http://dx.doi.org/10.23925/2179-3565.2020v11i3p13-34

RISUS - Journal on Innovation and Sustainability
volume 11, número 3 - 2020
ISSN: 2179-3565
Editor Científico: Arnoldo José de Hoyos Guevara
Editor Assistente: Rosa Rizzi
Avaliação: Melhores práticas editoriais da ANPAD

THE AGE OF BIG DATA: MAIN IMPLICATIONS ON SECURITY AND PRIVACY AND NEW TECHNOLOGIES THAT CAN HELP INVESTIGATIVE PROCESSES AND DETECTION OF REAL-TIME FRAUD

A era do Big Data: principais implicações sobre segurança e privacidade e as novas tecnologias capazes de auxiliar processos investigativos e detecção de fraudes em tempo real

Alessandro Marco Rosini Filho[1], Alessandro Marco Rosini[2], Angelo Palmisano[3].
Mackenzie/SP[1], UNIAN/SP, UNIVAG/MT, UNIFACVEST/SC[2], UNIVAG/MT[3]
Email: ale07.rosini@gmail.com, alessandro.rossini@hotmail.com, angelopalmisano@uol.com.br

ABSTRACT

With the great expansion in the volume of data created in recent years, coupled with the speed at which the most distinct information is generated from multiple and different origins, this work developed through qualitative method, based on a review and descriptive analysis of main publications made on the topic Big Data, based on security and privacy, in order to understand and analyze the different thoughts on the subject. We sought to submit the concepts involving this phenomenon termed as Big Data, as well as expose the issues and challenges surrounding the security and privacy of data stored and used by organizations that adopt technologies capable of extracting determining values on existing data, detecting that there are still many aspects that effect which must be carefully evaluated by the organizations that have this type of technology, or aims to implement it. In this concern, it was of interest in this research and check to see that these new technologies can make a decisive contribution to security optimization engine of technological environments and own technology itself, can bring important advances with regard to the prevention and detection fraud, more efficiency and agility in investigative processes and add more strategic intelligence with respect to public safety, aimed at combating crime.

Keywords: Big Data, IT security, Systems investigation, Fraud detection and prevention, Data privacy and security.

ACEITO EM: 09/07/2020
PUBLICADO: 30/09/2020

RISUS - Journal on Innovation and Sustainability
volume 11, número 3 - 2020
ISSN: 2179-3565
Editor Científico: Arnoldo José de Hoyos Guevara
Editor Assistente: Rosa Rizzi
Avaliação: Melhores práticas editoriais da ANPAD

A ERA DO BIG DATA: PRINCIPAIS IMPLICAÇÕES SOBRE SEGURANÇA E PRIVACIDADE E AS NOVAS TECNOLOGIAS CAPAZES DE AUXILIAR PROCESSOS INVESTIGATIVOS E DETECÇÃO DE FRAUDES EM TEMPO REAL

The age of Big Data: main implications on security and privacy and new technologies that can help investigative processes and detection of real-time fraud

Alessandro Marco Rosini Filho[1], Alessandro Marco Rosini[2], Angelo Palmisano[3].
Mackenzie/SP[1], UNIAN/SP, UNIVAG/MT, UNIFACVEST/SC[2], UNIVAG/MT[3]
Email: ale07.rosini@gmail.com, alessandro.rossini@hotmail.com, angelopalmisano@uol.com.br

RESUMO

Com a grande expansão no volume de dados criados nos últimos anos, aliado à velocidade com que as mais distintas informações são geradas a partir de múltiplas e distintas origens, este trabalho desenvolvido por meio de método qualitativo, baseado em uma análise crítica e descritiva acerca das principais publicações realizadas sobre o tema Big Data, baseado na segurança e privacidade, com o objetivo de conhecer e analisar os diferentes pensamentos sobre o tema. Buscou-se ainda apresentar os conceitos que envolvem este fenômeno denominado como Big Data, bem como expor os aspectos e desafios que envolvem a segurança e privacidade dos dados armazenados e utilizados pelas organizações que adotam as tecnologias capazes de extrair valores determinantes sobre os dados existentes, detectando que existem ainda muitos aspectos neste sentido que precisam ser cuidadosamente avaliados pelas organizações que possuem este tipo de tecnologia, ou almejam implementá-la. Além desta preocupação, foi de interesse desta pesquisa verificar e constatar que estas novas tecnologias podem contribuir de forma determinante como mecanismo de otimização da segurança de ambientes tecnológicos e da própria tecnologia em si, podendo trazer importantes avanços no que diz respeito à prevenção e detecção de fraudes, mais eficiência e agilidade em processos investigativos e agregar mais inteligência estratégica no que diz respeito à segurança pública, voltada para o combate à criminalidade.

Palavras–Chave: Big Data, Segurança de TI, Investigação de sistemas, Detecção e prevenção de fraudes, Segurança e Privacidade de dados.

INTRODUÇÃO

Os aspectos que envolvem o termo Big Data têm sido um dos assuntos mais abordados no mundo corporativo desde o ano passado. (Mcduling, 2013)

Estima-se que no ano de 2013 o mercado voltado para o Big Data movimentou cerca de 70 bilhões de dólares, e a previsão é que este segmento chegue crescer até 40% ao ano até 2015. (Barreira Junior, 2013)

O termo Big Data não se restringe apenas a uma grande quantidade de dados. As tecnologias que envolvem este conceito possibilitam encontrar importantes percepções nos dados existentes, além de contribuir com a captura e análise de dados futuros, trazendo mais agilidade e confiança para antecipar riscos ou responder a desafios nos mais distintos segmentos, dos quais possam ser aplicados os recursos do Big Data.

É possível afirmar que cada vez mais o mundo digital se aproxima da realidade, uma vez que tudo, ou todas as ações cotidianas, eventos, fatos, entre outros, podem transformar-se em dados armazenados eletronicamente.

Considerando este potencial de intenso volume, aliados à velocidade e à variedade as quais os dados eletrônicos são produzidos, esta pesquisa busca a resposta para algumas questões pontuais: será que é possível extrair valor destas informações de maneira rápida e eficiente ao ponto de se detectar, ou até mesmo prever fraudes? É possível que sejam otimizados processos investigativos por meio do Big Data? É possível que as tecnologias Big Data contribuam para melhorar a segurança de ambientes tecnológicos, ou até mesmo de pessoas? A existência de tantas informações armazenadas pode trazer algum risco para as empresas que as consomem?

O Big Data, aplicado aos diferentes segmentos, traz consigo alguns desafios específicos no que diz respeito à privacidade e segurança dos dados existentes dentro de um ecossistema de tecnologia da informação voltado para lidar com grandes volumes de dados, que precisam ser considerados pelas empresas para evitar possíveis danos, multas ou processos litigiosos.

No que diz respeito à possibilidade das tecnologias orientadas ao Big Data, por meio de sua capacidade analítica, acredita-se que seja possível efetuar a análise e correlação de históricos de dados existentes com dados capturados em tempo real, aliados com a capacidade do rápido processamento de dados juntamente com a competência em lidar com diversificados tipos de dados de maneira rápida e eficiente, é possível supor que estas características possam contribuir para satisfazer os problemas de pesquisa deste trabalho, como agregar melhorias para processos investigativos, mecanismos de defesa e detecção e prevenção de fraudes.

Essa pesquisa tem como objetivo inicial expor os conceitos e tecnologias que estão diretamente ligadas ao termo Big Data. Busca-se também, identificar os possíveis impactos sobre a segurança e privacidade dos dados consumidos por esta nova geração de tecnologias, além de expor os principais pontos de desafio para garantir a proteção dos dados distribuídos nos chamados ecossistemas Big Data.

Por fim, o presente estudo procura mostrar as possíveis aplicações dessas novas tecnologias em segmentos que possuem, de certo modo, uma forte relação com a computação forense, como no caso de prevenção e detecção de fraudes, novos recursos que possam ser aplicados a fim de otimizar processos investigativos, e como mecanismo de defesa e combate ao crime.

A grande maioria das organizações enxergam o Big Data como uma maneira de transformar seus dados em vantagem competitiva, extraindo valor de muitas informações hoje não exploradas. É um segmento que promete atrair grandes investimentos nos próximos anos, e consequentemente trará oportunidades para quem for conhecedor das diversas tecnologias envolvidas e souber extrair valor e utilidade destas informações. No segmento de segurança da Informação, detecção e investigação de fraudes isto não é diferente, pois algumas ferramentas começam a surgir, utilizando-se também deste tipo de tecnologia. Um típico exemplo dessa delicada situação na área política é a confusão proposital criada através dos fake news como indicam as recentes declarações da Kaspersky em Forbes Colombia *De la ciberseguridad a la ciberinmunidad*[1], bem como a

[1] https://forbes.co/2020/04/14/tecnologia/la-opinion-de-kaspersky-en-forbes-colombia-de-la-ciberseguridad-a-la-ciberinmunidad/

manifestação da própria Facebook[2] que decidiu excluir páginas da CLS Strategies dos EUA que acostumava vender fake news para a direita latino-americana.

A significativa atenção que o assunto Big Data tem demandado para o universo de tecnologia da informação atualmente, desperta um forte interesse em detectar-se em que sentido esta nova abordagem em lidar com massivos e distintos volumes de dados pode contribuir para melhorar os mecanismos de defesa, detecção e investigação, tanto de ambientes tecnológicos, quanto, por meio da própria tecnologia empregada, os mecanismos de defesa para a sociedade como um todo, preocupando-se também com os aspectos que envolvem a proteção e privacidade destes dados coletados e utilizados para estas e outras finalidades, as quais possam ser aplicadas as tecnologias orientadas ao Big Data.

1 BIG DATA

1.1 A origem do termo Big Data

O termo Big Data foi utilizado cientificamente pela primeira vez em Outubro de 1997 (Press, 2013), para destacar a ocorrência de grandes conjuntos de dados capazes superar as capacidades de memória principal, disco local e até mesmo disco remoto, no contexto de visualização computacional (Cox; Ellsworth, 1997).

Com a rápida propagação da Internet nos anos 90, a quantidade de dados armazenados eletronicamente se fazia expandir de forma exponencial, o que provocou a necessidade das empresas criarem modelos para o gerenciamento destas informações. Um dos setores em franca expansão da última década do século passado, o e-commerce, observou que o gerenciamento de dados deveria ser feito sobre três aspectos: Volume, Velocidade e Variedade (Laley, 2001). Estes "3 Vs" são hoje as três dimensões para a definição do termo Big Data (Press, 2013).

A constante evolução e diminuição de custos dos equipamentos tecnológicos, fez com que, em 2008, Bryant (2008) destacasse o grande potencial da tecnologia que envolve o Big Data:

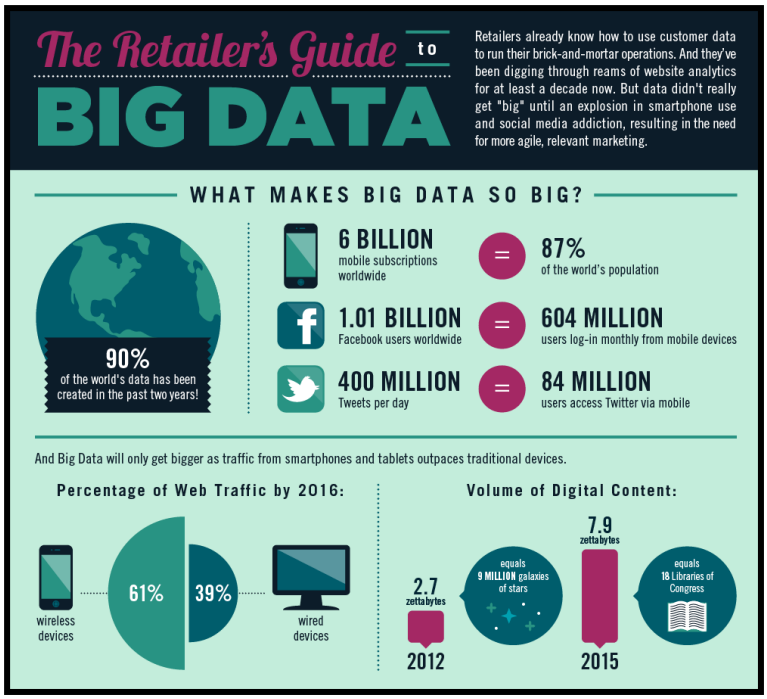
> "Assim como os sistemas de busca têm transformado a maneira de acesso à informação, outras formas de big data podem e irão transformar as atividades das empresas, pesquisadores científicos, médicos e operações de defesa e inteligência de todas as nações. Big data é talvez a maior inovação em computação na última década. Apenas começamos a ver o seu potencial para coletar, organizar e processar dados."

Conforme apontado pelo o ODCA (2012, p.6), 90% de todos os dados existentes no mundo foram criados somente entre os anos de 2010 e 2012, o que demonstra uma forte ligação com a popularização de mídias sociais e *smartphones*.

A Figura 1 reforça o quanto o crescimento da mobilidade por meio dos aparelhos celulares e as redes sociais como *Facebook* e *Twitter* contribuem para a criação de dados na atualidade, e ainda exibe uma projeção para o volume de conteúdo digital existente no ano de 2015, que crescerá motivado pelo ganho de mobilidade dos aparelhos eletrônicos tradicionais.

[2] https://revistaforum.com.br/global/facebook-exclui-paginas-de-consultora-dos-eua-que-difundia-fake-news-para-a-direita-latino-americana/

Figura 1 - Infográfico sobre Big Data



Fonte: Koetsier (2012, *apud* Monetate)

1.2 Conceitos sobre Big Data

O termo Big Data pode ser definido como:

> "Dados que excedem a capacidade de processamento de sistemas de banco de dados convencionais. A quantidade de dados é grande, e são gerados de maneira muito veloz, ou não se enquadram nos padrões da arquitetura de banco de dados atuais." (Reilly, 2012, p.4)

Segundo Hurwitz (2013, p. 16), a capacidade de gerenciar um enorme volume entre diferentes tipos de dados, a uma velocidade e prazo esperados de forma a permitir a análise e resposta em tempo real, é entendida como Big Data.

De acordo com o ODCA (2012, p.5), Big Data refere-se a:

> "Grandes quantidades de dados, dos quais o tamanho e variedade estão além das capacidades de processamento de dados tradicionais para se capturar, gerenciar e analisar em tempo hábil. Big Data vem de todos os lugares e as fontes mais comuns incluem dados gerados por máquina a partir de sensores, dispositivos RFID, arquivos de log, os sinais de GPS de telefone celular, mídias digitais (tanto on-line quanto off-line), sites de mídia social e registros sub-transacionais de transações on-line."

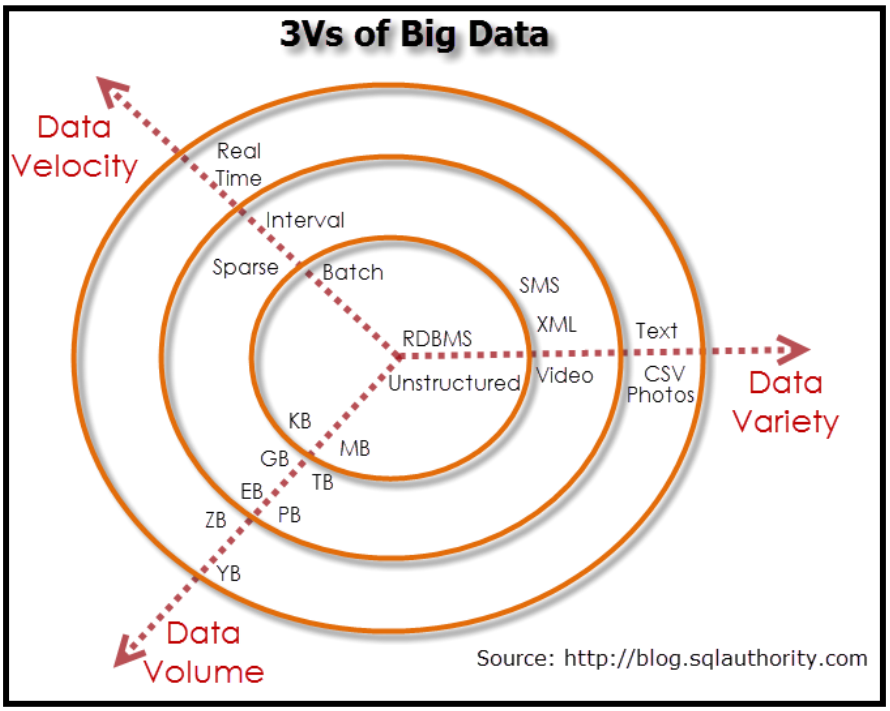
Na visão de Mayer e Cukier (2013, p.14), o termo Big Data é entendido como:

> "coisas que podemos fazer em grande escala que não podem ser feitas em escalas menores, para extrair novas percepções ou criar novas formas de valor, de modo a mudar os mercados, as organizações, a relação entre os cidadãos e os governos, e muito mais."

Em todas as definições sobre Big Data, é possível detectar que existe uma relação entre a grande quantidade de dados (intenso volume), a variedade das diferentes fontes destes dados, e a enorme rapidez (velocidade) com que são produzidos, o que caracteriza o termo Big Data como a combinação dos "*3 Vs*", Volume, Velocidade e Variedade quanto a existência de dados armazenados eletronicamente. (ODCA, 2012 p.5).

De acordo com o ODCA (2012, p.6), Volume refere-se a quantidade de dados gerados em relação ao espaço de armazenamento necessário para comportar a porção de dados existentes, que variam de muitos Terabytes à Petabytes, dependendo do negócio. Velocidade está relacionado ao tempo demandado para se processar e gerar respostas sobre os dados existentes, o que varia entre os diferentes tipos de negócio, pois dependendo do setor, uma utilização dos dados eficiente significa a captação, processamento e resposta em tempo real ao evento/dado gerado. A Variedade se aplica à complexidade dos diferentes tipos e origens de dados, sejam estes estruturados, ou não estruturados. A Figura 2 ilustra os "3 Vs" da perspectiva dos conceitos de Big Data:

Figura 2 - Big Data 3 Vs



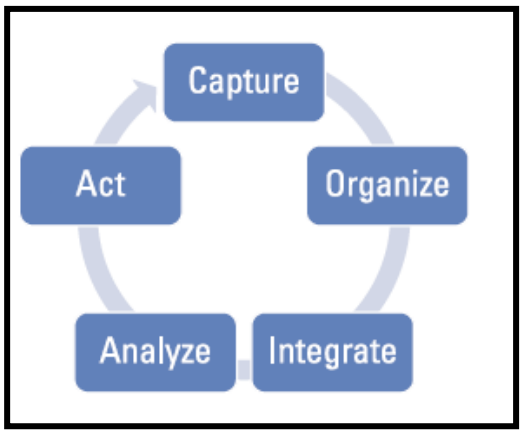
Fonte: Dave (2013)

Como é possível visualizar na figura 2, a velocidade dos dados, conforme ilustrado pela flecha a noroeste que antes eram processados por rotinas agendadas, agora, com os recursos existentes nas soluções orientadas a Big Data, é possível ter respostas em tempo real sobre as diferentes fontes e tipos de dados, inclusive esses não estruturados, ilustrados pela "Variedade de Dados" ilustrada pela flecha a leste, como o conteúdo de documentos, vídeos, imagens, planilhas, comentários em redes sociais, entre outros. A flecha a sudoeste, representa o "Volume de dados", expondo a ordem de grandeza em relação a quantidade de dados eletrônicos existentes e disponíveis no mundo para serem analisados sobre a escala que variava de *Kilobytes*, *Megabytes* e *Gigabytes*, até chegar no *Petabyte*, *Zetabyte* e *Yotabyte*.

1.3 Big Data: arquitetura e tecnologias

Para atingir os objetivos da utilização das tecnologias orientadas a Big Data, Hurwitz (2013, p.16) alerta que é necessário levar em consideração os requisitos funcionais para o Big Data, conforme ilustrado por meio da Figura 3 a seguir:

Figura 3 - Big Data: Ciclo de gerenciamento

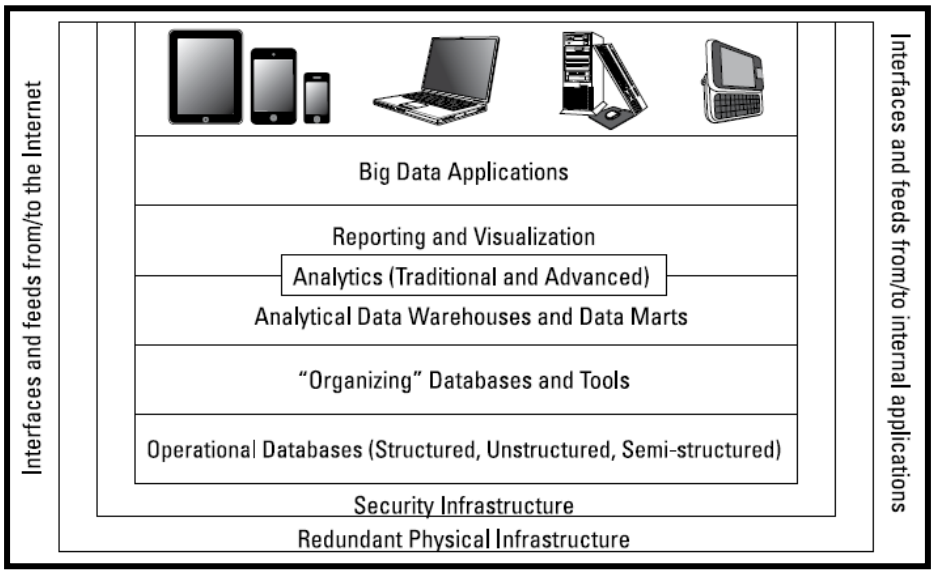


Fonte: Hurwitz (2013, p.17)

A Figura 3 mostra que primeiramente os dados precisam ser capturados, para em seguida serem organizados e integrados. Após a fase de integração, os dados são analisados, de acordo com o escopo do problema definido, e finalmente uma ação é gerada fundamentada na fase de análise. (Hurwitz, 2013, p.17)

Segundo Hurwitz (2013, p.18), a arquitetura tecnológica que envolve todo este fluxo lógico do Big Data inclui uma grande variedade de serviços que possibilitam, de forma efetiva, o uso dos diferentes tipos de fontes de dados. Esta arquitetura é representada pela Figura 4, conforme a seguir:

Figura 4 - Arquitetura Tecnológica do Big Data



Fonte: Hurwitz (2013, p.18)

Nas laterais da figura, estão os *feeds* e interfaces internas (direita) e externas (esquerdas), representado as diferentes fontes de dados internas e externas, e suas múltiplas origens. (Hurwitz, 2013, p.18)

A base do diagrama se refere à uma infraestrutura física redundante, que é indispensável para prover a escalabilidade e disponibilidade necessária para a qualquer sistema orientado ao Big Data. Para fornecer redundância, esta infraestrutura é baseada em computação distribuída, com o mesmo dado replicado em múltiplas

localidades e acessado/consumido por diferentes ferramentas de análise e aplicações Big Data. (Hurwitz, 2013, p.19)

O segundo nível do diagrama representa a Segurança sobre os dados que envolvem a toda infraestrutura física e lógica, englobando aspectos de conformidade de normas de segurança da informação, para garantir controle de acesso e proteção aos dados existentes. (Hurwitz, 2013, p.19)

De acordo com Hurwitz (2013, p.20), o terceiro nível da arquitetura Big Data refere-se às fontes de dados Operacionais, que neste contexto, além de trabalhar com dados altamente estruturados, também precisam armazenar e tratar dados de diferentes tipos de origens, que eventualmente possuem em seu conteúdo dados não estruturados, como imagens, textos publicados em mídias sociais, vídeos, entre outros. Um exemplo de tecnologia aplicada para atender a esta necessidade é o *NoSQL*. Para lidar com Big Data, é necessário possuir tanto base de dados relacionais quanto não relacionais.

O quarto nível do diagrama de uma arquitetura Big Data está relacionado às ferramentas e aos serviços de organização de dados, utilizados para gerenciar o grande volume de dados captados. Segundo Hurwitz (2013, p.22), estas soluções são baseadas em computação distribuída, permitindo o processamento de dados em paralelo, para garantir o processamento de grandes porções de dados a um baixo custo e de maneira eficiente. Exemplos de ferramentas e serviços que garantem esta função são o *MapReduce, Big Table, Hive* e *Apache Hadoop*.

O quinto nível da arquitetura Big Data refere-se aos *data warehouses* e *data marts* analíticos, que armazenam de maneira organizada os diferentes conjuntos de dados para que possam ser acessados de acordo com a demanda do negócio, provendo compressão de dados, vários níveis de particionamento e uma arquitetura de processamento em paralelo robusta. (Hurwitz, 2013, p.22)

O nível intermediário entre o quinto e o sexto nível refere-se à capacidade de análise, que pode ser Tradicional, ou Avançada. A análise tradicional está relacionada à percepções mais simples, que buscam identificar situações anômalas, realizar monitoramento em tempo real, ou simples visualizações, enquanto que a análise avançada envolve o contexto de Big Data com algoritmos mais complexos que incluem modelos estatísticos complexos, aprendizagem de máquina, redes neurais e análise de textos. (Hurwitz, 2013, p.143)

O sexto nível da arquitetura tecnológica que envolve o Big Data está inserido no aspecto de relatórios e visualização, que de acordo com Hurwitz (2013, p.23), "tornam-se ferramentas capazes de analisar e apresentar todas as relações entre os dados captados e processados e o impacto destas relações sobre o futuro".

O sétimo nível do diagrama representa as diferentes a áreas de aplicação para o Big Data, como marketing, indústria, saúde, segurança, serviços de utilidade pública, entre outros. (Hurwitz, 2013, p.23)

O oitavo nível representa exemplos das diferentes formas de dispositivos responsáveis por gerar os dados que serão consumidos, avaliados e armazenados. (Hurwitz, 2013, p.23)

1.4 Onde aplicar o Big Data

De acordo com o ODCA, (2012, p.8) o Big data pode agregar valor para quase todos os segmentos da indústria, otimizando os processos de tomada de decisão e trazendo novas percepções ao negócio. Dentre os segmentos em que é possível aplicar as tecnologias do Big Data, pode-se destacar:

a) As mídias sociais, como Facebook e Twitter;
b) Os mecanismos de busca, como Google e Yahoo;
c) As empresas do ramo de seguros, bancos e operações financeiras;
d) As empresas do setor de telecomunicações, marketing, provedores de Internet e provedores de serviços móveis;
e) As empresas do setor varejista e análise dos pontos de vendas;
f) Na otimização do processo de manufatura;
g) Nas indústrias de energia;
h) Na área da saúde;
i) Nas operações de TI;
j) Em pesquisa e desenvolvimento;
k) Em áreas de interesse de governos, como segurança pública e transporte.

Vamos discutir a frente os cuidados necessários para a proteção das tecnologias Big Data e sua aplicação em contextos relacionados à computação forense, envolvendo aspectos como fraudes corporativas, sua utilização em processos investigativos e como mecanismo de defesa.

2 A PROTEÇÃO DO BIG DATA

Iremos discutir nessa secção os requisitos existentes para assegurar a proteção de dados trafegados ou armazenados em todos os componentes de uma infraestrutura Big Data, bem como expor as implicações sobre a privacidade e confidencialidade dos dados utilizados.

2.1 Implicações de segurança e privacidade

Conforme citado por Savitz (2013), estima-se que até o ano de 2020, cerca de 200 bilhões de objetos estarão conectados à Internet, e que a possibilidade de se extrair significados dos dados gerados por estes objetos poderá, dentre muitas coisas, ajudar a diminuir a criminalidade, evitar acidentes de trânsito, economizar energia e evitar desperdícios, dentre outros.

As empresas que por ventura adotarem soluções Big Data assumirão um papel de destaque no âmbito organizacional, pois estas são se prepararam para coletar, armazenar, correlacionar, interpretar e apresentar resultados valiosos sobre todo o dado capturado. Porém, de acordo com Hurwitz (2013, p.228), as fontes de dados do Big Data podem não ser totalmente seguras, protegidas, além de conter diversos tipos de dados pessoais, que podem expor estas empresas.

Hurwitz (2013, p.230), afirma que "as empresas devem se certificar de que as novas fontes de Big Data não as exponham a ameaças não antecipadas, ou riscos de governança".

Segundo Tankard (2012), empresas que centralizarem os dados num único lugar poderão despertar a atenção de atacantes, tornando este cenário um alvo valioso às pessoas que tenha interesse nesses assuntos; que por sua vez terão grandes quantidades de informações expostas. Tankard (2012) afirma ainda que é importante que o armazenamento destes dados seja controlado e protegido adequadamente.

Ainda de acordo com Tankard (2012), é importante estar atento a normas e padrões regulatórios de países e setores envolvidos, e em especial às leis de proteção de dados, e alerta para que as empresas que se utilizam de sistemas Big Data façam uma minuciosa classificação das informações, e apliquem controles apropriados, como a imposição de períodos de retenção de dados e descarte adequado de informações, para que estejam sempre em conformidade com as leis e normas existentes.

Segundo Hurwitz (2013, p.227), é preciso ter cuidado durante o consumo/coleta de dados não estruturados, como no caso de mídias sociais, onde é preciso certificar-se de que não existam links maliciosos, ou vírus em seu conteúdo, pois isto pode trazer grandes riscos à organização. Os dados de mídias sociais devem ser consumidos de fontes confiáveis, que fazem o adequado monitoramento do comportamento dos usuários, a fim de prevenir que estes não propaguem conteúdo malicioso.

De acordo com o CSA (2013), a necessidade de preocupação com a segurança e privacidade dos dados em um ambiente Big Data se torna crítica, por conta da velocidade, variedade e volume dos dados, aliada à grande diversidade de tipos de fontes de dados. O massivo uso de computação em nuvem, aplicado a diferentes plataformas de software, acaba propagando os dados existentes em diversas redes de computadores, o que também acaba aumentando a área de risco de ataques e vulnerabilidades de segurança em todo o sistema.

O grande desafio da proteção de dados numa infraestrutura que lida com Big Data é que os controles de acesso à informação devem ser aplicados de acordo com o dado em si, e não somente aos sistemas e aplicações que o armazenam. Uma das formas de se atingir este controle é protegendo os dados mais críticos e sensíveis, tornando-os inelegíveis, por meio de técnicas já bastante utilizadas comumente, como por exemplo, a criptografia (Tankard, 2012).

Apesar de a criptografia ser uma técnica bastante utilizada para garantir a proteção dos dados, ela não garante que estes sejam invioláveis, pois os dados podem ser interceptados antes de aplicá-la. A criptografia dos dados acaba elevando o tempo e o custo para o processamento das informações, e em alguns casos pode não ser a técnica mais eficiente, conforme mencionado por (Hurwitz, 2013, p.228).

De acordo com Hurwitz, (2013, p.228), além da criptografia, outras técnicas podem ser aplicadas, como a *anonimização* de dados, que ajudam a não associar determinados dados sensíveis, como número de cartão de crédito, dados bancários, nome completo ou documentos a indivíduos, e a *tokenização*, que por meio de dispositivos denominados *token* ajuda a mascarar os dados sensíveis existentes, de modo a evitar que pessoas não autorizadas (que não possuem o *token*) consigam visualizar o dado real.

2.2 O controle das informações em ambientes Big Data

De acordo com a Oracle (2011), a "Governança de Dados é responsável por estabelecer as especificações desejáveis sobre a avaliação, criação, armazenamento, uso, arquivamento, e descarte de dados e informações." A Governança de dados atua também no controle de processos, padrões e métricas para que a organização atinja seus resultados, assegurando o uso efetivo e eficiente das informações e dados existentes.

Uma governança efetiva de dados e informações não somente assegura o uso eficiente das informações, como também é importante para mitigar riscos, conforme pontua Murphy (2012).

No contexto de Big Data, é muito importante que as empresas tenham controle acerca dos dados consumidos dentre as diversas fontes de dados existentes, tendo seus processos de segurança da informação, privacidade e governança estendidas para os sistemas Big Data, de maneira a considerar os seguintes aspectos conforme destacados por Hurwitz (2013, p.229):

a) Determinar quem poderá acessar as novas fontes de dados antes e depois de serem devidamente analisadas e entendidas;
b) Entender como estes dados serão segregados dos dados de outras empresas;
c) Se os dados consumidos forem adquiridos de outras empresas, a utilização deverá estar aderente aos termos de uso e regras existentes em contrato;
d) Verificar onde o dado será fisicamente armazenado, pois existem países que possuem regras de privacidade mais restritivas, que necessariamente deverão ser atendidas para evitar multas ou punições;
e) Em caso de armazenamento em nuvem, verificar os termos oferecidos em contrato, bem como as regras de privacidade do local onde o dado estará armazenado, e as políticas e níveis de segurança estabelecidos pelo fornecedor do serviço.

É importante ressaltar que antes de se definir as políticas de segurança e governança da informação a empresa precisa saber que tipo de dados ela irá utilizar, processar e analisar, pois se os controles e processos de governança e segurança da informação não estiverem aderentes às leis e normas dos países de atuação dessas empresas, e segmentos de onde são adquiridos e aplicados os dados, já que o risco de causar sérios danos ao negócio é iminente, como por exemplo, a perda de confiança e credibilidade por parte de clientes e fornecedores, ou até mesmo punições e multas por parte dos Governos.

Ainda neste sentido, é importante que as políticas e normas de governança e segurança da informação sejam sempre monitoradas, de forma a garantir que todas as permissões de acesso sejam sempre revisadas, atualizadas e documentadas, pois desta forma poderão facilitar o trabalho de auditorias internas e externas, ou até mesmo em casos de litígio, onde se faz necessária uma perícia forense.

2.3 Os principais desafios de segurança e privacidade em ambientes Big Data

A infraestrutura de sistemas Big Data é baseada em computação distribuída e banco de dados não relacionais. Neste sentido, é importante garantir a segurança de toda esta infraestrutura, para assegurar que possíveis ataques não tenham acesso a informações sensíveis tanto para o negócio, quanto à privacidade dos dados. A proteção adequada dos componentes de infraestrutura ajuda a minimizar os riscos externos, porém é importante criar controles internos que permitam que apenas os indivíduos confiáveis possuam acesso às informações mais restritas.

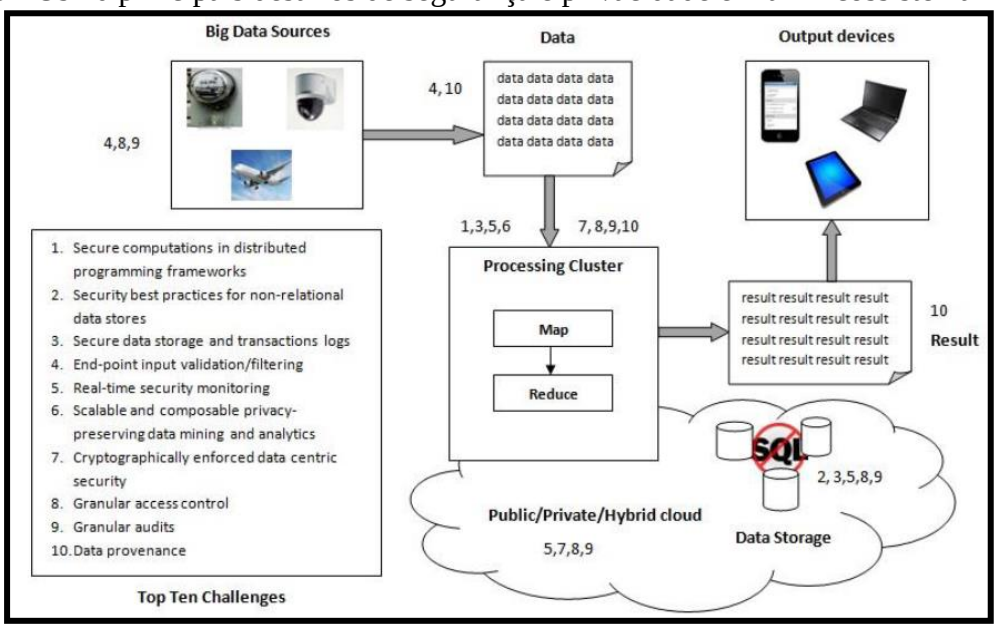
Vale ressaltar também a importância de se aplicar controles e filtros adequados nas fontes de dados consumidas pelo Big Data, para que estas não direcionem a um produto final impreciso e não confiável.

Estes pontos ressaltados são ratificados pelo CSA (2013, p.7):

> "Para garantir a segurança da infraestrutura de sistemas Big Data, os componentes de computação distribuída e o armazenamento de dados devem ser protegidos. Para proteger os dados em si, a difusão da informação deve ser preservada de acordo com a expectativa de privacidade necessária, e os dados mais sensíveis devem ser protegidos por meio do uso de criptografia e controles granulares de acesso. Gerenciar um enorme volume de dados exige soluções escaláveis e distribuídas, tanto para garantir a segurança dos depósitos de dados, quanto para permitir auditorias eficientes. Dados de fluxo contínuo que emergem de diversas fontes devem ser verificados quanto à integridade, e podem ser usados para realizar análises em tempo real em incidentes de segurança para assegurar a saúde da infraestrutura."

Para garantir que um ecossistema Big Data esteja adequadamente protegido contra ameaças externas e pronto a satisfazer requisitos de leis e normas regulatórias, o CSA (2013, p.6) aponta algumas questões específicas, classificando estas questões como os 10 principais desafios de como garantir a Segurança e a Privacidade em ambientes Big Data, conforme ilustrado na Figura 5:

Figura 5 - Os 10 principais desafios de segurança e privacidade em um Ecossistema Big Data



Fonte: CSA (2013, p.6)

A Figura 5 mostra o fluxo lógico dos dados em Sistemas Big Data desde sua origem, entrada, processamento, armazenamento e saída, e ainda situa os pontos críticos entre o fluxo dos dados em relação aos 10 desafios de segurança e privacidade, que serão detalhados a seguir.

(1). Processamento seguro em estruturas de computação distribuída

Pelo contexto da Figura 5, é importante garantir o processamento seguro em estruturas de computação distribuída, o que significa que os membros do *cluster* de processamento do ecossistema Big Data necessitam de mecanismos e controles de segurança que os tornem íntegros e confiáveis.

Conforme citado no item 2.3, uma das tecnologias mais aplicadas em sistemas Big Data é o *MapReduce*, que se baseia em computação distribuída para realizar a "divisão" dos dados e efetuar o processamento em paralelo. O CSA (2013, p.9) afirma que é necessário garantir a confiabilidade dos *Mappers* (nós dos *clusters*

envolvidos no processamento dos dados) e a proteção os dados em si, independente da existência de *Mappers* não confiáveis.

Para que um *Mapper* possa ser classificado como "confiável", é preciso implementar controles de acesso que garantam a visibilidade apenas de arquivos autorizados, bem como implementar mecanismos de autenticação periódicas com o nó "master", para garantir que estes componentes estejam sempre em conformidade com as políticas de segurança estabelecidas. (CSA, 2013, p.9)

(2). Boas práticas de Segurança para Bancos de Dados não Relacionais

Os bancos de dados não relacionais, como o *NoSQL* fazem parte de uma infraestrutura Big Data, muito por conta de sua característica de lidar com diversos tipos de dados estruturados e não estruturados de maneira escalável e com performance adequada para processamento em tempo real dos dados.

Entretanto, conforme Denoncourt (2012), os sistemas de banco de dados não relacionais foram desenvolvidos justamente para atender a desafios de performance e escalabilidade dos bancos de dados tradicionais, tendo seus mecanismos nativas de segurança pouco explorados e desenvolvidos.

Neste sentido, como existem vulnerabilidades de segurança conhecidas em bases de dados *NoSQL*, é importante que mecanismos que garantam a integridade dos dados sejam implementados na camada de aplicação. Os dados armazenados nestes bancos de dados, devem de preferência, estar criptografados. Outro ponto importante a se considerar para é a utilização de conexão SSL/TLS entre clientes/servidores e as aplicações e o banco de dados, para garantir a confidencialidade dos dados em transição. (Csa, 2013, p.12)

É importante ainda manter ativos os serviços de criação de *logs* para auditoria destas bases, pois por meio destes mecanismos, é possível identificar possíveis atividades maliciosas, ou acessos indevidos.

A utilização de criptografia nos dados é importante para garantir a privacidade, porém nota-se que dependendo do nível de criptografia utilizado e a quantidade de transações ativas, a criptografia pode gerar impactos negativos sobre a performance das bases de dados não relacionais.

(3). Proteção dos logs transacionais e armazenamento de dados.

Em sistemas Big Data, devido ao grande volume de dados existentes, usualmente os dados e *logs* transacionais são armazenados em mídias de armazenamento multicamadas, que efetuam o armazenamento de dados, de acordo com a frequência de uso destas informações. Esta prática contribui com a redução de custos de armazenamento, pois as informações menos utilizadas acabam sendo alocadas fisicamente em mecanismos de mais baixo custo de armazenamento, como por exemplo, o uso de serviços de computação em nuvem, o que na visão do CSA (2013, p.14) pode representar uma ameaça, pois a tendência é que quanto mais baixo o custo de armazenamento, maiores as chances das políticas de segurança e proteção aos dados não estarem devidamente aplicadas, ou até mesmo não existirem.

Neste sentido, é importante ressaltar mais uma vez que dependendo da frequência do uso de dados mais críticos, que conforme visto no item 3.1 necessitam de políticas de segurança e controles de acesso mais rigorosos, onde estes podem ser armazenados nestas camadas "inferiores", que podem não apresentar níveis de confidencialidade e integridade adequados e acabar expondo a empresa a riscos desnecessários.

Destaca-se também a relevância do adequado armazenamento dos arquivos de *log* quanto à disponibilidade, integridade e períodos de retenção, importantíssimos em auditorias e eventuais incidentes de segurança, ou em casos litigiosos, que demandem uma análise forense. Restringir o acesso e garantir a proteção dos arquivos de *log* também é importante para garantir que estes não sejam manipulados ou forjados intencionalmente.

(4). Aplicação de filtros e validação dos dispositivos de entrada.

De acordo com o CSA (2013, p.17), é importante que as empresas considerem monitorar os dispositivos de entrada, a fim de garantir que as fontes de dados sejam confiáveis, ou que ao menos seja possível filtrar dados irrelevantes, ou maliciosos.

Em relação aos dispositivos de entrada, o CSA (2013, p.17) aponta como ameaça as possíveis formas de ataque, que variam entre o comprometimento do dispositivo em si, como a instalação de algum *malware*, capaz de gerar dados incorretos, a clonagem destes dispositivos, para que estes gerem uma série de dados falsos, ou até mesmo a manipulação do ambiente de coleta destes dispositivos, como por exemplo, o aquecimento ou resfriamento proposital no caso de sensores térmicos utilizados na análise preditiva de temperatura.

Como resposta a estas ameaças, o CSA (2013, p.17) pontua que é necessário aplicar algumas técnicas no desenvolvimento destes dispositivos que garantam a autenticidade e integridade dos dados. A utilização de autenticação e certificação digital entre estes dispositivos e os sistemas que efetuam o transporte e coleta dos dados também pode ser uma forma de mitigar estes riscos expostos.

Na Figura 5, é possível identificar que este fator não está somente atrelado aos dispositivos de entrada em si, como também durante o transporte das informações geradas por eles até os sistemas onde os dados serão processados.

Neste sentido, é importante garantir que estes dados sejam transmitidos de maneira segura, para que nenhum atacante tenha acesso às informações trafegadas, e muito menos tenha a possibilidade de injetar dados maliciosos, que modifiquem o produto da análise dos dados verdadeiros.

(5). Monitoramento de Segurança em tempo real.

O monitoramento em tempo real numa infraestrutura Big Data possui dois aspectos importantes. O primeiro é garantir o funcionamento adequado de toda a infraestrutura Big Data, como por exemplo, verificar em tempo real a saúde e desempenho dos nós de processamento, conforme pode ser observado na Figura 5, onde este aspecto está presente durante o processamento e armazenagem do dado em si. O segundo é monitorar o ambiente para detectar possíveis vulnerabilidades, como por exemplo, a existência de dispositivos de entrada maliciosos, que possam minar, ou comprometer toda a funcionalidade analítica do ambiente, utilizando as próprias ferramentas de Big Data para o monitoramento em tempo real e geração de alertas e incidentes que possam ser considerados como ameaças.

Entretanto, conforme a visão do CSA (2013, p.20), "monitorar um ambiente Big Data envolve questões técnicas, éticas e legais", pois as ferramentas de monitoramento existentes ainda estão em franca expansão e desenvolvimento, não possuindo um alto grau de maturidade. Como questões legais e éticas, é importante destacar que as informações mais restritas dificilmente poderão ter seu conteúdo monitorado.

É importante destacar que o Big Data pode ser aplicado como ferramenta de segurança e monitoramento não somente para garantir a segurança e disponibilidade de seu próprio ecossistema, como também para auxiliar empresas de diversos segmentos a identificar incidentes de segurança, ou até mesmo fraudes em tempo real.

(6). Preservação da Privacidade de forma escalável e modular na mineração de dados em sistemas analíticos

Conforme cita o CSA (2013, p.23), para proteger a privacidade dos indivíduos, boas práticas de prevenção e detecção de abusos e violação de privacidade devem ser implementadas e continuamente monitoradas nos sistemas analíticos e de mineração de dados.

Em relação à elaboração das práticas de prevenção e detecção de violação de privacidade, é relevante considerar fatores internos e externos à organização para criar os mecanismos de monitoramento, já que funcionários mal intencionados podem "quebrar" os controles existentes e obter acesso a informações sensíveis.

Para dificultar que possíveis atacantes (externos) possam violar a privacidade de usuários, é importante manter os componentes de infraestrutura sempre atualizados, em conformidade com os últimos recursos existentes no que diz respeito a segurança. Técnicas como a criptografia, aplicação de controles de acesso e mecanismos de autenticação também devem ser implementadas para diminui riscos tanto externos quanto internos, além da criação e preservação de *logs* de acesso aos dados, que podem ser analisados para identificar o que eventualmente possa ter sido violado, bem como a origem do acesso.

(7). Criptografia aplicada ao controle de acesso e comunicação segura.

Atualmente, existem duas formas distintas de se controlar a visibilidade dos dados entre diferentes tipos de sistemas, indivíduos, ou organizações. A primeira delas é a limitação de acesso ao sistema operacional em si, por meio da criação de identidades específicas, com um baixo grau de privilégios de acesso utilizados para a comunicação dos sistemas em execução. A segunda maneira comumente adotada para restringir o acesso aos dados entre usuários e sistemas de arquivos é a utilização de criptografia. (CSA, 2013, p.25)

De acordo com o CSA (2013, p.25), a utilização do acesso controlado por privilégios de sistemas tem sido a forma mais aplicada para restringir a visibilidade dos dados, mas esta técnica possui um vasto histórico de ameaças constantemente exploradas por atacantes, como ataques de *buffer overflow* e de escalação de privilégios, onde determinadas vulnerabilidades de sistemas operacionais e aplicações em uso podem ser exploradas.

Conforme visto no item 6, é possível minimizar estas vulnerabilidades, mantendo os sistemas sempre atualizados com as últimas implementações de segurança, mas ainda assim esta forma de proteção é a que oferece mais riscos à visibilidade dos dados.

A criptografia, entretanto, oferece um grau de confiabilidade maior, pois ainda que cientificamente provado que é possível efetuar a quebra de qualquer mecanismo de criptografia, este se torna um processo muito mais sofisticado, o que demandaria um ataque muito mais direcionado, exigindo conhecimento e tempo muito maior até obter êxito na quebra da criptografia e visibilidade aos dados.

De acordo com o CSA (2013, p.27), existem pesquisas com o objetivo de otimizar as técnicas de criptografia atuais, de forma a implementar algoritmos que se baseiam em assinatura de grupos ou identidades para dificultar ainda mais a ação de possíveis atacantes. Porém, como é citado também pelo CSA (2013, p.28) estas técnicas em desenvolvimento ainda apresentam altos custos computacionais, o que no momento não traz tanta viabilidade de implementação, dependendo dos recursos existentes.

(8). Controle de acesso granular.

A Figura 5 mostra que é importante manter o controle de acesso granular aos dados desde sua entrada nos sistemas big data, através das diferentes fontes de dados, passando pela fase de processamento, e em seus diferentes meios de armazenamento.

De acordo com o CSA (2013, p.30), uma maneira de melhorar a eficiência do controle de acesso granular é aplicá-los na camada de infraestrutura, minimizando assim alguns ajustes em determinadas aplicações. Outras questões abordadas nos itens anteriores, como a garantia da segurança da infraestrutura, por exemplo, tornam-se pré-requisito para um eficaz controle de dados granular.

(9). Auditorias granulares.

Em ambientes Big Data, a existência de "falsos positivos" no que diz respeito ao monitoramento de segurança em tempo real, conforme visto no item 5 deste mesmo capítulo, existem e podem até ser consideradas frequentes. Porém, conforme cita o CSA (2013, p.31) também existem os incidentes considerados "verdadeiros positivos", ameaças, incidentes de segurança ou ataques, que acabam não sendo detectados em tempo real, e é neste contexto que as auditorias granulares acabam sendo importantes, já que estas são cruciais para o entendimento dos eventos ocorridos no ambiente Big Data, além de ajudarem no monitoramento e verificação de conformidade com leis e normas, e em possíveis investigações forenses.

De acordo com o CSA (2013, p.32), o produto de entrada das auditorias, como *logs* de aplicação, *logs* de sistemas operacionais, ou até mesmo o tráfego de rede entre os sistemas e componentes da infraestrutura Big Data devem ser analisados, através de ferramentas forenses, ou de *SIEM - Security Information and Event Management* - que são responsáveis pela coleta, correlação e análise de todos os eventos produzidos nos componentes existentes.

(10). Proveniência de dados.

O termo proveniência de dados é definido por Buneman (2000, p.1), como o "processo de monitorar e registrar as origens dos dados durante todo o seu percurso entre bases de dados".

Na visão do CSA (2013, p.34), a proveniência de dados é considerada fundamental para os processos de auditoria, para garantir a confiança e a detecção de falhas em aplicações Big Data. Neste sentido, é importante ressaltar que os dados de proveniência, principalmente aqueles atrelados aos dados mais sensíveis que requerem um alto grau de confidencialidade, necessitam de proteção através de criptografia e de controles de acesso granular, pois novamente a exposição destes dados pode acabar trazendo riscos às normas de privacidade atribuídas aos dados existentes.

3 APLICANDO O BIG DATA COMO MECANISMO DE DEFESA E ARTEFATO DE INVESTIGAÇÃO

Nesta seção serão abordadas possíveis aplicações práticas para o Big Data, sobre como as tecnologias baseadas nestes princípios de processamento de grandes volumes de dados dinâmicos e distintos podem ser utilizadas como mecanismos de defesa, combate a fraudes corporativas, mapeamento e identificação de ameaças e vulnerabilidades em sistemas de informação.

3.1 Detecção de fraudes

Atualmente, pesquisas apontam que, no mundo, cerca de U$ 3,5 trilhões em receitas são perdidas pela ocorrência de fraudes (Griffin, 2012em Acfe).

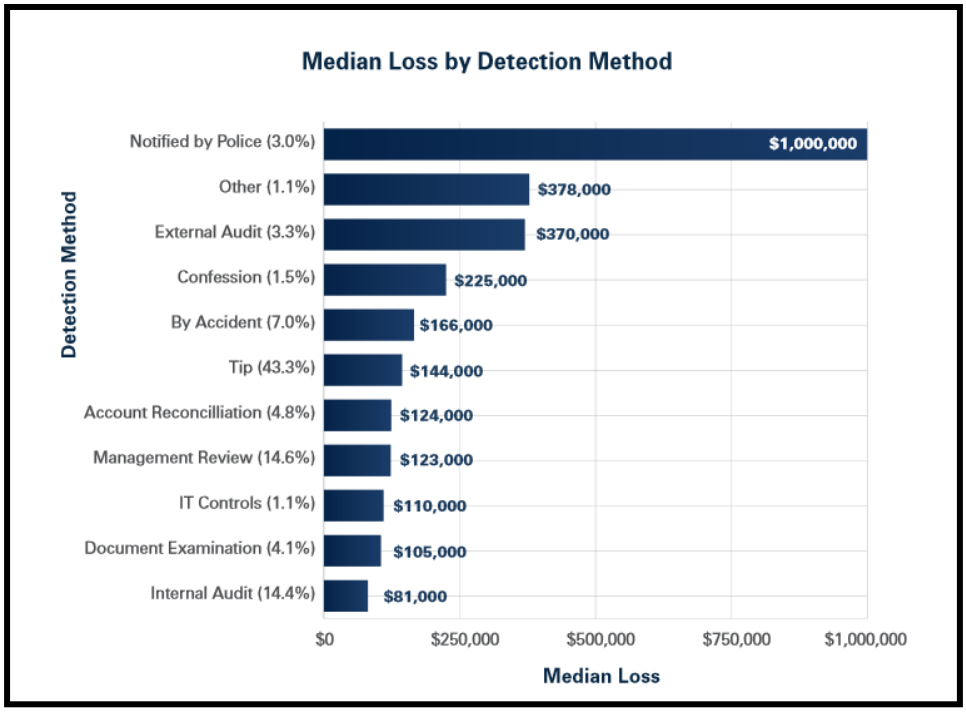
Este dado é preocupante, pois a quantia que se perde é tão grande que muitas empresas, principalmente do setor financeiro, bancário e de seguros, as mais afetadas, entendem que as fraudes deixaram de ser consideradas como apenas um custo do negócio, e passaram investir fortemente no combate a elas.

De acordo com o ACFE (2012, p.15), uma pesquisa mundial revelou que fraudes descobertas como resultado de denúncias à polícia são aquelas que apresentam o maior custo em perdas. Nestes casos, o custo total da fraude atinge em média cerca de U$1.000.000,00 (um milhão de dólares), enquanto que fraudes detectadas por auditorias externas, geram prejuízos de cerca de U$370.000,00 (trezentos e setenta mil dólares).

Esta mesma pesquisa revelou ainda que a maneira mais eficiente de detecção, no sentido de perdas com fraudes, é por meio de auditorias internas, o que mesmo assim ainda traria um prejuízo médio de cerca de U$81.000,00 (oitenta e um mil dólares).

Entretanto, como pode ser visualizado na Figura 6, em média apenas 14,4% das fraudes são detectadas por auditorias internas, e o método mais comum de se tomar conhecimento da existência de uma fraude é por meio de denúncias internas.

Figura 6 - Custo médio de perdas com fraudes de acordo com seu método de detecção.



Fonte: ACFE (2012, p.15)

Neste sentido, torna-se evidente o alto risco existente no que diz respeito a descoberta de fraudes corporativas, já que, conforme a pesquisa, a maneira mais comum de se tomar conhecimento de uma fraude é por meio de algum tipo de delação por parte de um colaborador, representando 43.3% das ocorrências.

Nota-se também que a detecção de fraudes é um processo lento e reativo, já que este mesmo estudo revela que, em média, uma fraude leva em torno de 18 meses até ser descoberta (ACFE 2012, p.13).

O processo de investigação de fraudes é baseado na análise de dados forenses. Atualmente, é mister atestar que praticamente todos dados possuem alguma forma de representação digital e estão armazenados em dispositivos eletrônicos.

Assumindo que estes dados estruturados, ou não estruturados, existam em dispositivos eletrônicos, estes podem ser utilizados por ferramentas tecnológicas capazes de analisar seu conteúdo e transformá-los em informações valiosas para mapeamento de riscos, em processos investigativos, ou até mesmo para antecipar possíveis desvios de condutas e evitar que estes se tornem grandiosos esquemas de fraudes e corrupção.

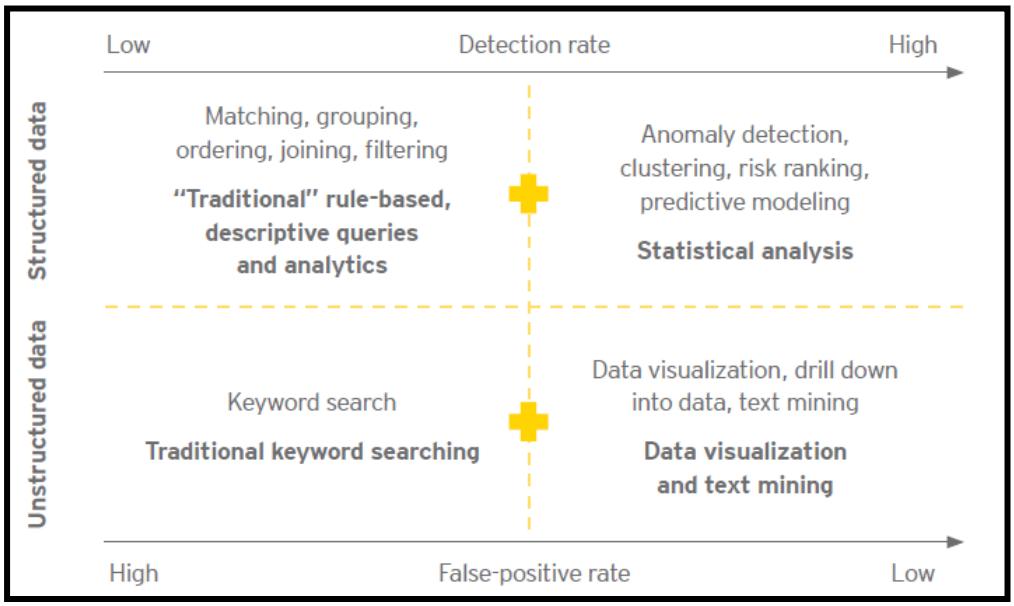
No contexto de análise de dados voltada para a detecção e prevenção de fraudes, o GFDAS (2014, p.6) define o termo *Forensics Data Analytics* como:

> "a habilidade de coletar e utilizar dados estruturados e não estruturados para identificar potenciais áreas de fraude ou corrupção, como operações indevidas, descumprimento de leis ou políticas da empresa, ou padrões anômalos e tendências nos dados."

Neste sentido, muitas empresas entendem que as emergentes tecnologias orientadas ao Big Data terão um papel importantíssimo no que diz respeito à detecção e prevenção de fraudes corporativas, conforme revelam as pesquisas realizadas pelo GFDAS (2014, p.2).

Baseando-se na arquitetura de um ecossistema Big Data, é possível notar que a utilização de soluções voltadas para este tipo de tecnologia podem contribuir para a otimização dos resultados obtidos nas análises de dados forenses, conforme o modelo de maturidade da capacidade de análise de dados forenses proposto na Figura 7:

Figura 7 - Modelo de Maturidade sobre a capacidade de análise de dados forenses



Fonte: GFDAS (2014, p.9)

Conforme pode ser observado na Figura 7, as características que mais se aproximam dos conceitos, aplicações e funcionalidades de sistemas orientados ao Big Data são aquelas que se localizam no quadrante superior à direita (Análise Estatística - Detecção de anomalias, classificação de riscos e modelos preditivos) e no quadrante inferior à direita (Visualização de dados e mineração de textos). Atingir estes pontos durante uma atividade de monitoramento, ou durante uma auditoria significam ganhos de eficiência e de precisão muito maiores do que em relação às práticas mais utilizadas atualmente, que mais se aproximam dos quadrantes do lado esquerdo, já que a quantidade de falso-positivos será muito menor, dada a grande quantidade de informações potencialmente cruzadas e analisadas pelas soluções Big Data.

Nota-se que além da capacidade de processamento em tempo real, os recursos oferecidos pela arquitetura tecnológica dos sistemas Big Data, voltadas para a análise de dados forense, também podem ser utilizados para processar informações específicas, e, ainda que de forma reativa, contribuir com a otimização do tempo gasto de determinadas tarefas computacionais. (GFDAS, 2014).

No universo corporativo, seja este o de uma organização pública, ou privada, atualmente poucas entidades utilizam todo o seu acervo de dados disponível para identificar possíveis comportamentos anômalos para um efetivo mapeamento e classificação de riscos para a ocorrência de fraudes, utilizando apenas alguns dados por amostragem em suas auditorias internas, o que de certa forma, ainda que de forma reativa, não deixa de ser efetivo. (GFDAS, 2014, p.23)

De acordo com o GFDAS (2014, p.16) as empresas pioneiras na adoção de soluções orientadas ao Big Data para prevenção de fraudes, ou com o objetivo de cumprir requisitos de conformidade com leis ou normas regulatórias, vêm atingindo resultados significativos quanto ao tempo gasto para se processar os dados existentes, reduzindo a complexidade do processo de análise e contribuindo com a agilidade na mitigação de riscos e precisão na tomada de decisões.

Esta mesma pesquisa revela ainda que, no momento, pouquíssimas empresas no mercado estão se utilizando da capacidade analítica e preditiva do Big Data para esta finalidade. (GFDAS, 2014, p.2)

Nesse sentido, ressalta-se que as tecnologias capazes de lidar com volumosos e variados tipos de dados certamente serão massivamente aplicadas no auxílio de prevenção e detecção de fraudes e desvio de conduta, que se detectados e resolvidos a tempo, além de contribuir com a credibilidade e transparência das empresas, poderão evitar severas multas e punições pelo descumprimento de leis, como por exemplo, o FCPA - *Foreign Corrupt*

Practices Act, lei de esfera civil e criminal em vigor nos EUA, que visa regular a ética dos negócios de empresas americanas, ou com ações nas bolsas de valores norte americanas em esfera mundial, a fim de evitar que estas efetuem pagamentos a membros ou entidades de governos estrangeiros em troca de vantagens comerciais ou econômicas.

Apenas nos primeiros meses de 2014, grandes empresas como a HP - *Hewlet Packard* e Alcoa, foram multadas em cerca de $108 milhões e $384 milhões, respectivamente, por violarem o FCPA, conforme mencionado pelo SEC - *Securities and Exchange Commission*, órgão do governo norte americano responsável por fiscalizar se as empresas estão em conformidade com leis como o FCPA. (Sec, 2014).

Nesse contexto, é importante ressaltar que o SEC tem investido pesado em tecnologias orientadas ao Big Data para otimizar seus resultados ao monitorar diversos tipos de transações, podendo também confrontar os dados pessoais dos envolvidos nestas transações analisadas, com o objetivo de detectar a existência de operações suspeitas de modo mais ágil, eficiente e eficaz. (Flitter e Lynch, 2014)

Além de instituições governamentais como o SEC, empresas do setor bancário e de meios de pagamento também estão buscando aprimorar a eficiência de seus sistemas de detecção de fraudes.

METODOLOGIA

Esse estudo foi desenvolvido por meio de uma análise descritiva acerca das principais publicações realizadas sobre o tema Big Data, baseado na segurança e privacidade, com o objetivo de conhecer e analisar as diferentes abordagens sobre o tema.

O método de pesquisa empregado foi o qualitativo, que busca descrever a complexidade de determinada situação e compreender os aspectos que envolvem o assunto abordado, por meio da imersão do pesquisador no contexto da pesquisa, buscando o contato direto com o tema explorado. (Moreira, 2002)

A escolha por este método de pesquisa é mais aderente a essa discussão, pois foca em aspectos da realidade, e busca mostrar os significados, motivos e expectativas do assunto para a propor as respostas cabíveis aos problemas e objetivos definidos.

A SEGURANÇA PÚBLICA E O BIG DATA

A utilização de recursos tecnológicos avançados aplicados nos departamentos de inteligência de forças armadas, polícias, e governos não é nenhuma novidade. Porém, a quantidade, velocidade e variedade de dados criados nos últimos anos acaba se tornando um desafio para que estas entidades consigam captar, processar e gerar resultados proveitosos a partir destes dados.

Neste aspecto, visando a extração de percepções úteis, por meio de mecanismos analíticos avançados sobre variados, volumosos e complexos fluxos de dados, certamente os sistemas capazes de lidar com Big Data também podem, devem e já são aplicados no segmento de segurança pública, conforme publicado por Prox (2013), que relata os resultados obtidos pelo departamento de polícia de Vancouver após a adoção de um novo sistema analítico, o CRIME - *Consolidated Records Intelligence Mining Environment*, capaz de analisar, unificar e apresentar resultados baseados no histórico de ocorrências correlacionado com dados de sistemas de informação geográficas, permitindo o mapeamento de crimes ao longo do tempo e um melhor direcionamento das equipes de policiamento para os locais de maior probabilidade para a ocorrência de crimes.

Com a adoção deste sistema de inteligência orientado ao Big Data, o departamento de polícia de Vancouver obteve uma redução de 24% sobre a ocorrência de crimes contra o patrimônio e de 9% da ocorrência de crimes envolvendo violência, provando assim a eficácia do sistema de inteligência adotado para a prevenção e coibição deste tipo de violência.

A Agência Nacional de Segurança dos Estados Unidos, popularmente conhecida como NSA, certamente está entre as organizações que mais se utilizam de sistemas capazes de lidar com Big Data.

Em 2011, a NSA decidiu tornar seu sistema de armazenamento e indexação de dados para lidar com Big Data, baseado em *NoSQL* e denominado *Accumulo*, um projeto de código aberto, que foi incorporado pela Apache. (Regalado, 2014)

Em 2013, com o vazamento de informações confidenciais da NSA, expostos ao mundo por Edward Snowden (ex funcionário da NSA), foi publicada uma apresentação do ano de 2008 que revelava o uso do sistema denominado XKeyscore, capaz de coletar quase toda a atividade na internet de um determinado indivíduo, como emails ou sites visitados, e até mesmo interceptar ligações telefônicas. O sistema de inteligência e espionagem espalhado em localizações estratégicas no mundo possui como objetivo monitorar comportamentos suspeitos, a fim de detectar possíveis terroristas, e conforme a apresentação divulgada por Snowden, cerca de 300 terroristas haviam sido presos até o ano de 2008 com a ajuda deste sistema. (Grego, 2013)

De acordo com estes casos de aplicações beneficiados pela capacidade analítica de sistemas capazes de lidar com Big Data, utilizados como mecanismos de defesa, investigação e detecção de fraudes corporativas e mapeamento de riscos em sistemas de informação, é importante ressaltar que estas tecnologias, ainda que bastante eficazes, podem ser consideradas relativamente prematuras.

CONSIDERAÇÕES FINAIS

Por meio desta pesquisa, foi possível obter uma visão clara e abrangente sobre um dos temas que mais vêm sendo discutidos por parte da alta gestão no universo corporativo.

É possível afirmar que a extração de valor e novas percepções a partir de dados existentes, porém pouco ou não explorados, pode ser considerada um dos grandes benefícios oferecidos pelas tecnologias capazes de lidar com o Big Data.

Constatou-se que a oportunidade de obter percepções úteis por meio da amplitude e otimização da capacidade analítica sobre os dados existentes, de modo geral, pode trazer grandes avanços e melhorias para as pessoas, já que segmentos como saúde, segurança e gestão pública afetam diretamente a vida de qualquer pessoa.

Como vimos, é possível afirmar que aplicar as novas tecnologias de processamento capazes de lidar com volumosos e diferentes tipos de dados, de modo a efetuar uma completa e profunda análise entre o histórico de dados e as informações geradas em tempo real de maneira rápida e eficiente, possibilita as organizações otimizarem, desde seus sistemas voltados à segurança da informação e mecanismos de proteção, até seus processos de auditoria e investigação forense.

Foi possível constatar também que a junção dos mais sofisticados algoritmos utilizados para análise preditiva e direcionados à detecção de anomalias com as tecnologias que lidam com Big Data, resulta em eficazes sistemas capazes de detectar, em tempo real, comportamentos anômalos em transações ou sistemas, contribuindo assim para um eficiente mapeamento de riscos e com a prevenção de fraudes em transações ou em ambientes corporativos.

Ainda que de maneira utópica, é possível comparar o emprego das tecnologias baseadas em Big Data aplicadas como mecanismo de defesa e combate ao crime em agências de espionagem e de polícias com o filme *Minority Report*, no qual a tecnologia existente, utilizada pela polícia no ano de 2054, permite prever o futuro e evitar que assassinos cometam seus crimes. Conforme os resultados obtidos pela polícia de Vancouver, vistos no item 4.3, é possível confirmar que a exploração do Big Data pode contribuir, e muito, no âmbito de segurança pública, por meio de toda a inteligência capaz de ser gerada a partir de dados já existentes.

Além de demonstrar as possíveis aplicações do Big Data, relacionadas com a área de segurança da informação, detecção e prevenção de fraudes, esse estudo procurou ressaltar as principais implicações sobre a preservação da privacidade dos dados coletados e armazenados em infraestruturas Big Data. Sobre este aspecto, é importante destacar que embora a preocupação com a privacidade dos dados não seja uma característica aplicada exclusivamente ao Big Data, é fundamental que as organizações façam o uso dos dados com muita ética e responsabilidade, além de aplicar os controles de necessários para os dados existentes nos chamados ecossistemas Big Data.

Como visto, existem alguns desafios específicos das tecnologias Big Data no que diz respeito à segurança e privacidade dos dados que devem ser muito bem avaliados antes do início de projetos que utilizem estes recursos, levando-se em conta, a existência de normas e leis sobre os segmentos e países cujos dados serão coletados e armazenados, já que uma possível brecha, ou violação de segurança pode acarretar, além de danos à entidade ou ao indivíduo, em severas multas e punições por parte de órgãos regulatórios.

Ainda sobre o aspecto de segurança e privacidade, é possível sugerir que os dados mais críticos para as organizações sejam monitorados também por sistemas capazes de lidar com Big Data, já que estes, por meio de suas particularidades, se mostraram eficazes e promissores sistemas capazes de melhorar o mapeamento de riscos e a detecção de ameaças e vulnerabilidades em sistemas de informação.

É importante destacar também que muitos profissionais da área de tecnologia da informação, ou das demais áreas requisitadas ainda não possuem uma visão clara, prática e consolidada sobre os conceitos e implicações do Big Data, o que certamente pode ser considerado como um empecilho para o sucesso de projetos voltados para este segmento, posto que como visto ao longo deste trabalho acadêmico, a aplicação prática destes recursos requer conhecimentos em múltiplas disciplinas, como infraestrutura, desenvolvimento, arquitetura e segurança de TI, além de conhecimentos sólidos na áreas de legislação e ciências da informação.

Sobre o mercado nacional, um aspecto importante a ser levado em consideração é o fato de muitas das tecnologias aplicadas aos sistemas orientados ao Big Data não serem amplamente difundidas e dominadas, o que pode ser considerado um risco para as organizações que pensam em iniciar projetos, já que a oferta de profissionais altamente especializados nestas tecnologias ainda é pequena. A boa notícia no Brasil é que a lei geral de proteção de dados pessoais LGPD, inspirada pela General Data Protection Regulation (GDPR) da UE, e aprovada no Senado em 2018 deve finalmente estar entrando em vigor agora em 2020; e o documento *prevê punições para eventuais abusos e define os direitos de usuários sobre os dados concedidos a terceiros, como a possibilidade do usuários solicitar a exclusão de suas informações pessoais de plataformas digitais de uma organização*[3].

Durante a análise de toda a bibliografia utilizada, foi possível constatar que este termo começou a ganhar mais destaque e visibilidade somente a partir do ano de 2011, como pode ser evidenciado a partir das datas das publicações utilizadas como referência para o desenvolvimento desta pesquisa. Ainda que um pouco mais de 3 anos represente um período considerável, do ponto de vista tecnológico certamente ainda existe muito a ser descoberto, inventado e melhorado no que diz respeito aos sistemas orientados ao Big Data.

Cabe observar a ausência de normas específicas para o segmento de Big Data difundidas entre os países, e recomendar a criação de convenções, acordos, ou até mesmo leis unificadas e compartilhadas entre os países, de forma que estas possam contribuir para um manuseio ético e transparente dos dados utilizados.

É possível sugerir também futuras pesquisas que possam contribuir para a criação de um *framework* capaz de tratar todos os desafios de segurança e privacidade adequada.

Por fim, a exploração do Big Data pode trazer notáveis benefícios e vantagens, nota-se que estes recursos tecnológicos ainda estão em franca expansão, e que portanto cabe sugerir que pesquisas mais aprofundadas nestes aspectos sejam efetuadas no futuro, para um melhor grau de mensuração de resultados.

REFERÊNCIAS

ASSOCIATION OF CERTIFIED FRAUD EXAMINERS (ACFE): Report to the nations on occupational fraud and abuse. 2012. Disponível em: <http://www.acfe.com/uploadedFiles/ACFE_Website/Content/rttn/2012-report-to-nations.pdf> Acesso em: 11 Abr. 2014.

BARREIRA JUNIOR, Eliseu. Como o 11 de setembro inaugurou a era do Big Data. Exame, 13 Set. 2013. Disponível em: <http://exame.abril.com.br/mundo/noticias/como-o-11-de-setembro-inaugurou-a-era-do-big-data?page=1>. Acesso em: 24 Abr. 2014.

BRYANT, Randal E. et al. Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. Computing Comunity Consortium, 2008. Disponível em: <http://www.cra.org/ccc/files/docs/init/Big_Data.pdf>. Acesso em: 14 Jan. 2014.

[3] https://olhardigital.com.br/noticia/senado-aprova-vigencia-imediata-para-a-lei-geral-de-protecao-de-dados/105902

BUNEMAN, Peter. et al. Data Provenance: Some Basic Issues. Universidade da Pensilvânia, 2000. Disponível em <http://db.cis.upenn.edu/DL/fsttcs.pdf>. Acesso em 05 Abr. 2014.

COX, M.; ELLSWORTH,D. Application-Controlled Demand Paging for Out-of-Core Visualization. Proceedings of the 8th IEEE Visualization '97 Conference, 1997. Disponível em: <http://www.evl.uic.edu/cavern/rg/20040525_renambot/Viz/parallel_volviz/paging_outofcore_viz97.pdf>. Acesso em: 14 Jan. 2014.

CLOUD SECURITY ALLIANCE (CSA). Expanded Top Ten Big Data Security and Privacy Challenges, 2013. Disponível em: <https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf>. Acesso em 08 Mar. 2014.

CLOUD SECURITY ALLIANCE (CSA). Big Data Analytics for Security Intelligence, 2012. Disponível em: <https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf>. Acesso em 20 Abr. 2014.

DAVE, Pinal. Big Data – What is Big Data. Journey to SQL Authority with Pinal Dave, 2013. Disponível em: < http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/>, Acesso em 17 Jan. 2014.

DENONCOURT, Don. How Safe Is NoSQL?. iProDeveloper, 2012. Disponível em <http://iprodeveloper.com/systems-management/how-safe-nosql>. Acesso em 27 Mar. 2014.

FLITTER, Emily; LYNCH, Sarah N. UPDATE 1-U.S. SEC's newest enforcement weapon: powerful software. Reuters, 2014. Disponível em: <http://www.reuters.com/article/2014/02/26/sec-enforcement-palantir-idUSL1N0LV1LZ20140226>. Acesso em 15 Abr. 2014.

GLOBAL FORENSIC DATA ANALYTICS SURVEY (GFDAS). Big Risks Require Big Data Thinking, 2014. Disponível em: <http://www.ey.com/Publication/vwLUAssets/EY-Global-Forensic-Data-Analytics-Survey-2014/$FILE/EY-Global-Forensic-Data-Analytics-Survey-2014.pdf>. Acesso em: 12 Abr. 2014.

GREGO, M. Ferramenta da NSA vê "quase tudo" que você faz na internet. Revista Exame, 31 Set. 2013. Disponível em: <http://exame.abril.com.br/tecnologia/noticias/ferramenta-da-nsa-ve-quase-tudo-que-voce-faz-na-internet?page=2>. Acesso em: 24 Abr. 2014.

GRIFFIN, R. Using Big Data to combat enterprise fraud: to combat fraud, more organizations are thinking big--employing new approaches around Big Data to convert the volumes of information available into useful insight and real action. Financial Executive, 01 Dez. 2012. Disponível em: <http://www.thefreelibrary.com/Using+Big+Data+to+combat+enterprise+fraud%3A+to+combat+fraud,+more...-a0313251600> Acesso em 10 Abr. 2014.

HURWITZ, Judith. et al. Big Data For Dummies. John Wiley & Sons, Inc. Nova Jersey, 2013.

KOETSIER, J; Big data: a retailer's guide to likes, tweets, reviews, customer data, and basically everything else (infographic), 2012. Disponível em <http://venturebeat.com/2012/11/19/big-data-a-retailers-guide-to-likes-tweets-reviews-customer-data-and-basically-everything-else-infographic/>. Acesso em 17 Jan. 2014.

LANEY, D. 3D Data Management Controlling Data Volume, Velocity and Variety, Application Delivery Strategies, Meta Group, 2001. Disponível em <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Acesso em: 14 Jan 2014.

MAYER, V.; CUKIER, K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Nova Iorque: Hardcover, 2013.

MCDULING, John. Proof that "big data" is one of the most overused corporate buzzwords of 2013. Quartz, 2013. Disponível em: <http://qz.com/151109/proof-that-big-data-is-one-of-the-most-overused-corporate-buzzwords-of-2013/> Acesso em: 28 Abr. 2014.

MOREIRA, Daniel Augusto. O método fenomenológico na pesquisa. São Paulo: Pioneira Thomson, 2002.

MURPHY, Barry. Information Governance Even More Important In The Era Of Big Data. Forbes, 11 Jul. 2012. Disponível em: <http://www.forbes.com/sites/barrymurphy/2012/11/07/information-governance-even-more-important-in-the-era-of-big-data/>. Acessado em: 19 Mar 2014.

Open Data Center Alliance (ODCA): Big data Consumer Guide, 2012. Disponível em: <http://www.opendatacenteralliance.org/docs/Big_Data_Consumer_Guide_Rev1.0.pdf> Acesso em: 15 Jan 2014.

ORACLE - Enterprise Information Management: Best Practices in Data Governance. Maio de 2011. Disponível em: <http://www.oracle.com/technetwork/articles/entarch/oea-best-practices-data-gov-400760.pdf>. Acesso em 19 Mar 2014.

PRESS, Gil. A Very Short History Of Big Data. Forbes, 5 Abr. 2013. Disponível em: <http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/> Acesso em: 14 Jan 2014.

PROX, Ryan. Establishing an Analytics Culture in Public Safety, Government Technology, 29 Ago 2013. Disponível em: <http://www.govtech.com/data/Establishing-an-Analytics-Culture-in-Public-Safety.html>. Acesso em: 24 Abr. 2014.

REGALADO, A. Spinoffs from Spyland: How America's eavesdropping agency commercializes technology. MIT Technology Review, 18 Mar. 2014. Disponível em: <http://www.technologyreview.com/news/525541/spinoffs-from-spyland/>. Acesso em: 24 Abr. 2014.

REILLY, O'. Big Data Now. Nova Iorque: O'REILLY Media, 2012.

SAVITZ, Eric. How Big Data Will Transform IT Security. Forbes, 26 Fev. 2013. Disponível em <http://www.forbes.com/sites/ericsavitz/2013/02/26/how-big-data-will-transform-it-security/>. Acesso em 22 Fev. 2014.

SECURITIES AND EXCHANGE COMMISSION (SEC). SEC Enforcement Actions: FCPA Cases. 2014. Disponível em: <https://www.sec.gov/spotlight/fcpa/fcpa-cases.shtml>. Acesso em: 14 abr 2014.

TANKARD, Colin. Big Data Security. Network Security, Volume 2012, Issue 9, Páginas 10–15, Set. 2012. Disponível em: <http://www.sciencedirect.com/science/article/pii/S135348581270082X>. Acesso em 22 Fev 2014.

TRUSTWAVE; 2013 GLOBAL SECURITY REPORT, 2013. Disponível em: <http://www2.trustwave.com/rs/trustwave/images/2013-Global-Security-Report.pdf>. Acesso em 21 Abr 2014.