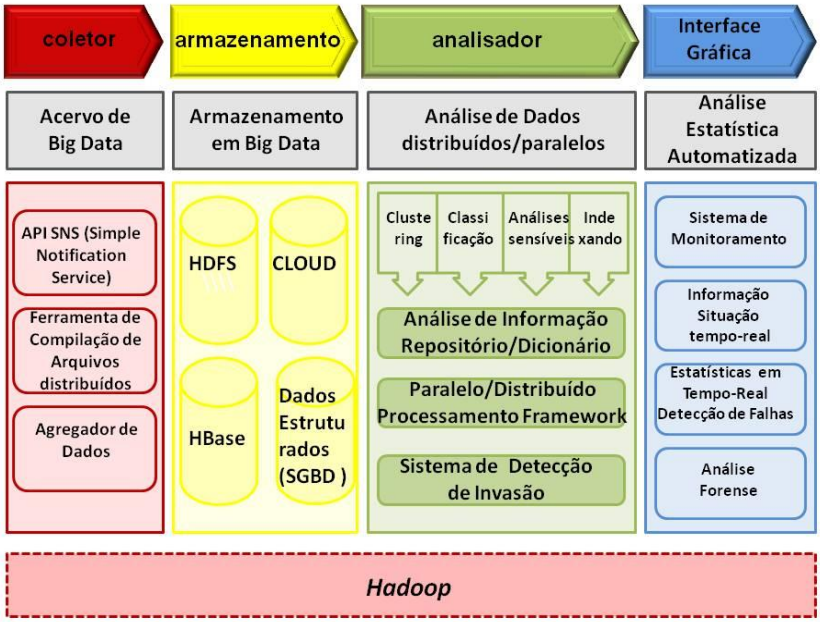
Fonte: (Marchal *et al.*, 2014)

Todas as informações de tráfego de DNS, HTTP, *honeypot* e registros de fluxo de números IP são coletados e analisados num sistema distribuído que tira proveito do grande volume em *Big Data*, utilizando uma plataforma *open source* como o Hadoop ou *Spark*.

A arquitetura integra duas partes, uma dedicada ao gerenciamento, armazenamento e processamento de dados de forma escalonável e distribuída, e outra dedicada à correlação desses dados, sendo aplicadas métricas com algoritmos específicos de probabilidade para calcular/avaliar a probabilidade e identificar se o nome de domínio, o pacote HTTP ou o fluxo de dados são maliciosos. Ocorrem alertas de identificação no sistema a partir da **pontuação** obtida através destes cálculos e caso o processo de correlação seja negativo o mesmo é finalizado

Outro sistema de IDS apresentado na Figura 12 produz alertas para uma suposta tentativa de ataque, verificando se o desvio entre um determinando evento e o seu comportamento padrão é superior a um limite pré-definido. Nessa figura apresenta-se uma plataforma de IDS sendo executado sobre a plataforma Hadoop, composta por quatro módulos principais: coletor, armazenamento, analisador e uma interface gráfica.

Figura 12 - Proposta de um IDS em Hadoop



Fonte: (Jeong *et al.*, 2012)

Segue uma breve explicação de cada módulo:

- **Módulo coletor:** executa busca de dados de *Web Pages*, redes socias, de ferramentas de extração de arquivos distribuídos e de agregadores de dados.
- **Módulo de armazenamento:** usado para armazenar e gerenciar volume de *Big Data* em *Storages*, e *Storages* de dados estruturados, fazendo uso de filtros de análise em tempo real.
- **Módulo analisador:** executa análises de *clusters*, e classifica os dados distribuídos/paralelos. Este módulo também desempenha um papel importante na análise de conteúdo, análise descritiva, análise preditiva, processamento de linguagem natural, *data mining* de textos, etc. O IDS é especialmente implementado para encontrar anormalidades utilizando o *MapReduce*.
- **Módulo GUI:** apresenta uma análise gráfica e automatizada dos resultados, obtendo informações de monitoramento em tempo real de estatísticas do sistema IDS.

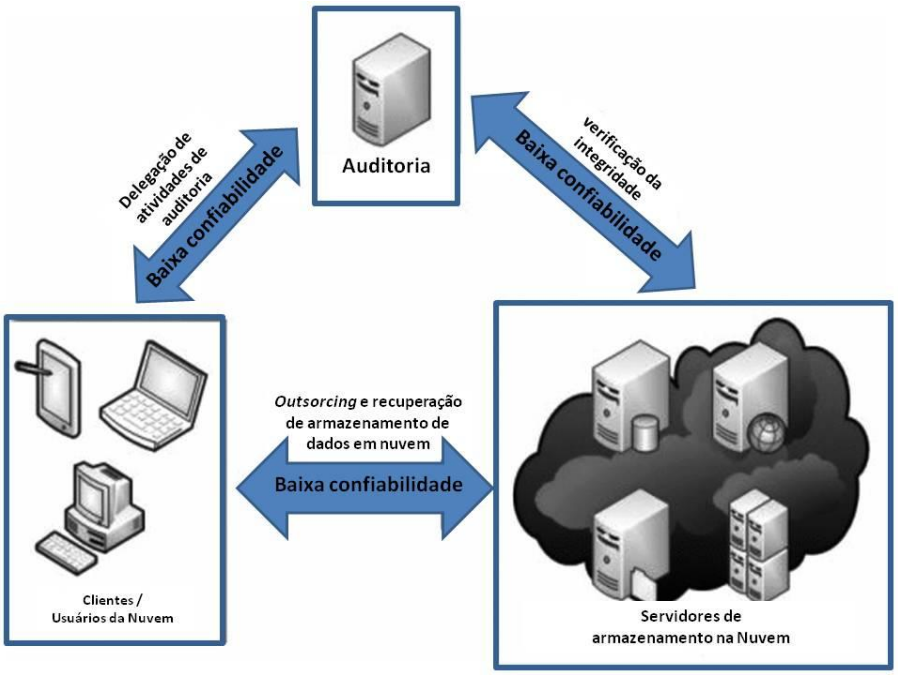
3.3.2 Auditoria de rede

Uma auditoria de rede consiste na verificação de dados de forma periódica em um processo regular de inspeção. Segundo (Liu *et al.*, 2015), para assegurar o uso confidencial ou sigiloso dos dados, têm sido utilizados requisitos funcionais e não funcionais (disponibilidade, consistência, integridade, identificação e agregação) em auditoria pública de *Big Data.*

Pelo fato de existirem múltiplos **nós de rede** num ambiente de *Big Data*, a disponibilidade é vista como uma grande vantagem, pois o acesso às informações é muito rápido. Porém, se alguns desses nós forem danificados, poderá acarretar problemas de integridade, pois não existirá uma confirmação correta da gravação dos dados. No entanto, o método tradicional para certificar a integridade dos dados torna-se desvantajoso em ambiente de *Big Data*, pois seu funcionamento consiste na obtenção de todos os blocos de dados no servidor para uma posterior análise.

Com isso, para resolver a complexidade na autenticação de análise de integridade de dados em *Big Data*, utiliza-se a técnica de *Multi-Replica* - Mur-DPA (múltiplas réplicas DPA) para uma auditoria pública de armazenamentos dinâmicos, conforme a Figura 13. O DPA - *Data Protection Act* tem como base, princípios que propoem uma adequada **movimentação** da informação. Estes princípios asseguram direitos específicos em relação às informações pessoais, além de exigir obrigações a essas organizações como determinadas responsabilidade no processamento de dados.

Figura 13 - Auditoria pública em um armazenamento dinâmico.



Fonte: (Liu *et al.*, 2015)

O sistema Mur-DPA contém uma estrutura de autenticação, com base numa estrutura de dados de árvores de dispersão chamada *Merkle Hash Tree* (MHT). Para suportar atualizações de dados e autenticações de índices de bloco dinâmico, são adicionados valores de classificação e de cálculo dos níveis MHT de nós. Esses valores são criados numa sequência do tipo *top-down*, e todas as réplicas de nível de nós para cada bloco de dados são organizados em uma sub-árvore da mesma réplica.

Como resultado, o sistema Mur-DPA produz uma verificação eficaz de atualizações de várias réplicas de dados de nós no mesmo instante, de maneira totalmente dinâmica e fazendo uso de autenticação de índices de blocos.

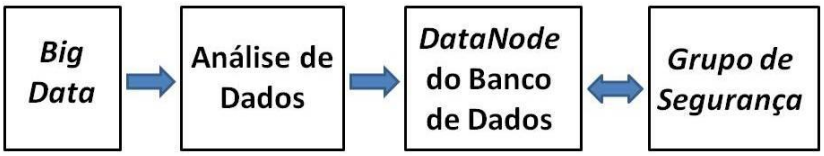
Na verificação de integridade, o Sistema Mur-DPA incorre muito menos sobrecarga de comunicação tanto para a verificação de atualização, como na verificação de integridade de conjuntos de dados em nuvem com várias réplicas, além de fornecer maior segurança contra os prestadores de serviços em nuvem não confiáveis.

3.4 Gerenciamento de chaves

Dada a grande variedade de dados estruturados, semiestruturados, e não estruturados em *Big Data*, se torna mais complexa a preservação de privacidade de dados, como mídias, textos, e-mails, arquivos XML, etc., quando comparados a dados estruturados, cujas consultas são realizadas pela linguagem SQL. Para que seja implantado um sistema que garanta um desempenho satisfatório de compartilhamento de arquivos entre servidores e usuários, é proposta a utilização de protocolos de autenticação rápida e dinâmica, com um gerenciamento de chaves simétricas e assimétricas.

O processo apresentado contém as etapas de análise de dados e a etapa de definições de níveis de prioridade de segurança (níveis de sensibilidade). A forma representativa na Figura 14 apresenta diferentes interfaces, iniciando a partir da fase de análise, com a filtragem de dados, fazendo uso de uma plataforma Hadoop*.* Depois é realizado um agrupamento dos dados, onde uma aplicação gera um *Datanode*, a partir de uma base de dados potencializando o carregamento dos dados a partir do Hadoop.

Figura 14 - Forma representativa do mecanismo de gerenciamento de chaves.



Fonte: (Islam e Islam, 2014)

Um algoritmo de escalonamento faz a interface do *Datanode* com a grupo pu suíte de segurança, selecionando o serviço mais adequado de acordo com um grupo pré-estabelecido (identificação, confidencialidade, integridade, autenticação, não repúdio) e o nível de prioridade ou sensibilidade (privativo, confidencial, público) de um pacote de segurança.

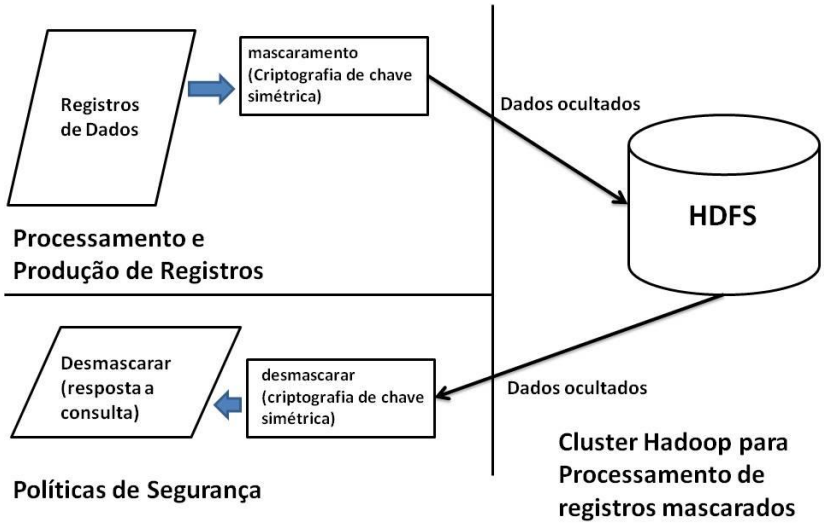
A classificação do *Datanode*, com base no nível de prioridade está associada para cada padrão ou norma, definida nos grupo de segurança, associado a um algoritmo. Por exemplo, para proporcionar privacidade de dados de um texto sigiloso, é selecionado o algoritmo 3DES de criptografia.

3.5 Mascaramento de dados

Em consequência da rapidez com que os dados são compartilhados, é cada vez mais necessária a utilização de métodos que fazem uso de ferramentas para ocultar informações de identificação pessoais ou confidenciais, tais como números de conta bancária e de cartões de crédito. Para isso, os dados devem estar num formato de anonimato e transferidos a locais seguros. Porém, dependendo dos algoritmos e das análises de inteligência de dados utilizado, uma identidade pessoal pode ser revelada, fazendo com que esses vazamentos de informações causem até mesmo a problemas éticos.

A arquitetura proposta na Figura 15 apresenta um procedimento de mascaramento de domínios confidenciais de registros de dados com criptografia de chave simétrica AES, e armazenamento em HDFS, para uma posterior análise. Quando for necessário um novo mascaramento, retornam-se os registros e as áreas que estão ocultas são decifradas com a mesma chave. Finalmente, a **qualidade** do mascaramento é medida por métricas, do tipo base k-*anonymity*. Essa métrica tem a propriedade de re-identificar correspondentes de uma tabela de microdados, preservando a veracidade da informação.

Figura 15 - Arquitetura de mascaramento de dados em *Big Data*

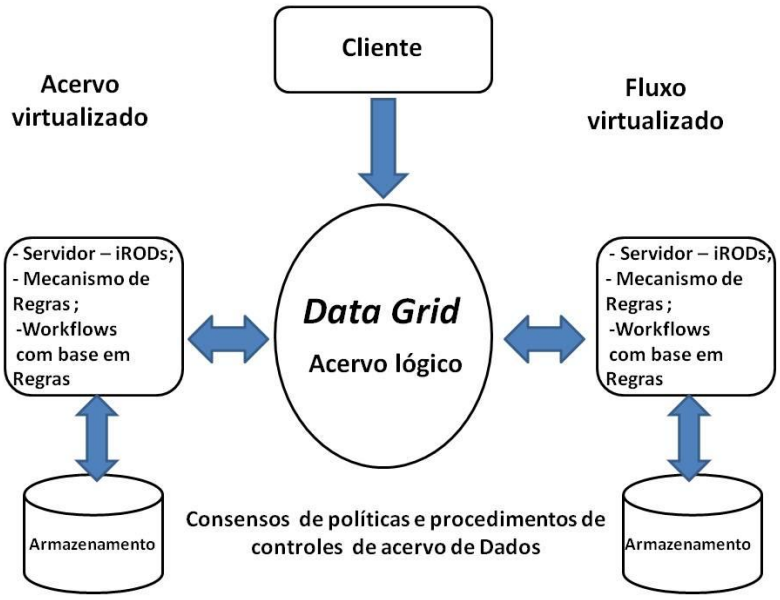


Fonte: (Sedayao *et al.*, 2014)

3.6 Gerenciamento em *DataGrid*

A arquitetura iRODS (Matturdi *et al.*, 2014), acrônimo de *Integrated Rule-Oriented Data System, ou Sistema Integrado de Dados Orientado a Regra* é um tipo de sistema de *DataGrid (Schnase et al., 2012) que possui middleware* de código aberto. Uma das principais funções do iRODS é conectar dados **não estruturados** com metadados. Está baseado em uma arquitetura cliente-servidor, com virtualização desses metadados e de mecanismo com base em políticas de regras, contendo mapeamento de *workflows* (fluxos de informação), conforme Figura 16.

Figura 16 - Fluxo *(Wokflow)* no gerenciamento iRODs



Fonte: ***(Schnase et al., 2012)***

O Sistema possui múltiplos recursos, tendo um gerenciamento de arquivos de forma autônoma, e fazendo uso de protocolos de segurança *Kerberos* e método de autenticação GSI (*Grid Security Infrastructure)* – protocolo de Infraestrutura de Segurança de Rede.

Os principais componentes, apresentados na Figura 16 são os servidores de banco de dados (*Storage*), responsável por armazenar e organizar as consultas de catálogos de metadados numa base de dados, um conjunto de servidores para alocação de recursos; um *DataGrid* na nuvem; mecanismo de regras distribuídas; e uma camada de armazenamento. Se os dados são armazenados em um disco rígido local, ou na nuvem, a camada de virtualização de arquivos no iRODS apresenta dados em formato de arquivos e pastas, dentro de um único *namespace.*

Figura 17 – Forma representativa da distribuição do Sistema Operacional

Fonte: (Rajasekar *et al.*, 2010)

Para o gerenciamento em alto nível, o Sistema disponibiliza uma *interface* gráfica de usuário para a interação com o *DataGrid* da nuvem, mantendo recursos de autenticação do usuário, autorização de acesso e auditoria de uso, representada na Figura 18.

Figura 18 - Arquitetura do Sistema iRODs

Fonte: (Rajasekar *et al.*, 2010)

O gerenciamento em *DataGrid* permite que os usuários acessem arquivos em ambiente distribuído, com base em seus atributos, replicando e sincronizando metadados, e conectando diferentes recursos de uma maneira lógica e abstrata.

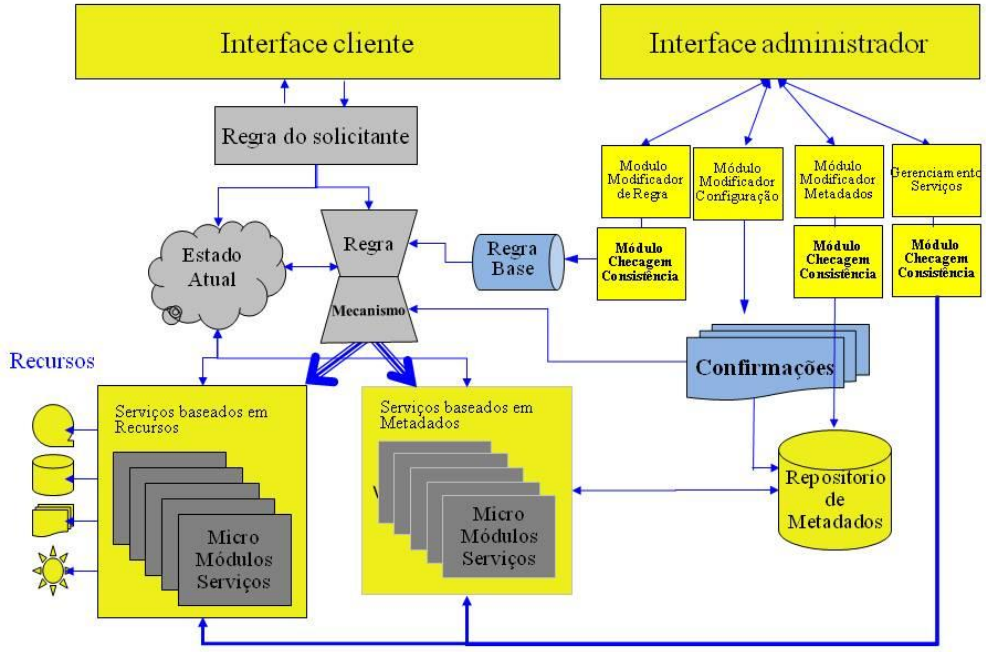
O Sistema possui uma arquitetura modular, que favorece um eficiente nível de proteção, e de implantação de políticas e procedimentos de segurança e privacidade de dados, considerando o uso em sistemas dinâmicos, descentralizados e distribuídos.

Este tipo de Sistema tem como característica principal uma flexibilidade no desenvolvimento de regras específicas de alocação, compartilhamento e gerenciamento de dados, de acordo com a necessidade organizacional.

O Sistema iRODS mantém os metadados em uma base de dados comum (Unix ou NTFS), fornecendo ao usuário ferramentas de consulta de metadados semelhantes aos utilizados na linguagem SQL, e uma hierarquia de arquivos e diretórios semelhantes ao Unix.

A Figura 19 apresenta as ferramentas do Sistema iRODS, onde acontece a junção de diferentes zonas de armazenamento, tornando os arquivos acessíveis em quaisquer locais de armazenamento, além de preservar os metadados durante a transferência para fins de gerenciamento, e de fornecer recursos nos repositório de dados, com disponibilidade, verificação de integridade e de validação.

Figura 19 - Funcionamento do Sistema iRODs



(Rajasekar *et al.*, 2010)

De maneira geral, o Sistema iRODS permite que os usuários façam operações complexas de gerenciamento e virtualização de dados distribuídos, com um controle de regras extensíveis de acessos, para cada extração de metadados, podendo efetuar publicações em bibliotecas digitais ou repositórios, garantindo que sejam nomeados, replicados ou arquivados com segurança, utilizando uma avançada tecnologia de criptografia e um sistema integrado de arquivamento de dados para permitir o rastreamento e monitoramento de acesso a dados, e para assegurar o cumprimento e a manutenção de proveniência de dados.

3.7 Segurança no Gerenciamento de Incidentes

A tecnologia de gerenciamento do Sistema SIEM, acrônimo de *Security Information and Event Management* (Sistema de Informação e Gerenciamento de Incidentes) tem como característica comparar ou correlacionar incidentes de rede, fazendo análise de alertas de segurança e de identificação de invasão. Seu foco principal é extrair, analisar e apresentar as variedades de dados.

No contexto de análise de dados, as primeiras gerações dos Sistemas de Identificação de Invasão (IDS) (Cárdenas, 2013) apresentavam procedimentos de segurança em camadas, realizando uma determinada reação em resposta a algum tipo de violação.

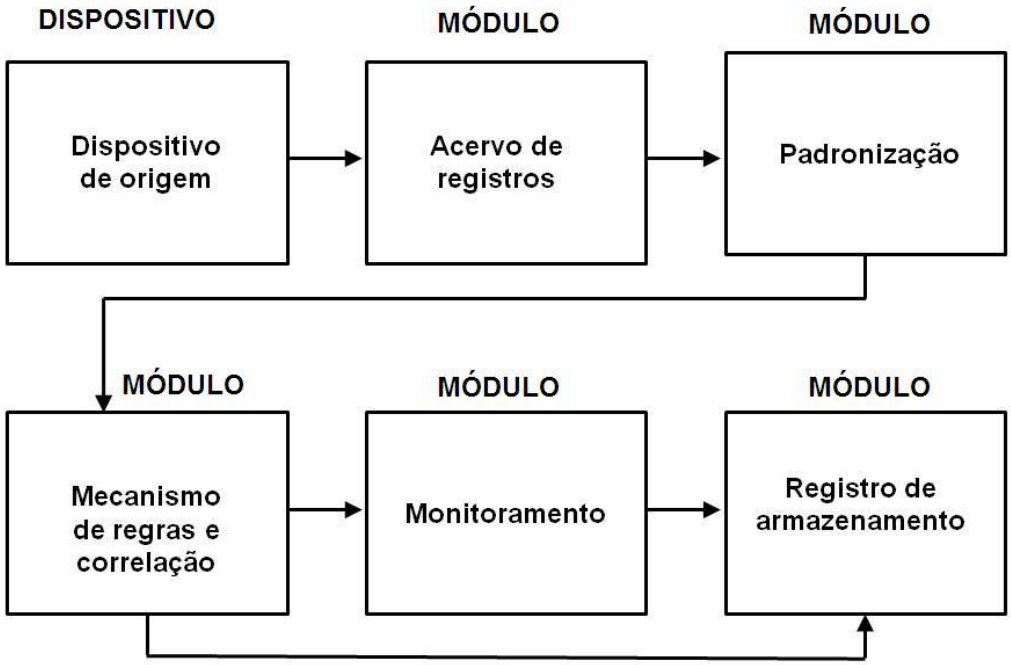
Atualmente os Sistemas SIEM gerenciam alerta de segurança através de regras de identificação de invasão, contendo alarmes com filtros de múltiplas fontes, e apresentando diversas informações para posteriores tomadas de decisão por analistas de segurança.

Apesar de existir uma imensa quantidade de informação disponível nas dimensões Volume e Variedade, torna-se vantajoso em *Big Data* a utilização de visualizadores de resposta de incidentes de eventos. Conforme Figura 20, essas interfaces executam consultas complexas de análise se registros e de *logs* de rede em grandes conjuntos de dados não estruturados. Também não há necessidade de excluir registros antigos porque estes podem ser reutilizados posteriormente para análise forense.

A descoberta de possíveis tentativas de invasão é uma atividade complexa e desafiadora para a segurança. Para o sucesso na análise de falhas ou identificação de ataques, tem se adotado procedimentos de análise de históricos de registros, de tráfego de rede ou de *web sites*.

Uma forma esquemática da arquitetura SIEM é representada na Figura 20.

Figura 20 - Arquitetura SIEM



Fonte: (Holik *et al.*, 2015)

Há alguns termos que aparecem na Figura 20 que merecem uma breve descrição:

Dispositivo de origem: Numa infraestrutura existem inúmeros dispositivos que servem como uma fonte de dados - servidores, aplicações críticas, dispositivos de rede, *firewalls*, etc.

Acervo de registros: módulo de extração e análise de dados obtidos a partir de dispositivos individuais. Ele utiliza o método *push*, armazenando em repositórios, que são enviados para o sistema pela própria fonte. Em seguida, o Sistema SIEM, serve como um receptor de dados e não intervém diretamente na gestão fonte. Um exemplo deste método é o protocolo *syslog*, que é combinado com um servidor *syslog*. A fonte fornece um endereço IP do servidor *syslog* e então recebe os dados enviados.

Padronização: Tem a função de converter os dados originais em uma forma padronizada e unificada.

Mecanismo de regras e correlação: aplicação de regras conforme correlação de descoberta de incidentes de segurança.

Registro de armazenamento: armazenamento de dados de registros de eventos de incidência. A maior prioridade no Sistema SIEM é salvar os arquivos de registros de dados, para garantir a rastreabilidade das informações e dos incidentes de violações registrados no SIEM.

Monitoramento: no conceito do Sistema SIEM é a interação com os dados no sistema. O sistema pode processar e avaliar a informação por si só, mas os administradores e usuários SIEM precisam acessar os incidentes globais.

Os sistemas SIEM proporcionam um avanço significativo na inteligência acionável de segurança, reduzindo o tempo para correlacionar, consolidar e contextualizar informações de violações diversas, como também para correlacionar dados históricos de longo prazo.

4 ANÁLISE DOS ESTUDOS

A preservação da segurança e da privacidade depende em grande parte das limitações tecnológicas envolvidas na capacidade de proteger, extrair, analisar e correlacionar conjuntos de dados potencialmente confidenciais ou importantes.

Os avanços em análise de *Big Data* também provêm ferramentas que se utilizadas de maneira trivial, possibilita a extração de dados sem autorização, e consequentemente habilitando violação de acesso a informações privadas. Como resultado, juntamente com o desenvolvimento de ferramentas de *Big Data*, é necessário também criar limites para evitar abusos.

4.1 Características reais de segurança e privacidade em *Big Data*

A seguir são listados alguns pontos que devem ser observados quando se discute segurança e privacidade de dados em *Big Data*.

- Para processar um grande volume de dados em ambientes distribuídos, é utilizada uma implementação do *framework MapReduce*. Em alguns cenários, o *MapReduce* pode não oferecer tanta confiabilidade e consequentemente poderá introduzir conteúdo com dados maliciosos;
- Num ambiente de *Cloud Computing*, os registros de dados são armazenados em diferentes níveis. *Big Data* produz uma auto-hierarquização no armazenamento de um grande volume de dados, criando novos desafios de segurança, pois passa a existir serviços de armazenamento não confiáveis e sem um controle sobre o local de *Storage*.
- Sistemas IDS no ambiente de *Big Data* representam outro desafio devido à produção de vários alertas para grandes volumes de dados, seguido consecutivamente por muitos alertas falsos positivos, dificultando este gerenciamento; (Zuech *et al.*, 2015)
- A privacidade dos dados é primordial, pois os dados recolhidos em operações dinâmicas de volume de *Big Data* contêm informações

privativas e o uso desses mecanismos por indivíduos de procedência suspeita dificulta o processo de preservação;

- A auditoria se faz necessária ara obter informações completas sobre um ataque. Porém, como em *Big Data* são enormes os níveis de volume de dados, os resultados podem conter muitos erros;
- Em sistemas tradicionais do tipo Data Warehouse, são estabelecidos padrões de segurança contra ataques, com implementações de mecanismos de análise de prevenção. Entretanto, em *Big Data,* devido ao grande volume, esses mesmos padrões são ineficientes;
- Empresas possuem especialistas para monitorar e aplicar procedimentos de segurança em tempo real, porém essa prática é inviável em *Big Data*.
- A utilização de criptografia em *Big Data* é um processo oneroso e, por vezes, pode até causar problemas de desempenho;
- Outro desafio em *Big Data* é fazer validação e filtragem. Usuários produzem conteúdos maliciosos e estes dados podem ser extraídos e armazenados.
- Metadados garantem uma maior facilidade na análise de diversas fontes. Porém, o crescente aumento de variedade dos dados causa erros nos resultados em relação a sua proveniência.
- Os sistemas tradicionais implementam o uso da esteganografia (técnica de esconder ou mascarar, sobrepondo informação) para proteger os dados em forma de imagens gráficas. Entretanto, essa técnica também é utilizada por indivíduos mal intencionados.
- Além da preservação da privacidade e confiabilidade, os dados utilizados em análise de *Big Data* podem incluir informações confidenciais e de propriedade intelectual. Arquitetos de sistemas devem garantir que os dados sejam protegidos e utilizados de acordo com as leis regulamentárias.

4.2 Análise na utilização dos mecanismos em segurança

Com base na literatura, é apresentada a Tabela 5 possibilidades de utilização dos mecanismos de segurança propostos no Capitulo 3, frente às características de *Big Data* explicadas no Capitulo 2. As ocorrências apresentam uma possível solução e o motivo da escolha ou da decisão de uso da técnica.

Conforme explicado anteriormente, as informações que constam na tabela são meramente sugestivas, tendo como base a literatura estudada.

Também existe a necessidade de mencionar casos publicados, onde são avaliadas as tecnologias mencionadas nesse trabalho. São constatadas por (*Jeong*, 2012) algumas limitações da plataforma Hadoop*,* descritas a seguir:

- Perda de desempenho ao executar uma extração de dados no input/output do sistema de arquivos HDFS, pois normalmente essa ação precisa ser realizada antes e depois de uma função / atividade (*Jobs*);
- Devido as frequentes alterações do volume de dados, torna-se incapaz de produzir relatórios em intervalos de segundos.
- Onde ocorrem frequentes inclusões, exclusões e atualizações de dados

A arquitetura Hadoop é menos eficiente que os tradicionais SGBD não relacional, ou seja, para dados menos volumosos, deve-se avaliar soluções tradicionais.

Porém, em um estudo de caso publicado há alguns anos pelo Banco americano *Zions Bancorporation (Cárdenas, 2013)*, foi publicada uma comparação de desempenho da plataforma Hadoop com as ferramentas de Sistemas SIEM (Chickowski, 2012), para análise de dados numa maior frequência de incidentes e volume de dados. Constatou-se que ao utilizar Hadoop em execução com *Hive*, houve uma maior rapidez no processamento de dados. As consultas de um mês de carga, que levavam de 20 minutos a uma hora para serem processadas passaram para cerca de um minuto, apresentando os mesmos resultados.

Tabela 4 - Problemas e soluções para cada característica em *Big Data*

Característica	Ocorrência	Solução	Motivo
Volume	Dificuldades para armazenar quantidade de dados em sistemas tradicionais (exabyte ou zettabyte)	Armazenamento no sistema de arquivos HDFS	Possui escalabilidade para armazenar e processar grandes conjuntos de dados de modo confiável
Variedade	Manipulação de diversos tipos de dados (textos, imagens e vídeos).	Arquitetura Hadoop	Aceita dados estruturados de forma relacional, semi estruturados e não estruturados
		Monitoramento	Utilizar aplicativos que produzem alertas para uma suposta tentativa de ataque
		Gerenciamento de Chaves	A utilização de protocolos com autenticação rápida e dinâmica favorece a antecipação de incidentes de segurança
		SGBD não relacional (NoSQL)	Com o frequente fluxo de dados no input/output, Hadoop torna-se menos eficiente.
Velocidade	Necessidade de alto poder de processamento	Processo distribuído e em paralelo (*MapReduce*)	Pode-se terceirizar o processamento em muitos *nodes* formando um aglomerado computacional.
Variabilidade	Assegurar integridade dos dados	Auditoria de rede	Algoritmos baseados em políticas com delegação de confiabilidade.
Veracidade	Preservar o anonimato informações de identificação pessoal ou confidenciais	Mascaramento de dados	Limita a exposição de dados dinâmicos
Volatilidade	Análise de ciclo de vida dos dados	Gerenciamento de *DataGrid*	Atua numa camada intermediária entre os armazenamentos relacionais e os dados processados em plataformas robustas, atendendo exigências de retenção de dados.
Validade	Armazenamento de códigos maliciosos	Segurança em *Cloud Computing*	Uso de chaves de codificação somente para usuários autorizados.
Visualização	Dados estatísticos de eventos	Gerenciamento de Incidentes (SIEM)	Existindo alta frequência nas alterações do volume de dados, torna-se ineficiente a produção de alertas de segurança em Hadoop
Valor	Custo de implementação da plataforma de *Big Data*	Hadoop	*Framework* de código aberto
Complexidade	Gerenciamento de metadados	Segurança em *DataGrid*	Possibilita alcançar escalabilidade de gerenciamento

5 CONSIDERAÇÕES FINAIS

Com objetivo de contribuir num futuro projeto de implantação de mecanismos de segurança e privacidade em *Big Data*, são listadas orientações mais significativas a serem consideradas, tais como:

a) Uso da técnica de mascaramento e criptografia de dados para evitar a vulnerabilidades das informações numa base de dados;
b) Analisar todos os *logs* (registros) de incidentes, para identificar com precisão atividades não autorizadas de sistema;
c) Preservar um sistema de gerenciamento de autenticação de acesso. Sempre considerar o princípio da segurança: a aplicação do menor privilégio no uso do sistema ao usuário é a sua identificação;
d) Conhecer inteiramente o sistema de segurança da aplicação que tem acesso a informações confidenciais;
e) Usar o sistema SIEM para obter os registros de dados revelando com clareza os eventos de incidência;
f) Monitorar e auditar dados com objetivo de extrair e revelar uma maior quantidade possível de fluxo de informações.

Conforme a tabela 5, todas as tecnologias estudadas neste trabalho se baseiam resumidamente nos métodos de aplicação de criptografia, autenticação e métrica, com a possibilidade de utilização em bancos de dados não relacionais, em *DataGrid*, *Cloud Computing* e em processamentos em massa com a plataforma Hadoop.

Tabela 5 - Estudo de Segurança e Privacidade em *Big Data*

Mecanismos	Objetivo	Método de aplicação
Segurança em Hadoop	Preservação da segurança específica para esta plataforma.	**(CRIPTOFRAFIA)** confiabilidade entre o usuário e o servidor de metadados, algoritmos de criptografia.
	Segurança HDFS	
Segurança em *Cloud Computing*	Segurança nos dados armazenados na nuvem	**(AUTENTICAÇÃO E CRIPTOFRAFIA)** e compressão de dados.
	Segurança no processo de armazenagem na nuvem	
Monitoramento de rede	Identificação de invasão de sistemas	**(MÉTRICA)** probalidade de identificação de códigos maliciosos
Auditoria de rede	Auditoria Dinâmica de armazenagem em *Big Data*	**(MÉTRICA)** Verificação de movimentação ou fluxo de informação na rede.
Gerenciamento de chaves	Segurança de dados não estruturados	**(CRIPTOFRAFIA)** Análise de dados (filtragem, *clustering*, classificação) e segurança (normas e algoritmos existentes)
Mascaramento de dados	Mascaramento de informações confidenciais	**(MÉTRICA)** decifragem e reidentificação de dados
Gerenciamento em *DataGrid*	Sistema Integrado de Dados Orientado a Regra (iRODs)	**(MÉTRICA)** Aplicação de regras de alocação, compartilhamento e Gerenciamento de Metadados
Gerenciamento de Incidentes de dados	Alertas de segurança (SIEM)	**(MÉTRICA)** Recolher, analisar e apresentar resposta de incidentes de eventos.

5.1 Melhores práticas em *Big Data*

A *Cloud Security Alliance* (CSA), organização norte-americana sem fins lucrativos, e o *Big Data Working Group* (BDWG), representado por empresas como *eBay*, *Cisco*, *Dell*, IBM, fizeram um levantamento com as melhores práticas (CSA, 2013) a serem aplicadas no objetivo de garantir segurança e privacidade em *Big Data*:

a) Autorizar o acesso de arquivos através de políticas de segurança predefinidas;

b) Manter os dados protegidos por algum algoritmo de criptografia;

c) Implementar sistemas de criptografia baseado em políticas de regras;

d) Utilizar *softwares data analytics* para encontrar conexões anômalas ao *cluster* Hadoop*;*

e) Implementar *softwares data analytics* para assegurar tanto a privacidade, como a confiabilidade;

f) Providenciar levantamento de informações de auditoria de dados.

5.2 Os desafios em *Big Data*

A seguir são apresentados os desafios em *Big Data*, dos quais foram apontados nos relatórios da CSA, e que possuem uma maior relevância para este trabalho, tais como:

a) Segurança em computação de processamento distribuído – *MapReduce:* devem-se assegurar medidas de prevenção de ataques no **mapeador** de dados, ou seja, quando acontecer a leitura dos blocos de dados, na execução dos cálculos que produzem os *outputs* da lista de pares de chave-valor. Mapeadores não confiáveis poderiam retornar com resultados errados quase impossíveis de identificar em grandes volumes de dados. Mapeadores de dados também podem conter vazamentos intencionais ou não intencionais que poderiam analisar um registro particular que compromete a privacidade.

b) Melhores práticas em segurança para banco de dados não relacionais (NoSQL), pois o foco está na análise de dados e os desenvolvedores incorporam segurança apenas no *middleware*, necessitando de suporte de segurança geral no projeto.
c) Segurança de Monitoramento em tempo real – devido ao crescente número de alertas de falsos positivos produzidos por dispositivos de segurança.
d) *Softwares Data Analytics* - precisam ter acesso escalonável e combinável na extração e na análise (*Data mining* e *Visual Analytics*), pois se realiza **anonimato da informação**, processo de criptografar ou remover informações de identificação pessoal a partir de conjuntos de dados, não é suficiente para manter a privacidade do usuário. Um invasor mal-intencionado pode extrair informações privadas de clientes.
e) Controle na procedência de dados – ficará cada vez mais complexo esse controle, na documentação dos dados de interesse, e em fornecer um histórico de registro e das suas origens, devido à grande variedade. Estes metadados são importantes nos casos em que os peritos forenses precisam fazer uma investigação e conhecer com precisão a fonte de dados.

5.3 Futuro do *Big Data*

No relatório mais recente do *Gartner Group* sobre o *Hype Cycle*, nota-se a inexistência de *Big Data* no gráfico de 2015, ao comparar com anos anteriores. Este relatório apresenta resultado de pesquisa sobre as tecnologias mais populares e emergentes no mundo.

A explicação mais provável desse acontecimento é que *Big Data* sempre foi um termo genérico, justificando a opinião de (Davenport, 2014), mencionada nesse trabalho. E com base nos relatório, nos próximos anos faz-se necessário dar uma maior importância nas tendências tecnológicas relacionadas à *Big Data* como a Internet das Coisas ou *Internet of Things* (IoT), que permanece consistentemente quase no pico do ciclo em vários anos seguidos.

Apesar da segurança dos dados **Segurança Digital** serem um fator crítico para governos e empresas privadas, ela aparenta ter menos importância ou aceitação popular, segundo a análise. Uma possível explicação para esse fato é que para os meios publicitários a Segurança da Informação é tratada como mais um item que agrega valor a tecnologia.

5.4 Trabalhos Futuros

Como possíveis trabalhos futuros, dando continuidade ao tema de segurança e privacidade em *Big Data*, pode-se apontar:

- Estudo da plataforma *Spark*: que da mesma maneira como *Haddop*, viabiliza processamento e distribuição de dados em grandes volumes. Porém a diferença é que esses dados são processados e armazenados em memória, fazendo com que a velocidade aumente mais de 100 vezes (Marchal *et al.*, 2014).

- Estudo de Internet das Coisas: relacionado ao conceito de objetos do cotidiano, que vão desde máquinas industriais até dispositivos portáteis, dos quais fazem uso de sensores internos utilizados para reunir informações pra serem usadas em tomadas de decisões através da rede. Nesse contexto, se faz necessário o estudo da incorporação do protocolo SSL (*Secure Socket Layer*), padrão global em tecnologia de segurança que gera um canal criptografado entre servidores web e dispositivo móvel.

- Estudo das correlações e implicações no crescimento dos dados produzidos pela IoT, conjuntamente com *Cloud Computing.*

REFERÊNCIAS

ADLURU, P.; DATLA, S. S.; ZHANG, X. Hadoop eco system for big data security and privacy**. Systems, Applications and Technology Conference (LISAT), 2015 IEEE Long Island, 2015, 1-1 May 2015. p.1-6.**

BERNADETTE, F. L. O., H.R. PONTES, J. C.S. NoSQL no desenvolvimento de aplicações Web colaborativas**, Universidade Federal de Pernambuco 2011.**

BRITO, R. W. Bancos de Dados NoSQL x SGBDs Relacionais:Análise Comparativa**. Faculdade Farias Brito e Universidade de Fortaleza 2010**

CHAU, M.; REITH, R.; UBRANI, J. Smartphone OS Market Share. **2015. Disponível em: < http://www.idc.com/prodserv/smartphone-market-share.jsp >. Acesso em: abril de 2016.**

CHENG, H. et al. Secure big data storage and sharing scheme for cloud tenants**. China Communications, v. 12, n. 6, p. 106-115, 2015. ISSN 1673-5447.**

CHICKOWSKI, E. A Case Study In Security Big Data Analysis**. 2012. Disponível em: < http://www.darkreading.com/analytics/security-monitoring/a-case-study-in-security-big-data-analysis/d/d-id/1137299 >. Acesso em: abril.**

CSA. Expanded Top Ten Big Data Security and Privacy Challenges**: Cloud Security Alliance BIG DATA WORKING GROUP 2013.**

CÁRDENAS, A. A. Big Data Analytics for Security Intelligence**: Cloud Security Alliance (CSA). Technical report 2013.**

DAVENPORT, T. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities**. Boston, Massachusetts: Harvard Business Review Press, 2014.DAVIS, K.** Ethics of Big Data. **Sebastopol, CA: O`Reilly Media, 2012.**

GARTNER. IT Glossary **Search. 2016. Disponível em: < http://www.gartner.com/it-glossary/big-data/ >. Acesso em: abril de 2016**

HASHEM, I. A. T. et al. The rise of “big data” on cloud computing: Review and open research issues. **Information Systems, v. 47, p. 98-115, 1// 2015. ISSN 0306-4379. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0306437914001288 >.**

HOLIK, F. et al. The deployment of Security Information and Event Management in cloud infrastructure. Radioelektronika **(RADIOELEKTRONIKA), 2015 25th International Conference, 2015, 21-22 April 2015. p.399-404.**

IBM. What is big data? **, 2015. Disponível em: < http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html >. Acesso em: fevereiro de 2016**

IRIZARRY, R.; PENG, R.; LEEK, J. The Big in Big Data relates to importance not size**. p. Simply Statistics A statistics blog, 2014. Disponível em: < http://simplystatistics.org/2014/05/28/the-big-in-big-data-relates-to-importance-not-size/ >. Acesso em: 28 Maio de 2016.**

ISLAM, M. R.; ISLAM, M. E. An approach to provide security to unstructured Big Data**. Software, Knowledge, Information Management and Applications (SKIMA), 2014 8th International Conference on, 2014, 18-20 Dec. 2014. p.1-5.**

J HAN, E. H., G LE, J DU. Survey on NoSQL Database **- PCN&CAD Center: Beijing University of Posts and Telecommunications 2011.**

JEONG, H. D. J. et al. Anomaly Teletraffic Intrusion Detection Systems on Hadoop-Based Platforms: A Survey of Some Problems and Solutions. **2012 15th International Conference on Network-Based Information Systems, 2012, 26-28 Sept. 2012. p.766-770.**
KEARNEY, A. T. Rethinking Personal Data **- World Economic Forum 2013. 2014. Disponível em: < http://reports.weforum.org/rethinking-personal-data >.**

KUONEN, D. The Big in Big Data relates to importance not size**. 2014. Disponível em: < http://simplystatistics.org/2014/05/28/the-big-in-big-data-relates-to-importance-not-size/ >. Acesso em: abril de 2016.**

LIU, C. et al. MuR-DPA: Top-Down Levelled Multi-Replica Merkle Hash Tree **Based Secure Public Auditing for Dynamic Big Data Storage on Cloud. IEEE Transactions on Computers, v. 64, n. 9, p. 2609-2622, 2015. ISSN 0018-9340.**

MARCHAL, S. et al. A Big Data Architecture for Large Scale **Security Monitoring. 2014 IEEE International Congress on Big Data, 2014, June 27 2014-July 2 2014. p.56-63.**

MATTURDI, B. et al. Big Data security and privacy: **A review. China Communications, v. 11, n. 14, p. 135-145, 2014. ISSN 1673-5447.**

MURTHY, P. et al. Big Data Taxonomy: Cloud Security Alliance **BIG DATA WORKING GROUP. Technical report 2014.**

NETWORK TECHNICAL AUTHORITY. *Big Data* : A NTA Technical Vision **DE&S Information Systems and Services: August. Rev 0.4' 2013.**

PERLROTH, N.; LARSON, J.; SHANE, S. N.S.A. Able to Foil Basic Safeguards of Privacy on Web**. 2013. Acesso em: 10 janeiro de 2016.**

RAJASEKAR, A.; MOORE, R.; HOU, C. iRODS Primer: integrated rule-oriented data system: **Book in Synthesis Lectures on Information Concepts Retrieval and Services 2010.**

SARALADEVI, B. et al. Big Data and Hadoop-a Study in Security Perspective. **Procedia Computer Science, v. 50, p. 596-601, // 2015. ISSN 1877-0509.**

Disponível em: < http://www.sciencedirect.com/science/article/pii/S187705091500592X >.

SCHNASE, J. L. et al. RENCI Announces E-iRODS **Consortium at SC12: RENCI Press Release 2012.**

SEDAYAO, J.; BHARDWAJ, R.; GORADE, N. Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: **Experiences and Issues. 2014 IEEE International Congress on Big Data, 2014, June 27 2014-July 2 2014. p.601-607.**

TAURION, C. Big Data. **Rio de Janeiro: Brasport, 2013.**

TECHNOLOGY, N. I. O. S. A.; NIST. The NIST Definition of Cloud Computing. **September 2011. Disponível em: < http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf >. Acesso em: abril de 2016.**

TERZI, D. S.; TERZI, R.; SAGIROGLU, S. A survey on security and privacy **issues in big data. 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST), 2015, 14-16 Dec. 2015. p.202-207.**

VIJAYAKUMAR, V. et al. Big Data, Cloud and Computing ChallengesBig Data Security Issues Based on Quantum Cryptography and Privacy with Authentication for Mobile Data Center**. Procedia Computer Science, v. 50, p. 149-156, 2015/01/01 2015. ISSN 1877-0509. Disponível em: < http://www.sciencedirect.com/science/article/pii/S1877050915005785 >.**

YU, S. et al. Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing**. INFOCOM, 2010 Proceedings IEEE, 2010, 14-19 March 2010. p.1-9.**

ZUECH, R.; KHOSHGOFTAAR, T. M.; RANDALL, W. Intrusion detection and Big Heterogeneous Data: a Survey: **Journal of Big Data, a Springer Open Journal 2015.**

GLOSSÁRIO

[Cluster] - consiste em diversos mainframes fazendo tarefas em conjunto, porém, agindo como se fossem em um único sistema.

***[Criptografia]* -** é um conjunto de técnicas para embaralhar ou esconder a informação de acesso não autorizado. O objetivo da criptografia é transformar um conjunto de informação legível, como um e-mail, por exemplo, em um emaranhado de caracteres impossível de ser compreendido. O conceito chave é que apenas quem tem a chave de decriptação seja capaz de recuperar, por exemplo, um e-mail em formato legível. Mesmo conhecendo todo o processo para esconder e recuperar os dados, a usuário sem autorização não consegue descobrir a informação sem a chave de decriptação.

***[Hbase]* -** é um banco de dados distribuído orientado a coluna, escrito na linguagem de programação Java. Possui integração com o Hadoop, e utiliza o *MapReduce* para distribuir o processamento dos dados.

***[RSA]* -** é um algoritmo de criptografia de dados. Seu nome se deve a três professores do Instituto de Tecnologia de Massachusetts (MIT), *Ronald Rivest, Adi Shamir* e *Leonard Adleman*, fundadores da atual empresa *RSA Data Security*, Inc.. É avaliada como uma das mais bem sucedidas implementações de sistemas de chaves assimétricas, alem de ser considerado um dos algoritmos mais seguros já inventados. Foi o primeiro algoritmo a possibilitar criptografia e assinatura digital, alem de apresentar inovação no uso de criptografia de chave pública.