Big Data: Ferramentas e Aplicabilidade

Natanael Galdino
natan.gald@hotmail.com
IESSA

Resumo:O Big Data representa a vasta quantidade de informação gerada diariamente através dos mais diversos dispositivos eletrônicos e o tratamento analítico dessa informação através de diversas ferramentas Tecnológicas, com o intuito de se obter padrões, correlações e percepções que podem auxiliar em tomadas de decisões nas mais diversas áreas. Com o a popularização da internet e o advento de diversos dispositivos tecnológicos, a geração de dados cresceu exponencialmente nos últimos anos. Com isso, e através de inteligências tecnológicas há tempos existentes como Business Intelligence, foram concebidas outras ferramentas destinadas a trabalhar com tipos diferentes de dados, principalmente aqueles não possíveis de serem administrados em sistemas relacionais. Este artigo se destina, através de uma revisão de literatura, a esclarecer pontos relevantes sobre Big Data como suas principais ferramentas e soluções, além de casos de uso bem sucedidos. Para isso foram utilizados como fontes livros, artigos publicados, vídeos explicativos e revistas que auxiliam na compreensão dessa tecnologia.

Palavras Chave: Big Data Analytics - NOSQL - Processamento - Dados - Clusters

1. INTRODUÇÃO

A quantidade de dados gerados pela humanidade nos últimos anos aumentou de forma exponencial. Segundo uma pesquisa recente (IBM , 2013), no ano 2000, 25%(vinte e cinco por cento) dos dados eram digitalizados, no ano de 2007, esse numero saltou para 93% (noventa e três por cento), e no ano de 2013, foi para 98% (noventa e oito por cento). Esse crescimento, devido principalmente a fatores como aumento do acesso a dispositivos eletrônicos e a popularização da internet, está gerando uma revolução no tratamento de dados.

A aplicabilidade do Big Data está no tratamento desse volume de dados, que vem de variadas fontes e que demandam alta velocidade de processamento, na busca por um valor (Taurion, 2013).

Esse valor, obtido através de correlações entre dados, pode se dar através de descoberta de padrões, preferências de usuários, aumento no número de vendas em determinada época do ano, descoberta de cura de doenças, entre diversos outros benefícios aplicáveis a diversas áreas de estudo.

Por ser um assunto relativamente novo, muitos artigos que mencionam o tema, o fazem de maneira conceitual e sem abranger alguns detalhes, que vão além de conceitos pontuais. A parte prática de Big Data é um ponto importante a ser mencionado para um maior entendimento. Nesse sentido esse artigo tem por objetivos: Apontar as principais diferenças entre os modelos tradicionais de tratamento de dados e os modelos de Big Data; apresentar ferramentas mais importantes de uso do big data que ajudam a esclarecer melhor como o Big Data funciona na sua essência; discorrer sobre de três casos de uso de sucesso que confirmam a eficiência e o impacto que essa nova tecnologia tem proporcionado à sociedade. Para tal, foram usadas diversas fontes de pesquisa, como livros, revistas, vídeos explicativos, sites, artigos publicados.

2. ENTENDENDO O BIG DATA

O aumento exponencial dos dados no decorrer dos anos através do advento da internet e de diversos dispositivos como celulares e computadores ocasionou uma revolução no que tange a gestão da informação. Segundo Santanchè (2014), o Big Data, embora tratado por muitos como solução, em si é um problema, pela quantidade e diversidade de dados, que será resolvido através das ferramentas de Big Data Analytics.

A origem dos dados vem basicamente de Web e redes sociais (dados de fluxo de cliques, blogs, posts, feeds de notícias), dados de transações (compras de cartão de crédito, registros de ligações e de reclamações nas empresas) dados de biometria (identificação automática, DNA, impressões digitais, reconhecimento facial) dados gerados por pessoas (privados e que devem ser protegidos por legislação, como documentos eletrônicos, exames e registros médicos, ligações telefônicas) e dados machine to machine (gerados diretamente por maquinas, como sensores, dispositivos de GPS e medidores). (Intel, 2015).

Os cinco VS, Volume (quantidade de dados acumulados), Variedade (meios de propagação e tipos de dados), Velocidade (taxa de transmissão de dos dados), Veracidade (se os dados são confiáveis) e Valor (resultado obtido no uso das

ferramentas de Big Data) denotam o objetivo de manter as plataformas e sistemas em harmonia de tal forma que gerem o resultado esperado. (Veja, 2013).

Os dados são qualificados em três categorias: dados estruturados, pertencentes a um SGBD relacional com esquema relacional associado, dados semiestruturados, que são irregulares ou incompletos não necessariamente de acordo com um esquema, compreensíveis por maquinas mas não por seres humanos, como documentos HTML e logs de web sites , e dados não estruturados, sem estrutura prévia nem possibilidade de agrupamento em tabelas, como vídeos, imagens e emails.(Intel 2015)

O desafio para as ferramentas de Big Data é entre outros a manipulação de dados semiestruturados e não estruturados no intuito de extrair valor destes através de correlações e outros processamentos de análise e então compreendê-los para que tragam valor ao determinado meio aplicável.

O tratamento dos dados é realizado com o apoio de algoritmos inteligentes, que são sequencias de instruções que permitem que se chegue a uma conclusão sobre que tipo de ação tomar. Esses algoritmos, são a "rede neural" do sistema e podem servir para fins diversos dependendo do propósito buscado pela corporação. Uma empresa pode compreender melhor o comportamento de um cliente, um médico pode saber se o paciente de uma clinica necessitará ser internado em determinado período ou de que maneira, é possível reduzir despesas dentro de uma empresa. A Amazon usa a inteligência de algoritmos para indicar produtos aos seus clientes. A Netflix segue o mesmo caminho indicando séries conforme as séries já assistidas por seus clientes.

Cezar Taurion(2013), em seu livro Big Data, faz uma analogia em que as ferramentas de Big Data, representarão para as corporações e para a sociedade a mesma importância que o microscópio representou para a medicina. Uma ferramenta de análise onde se pode extrair informações, prever incidentes e ter a capacidade de corrigi-los quando existentes, ou até mesmo evitá-los.

Os algoritmos de sistemas preditivos, que com base em dados processados "predizem" um fato com grandes probabilidades de ocorrer, são um grande desafio a ser superado nessa lacuna que existe entre aplicabilidade em tempo real, e análise de dados anteriores para se tomar decisões. Os sistemas relacionais de bancos de dados, há tempos aplicados em empresas e rendendo sucesso nesse ponto, tornam-se incapazes tanto de trabalhar com o imenso número de informações quanto fazer análises preditivas e em tempo real. Nesse conceito a streaming computing, que trabalha com dados em tempo real e grande fluxo de dados, como, em sistemas de trânsito, que monitoram o tráfego de veículos em determinada cidade, e que transmitem ao usuário qual a melhor rota a ser tomada para chegar ao seu destino, através de seus algoritmos, traz soluções práticas e rápidas aos seus usuários. (Taurion, 2013).

No entanto, deve-se seguir o princípio de que não existe a melhor ferramenta, mas sim a que melhor se adéqua as necessidades da corporação. Para algumas corporações, o uso de ferramentas tradicionais SQL, com sistemas preventivos, que comparam vendas em períodos do ano, para projetar promoções, por exemplo, já são suficientes para o negocio. Portanto, há que se considerar as necessidades de negocio para adotar a ferramenta apropriada e que traga o resultado esperado.

3. SISTEMAS TRADICIONAIS X SISTEMAS DE BIG DATA ANALYTICS

O gerenciamento de informações há tempos é um conceito adotado em corporações que desejam aperfeiçoar seus processos através de métricas de recolhimento e tratamento de dados.

A diferença no processamento de dados de modelos tradicionais (SQL) para modelos de Big Data Analytics, começa pela diferença entre escalabilidade vertical e horizontal.

Na escalabilidade vertical, usada em sistemas SQL, para poder ter um melhor poder de processamento, investe-se em máquinas com tecnologias mais avançadas e consequentemente mais caras, assim aprimorando o processamento dos dados. Na escalabilidade horizontal, usa-se computação paralela em que maquinas de nível intermediário "commodities", que são usadas em conjunto para processar uma quantidade de dados que apenas uma delas seria incapaz de processar, assim, reduzindo custos e possibilitando o processamento de grandes volumes de dados. (Coelho, 2004).

Nos modelos tradicionais, o conceito de Business Intelligence, que em síntese, é uma técnica de gerenciamento de negócios orientado à análise de informações, com o intuito de conhecer fatos que afetam positiva ou negativamente o negócio, sendo um forte auxiliar nas tomadas de decisões.

A ferramenta ETL (Extração, Transformação e Carregamento), seguindo o principio do Business Intelligence, é uma tecnologia usada em muitas corporações, e que faz a coleta de dados de todos os tipos e formatos, transforma-os, através de algoritmos, aplicando princípios de correlações entre esses dados e carrega-os em um ambiente de visualização, em que administradores da alta gerência, podem visualizá-los, podendo extrair informações que os permitirão ações de melhoria nos processos organizacionais.

No processo de análise de informações, a ferramenta OLAP (Processo analítico Online), auxilia na tomada de decisões através de cubos multidimensionais que oferecem diferentes perspectivas sobre informações da empresa como regiões e períodos em que determinados produtos são mais vendidos, padrões de consumo dos clientes, entre outras analises. (Intel, 2016);

A quantidade de dispositivos somada aos diversos formatos de arquivos, e a necessidade da extrair de valor dos mesmos, mostrou a limitação dos modelos relacionais, que serviam bem para o tratamento de dados estruturados, mas não possibilitavam o tratamento de dados semiestruturados ou não estruturados. Esse motivo foi um dos principais motivadores da busca de ferramentas NOSQL, que trabalham com bancos de dados não relacionais. Além da maior quantidade de dados, sistemas NOSQL são preparados para trabalhar em sistemas instáveis em relação aos modelos RMDBS (Sistema de gestão de Bancos de dados Relacionais), tendo um processamento mais complexo. Ainda no modelo NOSQL, os dados oriundos de diversos dispositivos desde aparelhos móbiles até servidores, são replicados em clusters onde são processados através de ferramentas Analytics, e posteriormente visualizados através de gráficos, dashboards, entre outras ferramentas de análise, tal qual no modelo ETL, também usado nos modelos relacionais. O processo conhecido

como retroalimentação, em que dados já processados são novamente usados em um segundo processamento contribui para manter informações atualizadas e confiáveis. (Intel, 2015)

3.1. PROPRIEDADES ACID E CAP

Essas estruturas definem o comportamento da base de dados. Nos modelos relacionais, busca-se manter as propriedades *Acid* (Atomicidade[1], Consistência[2], Isolamento[3] e Durabilidade[4]. No entanto em modelos não relacionais, onde o fluxo de dados é maior, torna-se impossível mantê-las, surgindo então as propriedades *Cap* (Consistency, *Availability* e *Partition Tolerance).* Sendo possível escolher apenas duas, a corporação deverá escolher se prefere um sistema sempre disponível, tolerante a falhas, ou consistente, em que todos os usuários terão a mesma informação ao mesmo tempo. Em redes sociais como Facebook, por exemplo, o tempo de visualização pode ser diferente entre usuários, portanto, a consistência pode ser colocada em segundo plano em detrimento da disponibilidade e tolerância a falhas. Já em lojas de comercio eletrônico como a Amazon, abrir mão da consistência, pode implicar vender produtos com estoque esgotado aos clientes, gerando atrasos, cancelamentos e prejudicando a imagem da empresa, então deve-se abrir mão da tolerância a falhas ou disponibilidade, para manter o sistema sempre consistente a todos os usuários. (Intel,2015)

Ao anular as propriedades *Acid,* os sistemas obterão as propriedades *Base*:

1) *Basically Avaliable*: Dados serão replicados e serão sempre consistentes;
2) *Soft State*: Dados inconsistentes serão tratados posteriormente;
3) *Eventually Consistent*: Garante a consistência em algum momento.

Existe um conceito, ainda novo no mercado que trata sobre modelos *NewSql*, os quais mantém o modelo ACID ao mesmo tempo que buscam manter o rendimento de banco de dados NOSQL, os quais com o decorrer dos anos, devem ser aprimorados até se
tornarem padrões, melhorando assim a todos os pontos do sistema, sem que seja necessário abdicar de nenhum, como ocorre no modelo CAP. (Intel,2015)

A seguir apresenta-se a imagem ilustrativa da propriedade CAP, com e o exemplo de bancos de dados NoSQL como Cassandra, CouchDB e Riak que são tolerantes a falhas e sempre disponíveis, e bancos de dados Nosql como MongoDB, Hbase e BigTable, que são consistentes e tolerantes a falhas.

---

[1] **Atomicidade**: Toda transação deve ser bem sucedida, ou não ser realizada, (Intel, 2015);
[2] **Consistência**: O banco de dados deve permanecer consistente ao realizar uma operação, (Intel, 2015);
[3] **Isolamento:** Várias operações são executadas ao mesmo tempo sem interferência de uma em outra, (Intel, 2015);
[4] **Durabilidade**: Transações completas devem persistir e não ser alteradas, (Intel, 2015).

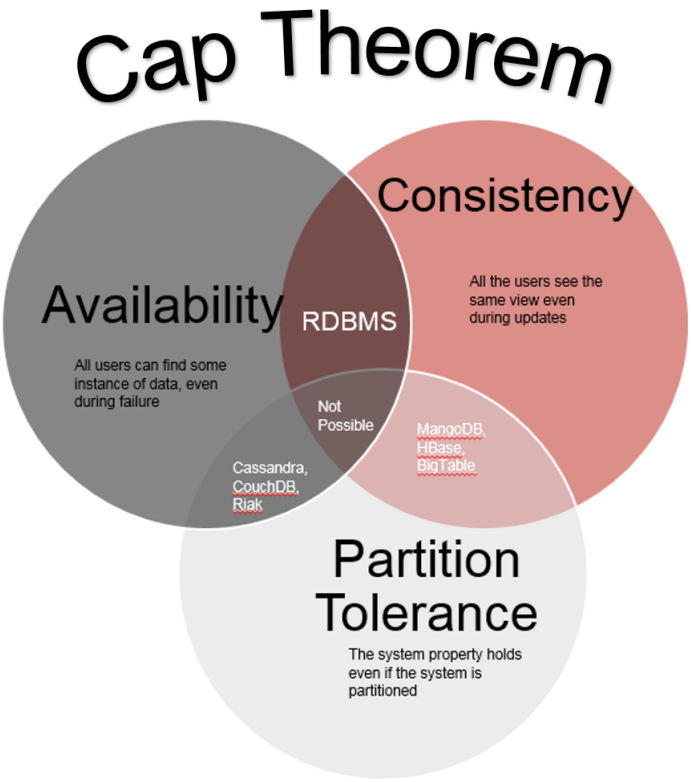

Figura 1: Exemplo do CAP, demonstrando a impossibilidade de se obter as três propriedades. RamaNathan(2014)

4. FERRAMENTAS

4.1. AMBIENTES EM NUVEM

A computação em nuvens (Cloud Computing) é uma grande aliada no uso de ferramentas de big data. A queda no preço de armazenamento ao longo dos anos, aliada à elasticidade que ambientes em nuvem oferecem facilitam o acesso a esses serviços até mesmo para corporações que não tem muito dinheiro para investir. Diferentemente de mainframes que custam pra empresa um valor considerável, e muitas vezes não é utilizado completamente, os ambientes em nuvem permitem o pagamento por hora e somente cobram pela quantidade de informação necessitada pela empresa. A escalabilidade permite que as configurações de nuvem, quanto ao número de visitas ao sistema, desempenho, processamento dos dados entre outros, seja aumentada somente quando a empresa realmente necessite disso, como em épocas em que as vendas aumentam, Natal e Black Friday, por exemplo, e posteriormente volte a operar com menos servidores, evitando gastos desnecessários com servidores que seriam usados apenas em um período do ano. (CPBR6, 2013)

4.2. HDFS

O Hadoop Distributed File System ou Sistema de arquivos distribuídos surge com a necessidade de se trabalhar com arquivos grandes. O HDFS faz a quebra em blocos desses arquivos e os distribui em diversos nós (máquinas), com replicação em grau três como segurança no caso de um nó falhar. O Name Node é a máquina responsável pelo gerenciamento dos outros nós, e envia informações (Heartbeats) para o código, em caso de um nó falhar, além fazer a redistribuição dos blocos de dados quando houver falha, sempre mantendo grau três. (Paiva, 2016)

4.3. YARN

É um gerenciador de recursos distribuídos do cluster. Através do Resource Manager, realiza a locação de recursos nos nós do cluster para a realização de tarefas das aplicações. Dessa maneira, cada aplicação sabe em que maquina os seus recursos estão alocados, e mantém o principio da localidade, que é realizar o processamento do código onde estão os dados. (Yarn, 2016)

4.4. MAP REDUCE

É o sistema analítico do Hadoop desenvolvido para operar com grandes volumes de dados. Segue o principio da localidade em que o código é enviado para o local onde os dados estão para ser processado. O processamento analítico é distribuído em vários servidores, dos quais se deseja tirar informação. Através de um processamento paralelo/distribuído, os dados são divididos em partições ou ficheiros através da função *Split*. Nesse processo, o Map reduce monta a separação dos dados em partições, mapeia as atividades em cada local e duplica em ambientes e depois faz as reduções. Durante o mapeamento através do processamento em cada nó da partição ou cluster, são formados pares valor chave enviados ao redutor, agrupando pares com as mesma características. Basicamente são três fazes, a saber: Map, onde todos os dados são reunidos; Shuffle, onde os dados são reunidos e organizados e Reduce, onde os dados são associados e correlacionados. Nem todos os algoritmos se encaixam nesse modelo. (Paiva, 2016)

4.5. HADOOP

É a ferramenta mais importante de Big Data. Através de nós de clusters usa computação distribuída com alta escalabilidade, tolerância a falhas e confiabilidade. Sendo uma plataforma Java de computação, ela é voltada para clusters e processamento de grande volume de dados A ideia principal do Hadoop é tratar essa grande quantidades de dados sem ter a necessidade de copiar esses dados em outro servidor, o que ocasionaria mais tempo e investimento. No processo Hadoop, os dados são tratados dentro dos servidores e em tempo real, gerando mais praticidade no processamento e economicidade de tempo e dinheiro. Busca manter a redundância e tolerância a falhas através da replicação dos dados, assim, se houver falha em um dos clusters (rodapé), haverá outro disponível para manter o processamento, além de poder executar um algoritmo, em qualquer uma das partições ou clusters, sendo esse algoritmo disseminado em outros nós de clusters, o que simplifica o processo e deixa o

sistema mais rápido. É formado basicamente pelo framework Map Reduce, pelo gerenciador de recursos distribuídos (YARN) e pelo sistema de arquivos distribuídos (HDFS). (Intel, 2016)

4.6. MPP

Massively Parallel Processing ou processamento massivo paralelo, é um paradigma de Big Data, feito para processar grandes quantidades de informações, é escalável em relação a quantidade de dados, e suporta linguagem SQL e tabelas relacionais, sua diferença quanto ao Hadoop reside no fato de que é um paradigma de estrutura rígida, e não permite trabalhar com imagens ou documentos de texto. Pode trabalhar em conjunto com Data Warehouse, fazendo operações paralelas. (Big Data Now, 2013).

4.7. HBASE

É um banco de dados Nosql que processa grandes volumes de dados de maneira rápida e em tempo real. Trabalha com o conceito chave – valor, em que cada dado é associado a outro trazendo uma característica similar ao modelo relacional com sua organização se dando em linhas, colunas, tabelas e famílias de colunas. No entanto não há a obrigatoriedade de esquemas, como ocorre no modelo SQL, portanto pode haver linhas sem determinadas colunas e vice-versa. Nesse modelo, diferentemente do SQL, os dados não são alterados, apenas somados, podendo haver varias versões sobre determinada chave ou valor. (Paiva, 2016)/(CPBR6, 2013)

4.8. SPARK

Ferramenta de processamento de dados que roda até 100 vezes mais rápido que o Map Reduce. Como o Map reduce não processa bem todos os algoritmos, o Spark atua sendo mais abrangente na questão de diferentes tipos de processamento. Também executa o código em paralelo. Sua principal diferença em relação ao Map Reduce é o fato deste persistir em disco. O Spark trabalha em memória, faz encadeamento de funções e só apresenta o resultado no fim do processamento. O driver, aplicação principal do Spark, faz alocação maquinas no cluster para processamento de funções. Pode trabalhar tanto com o paradigma SQL quanto o NOSQL. (Paiva, 2016)

4.9. MACHINE LEARNING

Machine Learning é o termo que designa o processo de ensinamento da maquina a "entender" dados que a princípio parecem não fazer sentido, processá-los e tirar algum valor disso. Pode-se usar machine learning, por exemplo, em redes sociais, posts ou tuites, com expressões diferentes das formais, por exemplo: "Pato passa em branco no jogo do tricolor", usam-se algoritmos para que a máquina entenda que "Pato" não é um animal e sim um jogador de futebol, "passar em branco" significa não

fazer gol e "tricolor" significa um time de futebol, nesse caso, pode-se medir o nível de satisfação dos torcedores em relação ao time, ou em casos parecidos, o nível de satisfação de clientes em relação a uma empresa, através do que eles postam nas redes sociais. Algoritmos de machine learning auxiliam principalmente a transformar dados que a principio seriam não estruturados, em dados estruturados.

Outra forma de usar machine learning é através de computação cognitiva, e biometria. Com base no comportamento de um indivíduo em frente ao caixa eletrônico, usa-se uma tecnologia kinect, que mapeia regiões do corpo do suspeito, e através de algoritmos de inteligência artificial, é possível reconhecer o perfil comportamental de um bandido ou fraudador de cartões, passando à segurança do local essas informações, pode-se melhorar a segurança do local. (Nogare, 2014)

5. CASOS DE USO

A aplicabilidade do Big Data Analytics pode ocorrer em diferentes ramos, trazendo melhoria a processos organizacionais e apoio a tomada de decisões, tal qual Business Intelligence, que através das informações coletadas, toma estratégias para um melhor desempenho na área aplicada, e indo mais além, podendo inclusive prever tendências com base na análise de dados.

5.1. SAÚDE

Vários algoritmos de predição podem ser implantados com base no grande número de informações disponibilizadas na área da saúde. É possível cruzar diversas informações como dados de poluição atmosférica, sintomas de determinada doença feitos em uma consulta médica, até mesmo postagens feitas em redes sociais de pessoas falando que estão com determinada doença. Toda essa informação pode ser correlacionada para poder chegar a conclusões como, em que região determinada doença está mais presente. Assim atuou a ferramenta Google Trends, quando o mundo sofreu com o surto de epidemia H1N1.

No Brasil, a INCOR (Unidade de imunologia do Instituto do Coração) faz uso do Big Data através de algoritmos disponíveis em banco de dados, do mundo inteiro para verificar a mutação do vírus do HIV, podendo perceber suas variações e assim desenvolverem vacinação mais eficazes contra essas variações, além de ser uma ferramenta auxiliar na busca da cura dessa doença. (Exame 2014)

5.2. EMPRESA DE TRANSPORTE AÉREO

Um dos maiores gastos que as empresas de transporte aéreo têm é o de combustível, sendo responsável por 30% de todas as despesas em um mês. Uma empresa dos Estados Unidos, sabendo desse fato decidiu aplicar Big Data nesse campo com o intuito de encontrar formas de economizar. Através de informações obtidas por meio de sensores acoplados no avião, em um voo transatlântico a empresa obtém 640 Terabytes de dados, e com esses dados em mãos, consegue prever diversas situações como tempo para uma nova revisão no avião, quando será necessário fazer abastecimento, entre outras informações que auxiliam na tomada de decisão. Notou-se

que com a melhoria de 1% na usabilidade do Big Data, geraria um lucro de 30 bilhões de dólares em 15 anos, com essa economia, pode-se fazer melhorias outras demais áreas da companhia gerando um benefício para o negocio como um todo. (Diálogo Intel, 2015)

5.3. SEGURANÇA PUBLICA

Depois dos atentados terroristas de 11 de setembro, as autoridades americanas, fizeram uma revolução em seus sistemas de segurança nacional, aplicando ainda mais o uso de tecnologias em seus processos de segurança através de ferramentas de Big Data.

No estado do Tennessee, uma ferramenta totalmente aplicada à segurança pública tem mostrado grande eficiência tanto no desvendamento de crimes, captura de criminosos, e também na prevenção de delitos. Através de diversas câmeras de segurança, sensores, informações de terceiros, e uma monitoração de dados por uma central, é possível controlar lugares suscetíveis a crimes, horário em que comunmente delitos ocorrem, e deslocar tropas para determinado local antes que o delito ocorra, ou em caso da fuga do criminoso, é possível organizar tropas policiais, com base na rota de fuga e por meio da central orientar a tropa sobre qual caminho mais eficiente a ser tomado para a captura do criminoso.

Em 2013, durante uma maratona na cidade de Boston, um atentado terrorista causou a morte de três pessoas e feriu outras 264. A policia local implantou um sistema de Big Data, que recolhia informações disponibilizadas por terceiros, com dados como várias filmagens que aconteceram durante a maratona por celular pessoal ou câmeras de segurança, análise de comportamento de indivíduos e, sobretudo, quem estava portando mochilas durante a maratona. Através do estudo de caso e correlação de diversos dados a policia conseguiu identificar e prender o terrorista. (KM e Canal Mais, 2015)

6. CONSIDERAÇÕES FINAIS

Como se pode notar, já existe uma gama de bons resultados alcançados pelas ferramentas de Big Data, que já servem inúmeras áreas de serviço e pesquisa, gerando bons resultados, e trazendo retorno financeiro e operacional. No entanto, há que se ressaltar, o pouco tempo de existência de tais ferramentas, e a necessidade de melhoria em alguns pontos como: interoperabilidade entre sistemas, algoritmos mais eficientes, mão de obra qualificada e melhor conhecimento da área de governança das empresas, sobretudo em países como o Brasil, onde o nível de envolvimento com Big Data pelas empresas ainda é pequeno, o que faz o investimento na área não ser o necessário, impedindo a empresa de obter resultados dessa tecnologia.

Um fator que deve ser considerado com base em princípios de veracidade e valor, é até que ponto o Big Data é definitivamente eficaz em suas conclusões. Renê de Paula (2013) menciona que não se deve considerar um grupo/fator homogêneo nas tomadas de decisões em Big Data. Dados os mais diversos grupos inseridos na sociedade com opiniões e gostos diferentes, as soluções Big Data que podem ser muito aplicáveis a um grupo, podem não fazer diferença alguma para outro. Tal relevância deve ser

considerada na criação de sistemas cada vez mais objetivos, tendo o cuidado de não chegar a conclusões errôneas, pois se pode existir margem de erro quando se trabalha com bancos de dados relacionais, e essa margem de erro pode parecer pequena, em um sistema de Big Data Analytics, no tratamento muitas vezes de terabytes ou exabytes de dados, essa margem de erro, pode apresentar um número significativo. Nesse ponto se faz necessário o uso de algoritmos cada vez mais eficientes, que evitem erros ou que busquem reduzi-los ao máximo possível.

Pela complexidade do assunto, e a impossibilidade de discorrer sobre vários outros temas que envolvem a abordagem de Big Data, destaca-se dois assuntos de importância a ser abordados em futuros artigos. O primeiro é a segurança da informação envolvendo Big Data. Uma vez que os dados tratados podem ser referentes à saúde de clientes, transações financeiras, dados de documentos pessoais, há que se manter uma política de segurança nesse sentido. O segundo é a falta de expertise na área que é outro dos grandes desafios para o Big Data, principalmente nas profissões de Engenheiro e Cientista de dados, os quais devem não somente entender cada ferramenta, mas saber aplicá-las aos negócios extraindo as mais diversas informações e gerando valor independente da área de atuação.

O crescimento ainda que gradual pelos problemas acima citados faz-se notável, e aos poucos vai se tornando parte da vida não só das corporações mas também da sociedade, facilitando processos e dando uma visão de como os problemas poderão ser resolvidos no futuro.

7. REFERÊNCIAS

CPBR6 – Big Data e computação em nuvem, 2013 ;
disponível em: https://www.youtube.com/watch?v=dOR3hGS_IZI

INTEL. Curso Big Data. 2015
Disponível em: dialogoti.intel.com/pt-br/curso/big-data

KM E CANAL MAIS BIG DATA - Les nouveaux devins – Documentaire,2015
Disponível em: https://www.youtube.com/watch?v=5mmQeb8mXVk

NOGARE, D. Conceitos e Ferramentas de Big Data, 2015
Disponível em: https://www.youtube.com/watch?v=ByHJTRyP4_A

O' REILLY MEDIA. Big Data Now Edição 2013

COELHO, O. P. Arquitetura: Princípios para alcançar Desempenho e escalabilidade em Aplicações,2004
Disponível em : https://msdn.microsoft.com/pt-br/library/cc518051.aspx

PAIVA, R. Curso de Big Data - Aula 2 - Principais Ferramentas (Hadoop, HBase e Spark),2016
Disponível em: https://www.youtube.com/watch?v=CjRkEywm1go

PAULA, R. Big Data, CSI Miami e por que eu ainda prefiro Law and Order, 2013
Disponível em : https://www.youtube.com/watch?v=Qvnq8j5qd4E

RAMANATHAN, S. Lambda Architecture for Big Data – Quick peek.2014
Disponível em: http://blogs.perficient.com/dataanalytics/2014/12/10/lambda-architecture-for-big-data-quick-peek/

Revista Exame. 5 coisas que o big data faz pela sua saúde. 2015

Disponível em : http://exame.abril.com.br/publicidade/dell/5-coisas-que-o-big-data-faz-pela-sua-saude

REVISTA VEJA. Edição 2321 – Entenda o que é Big Data,2013

SANTANCHÈ, A. - NoSQL e Big Data - Aula 27 - Bancos de Dados 2015.2
Disponível em : https://www.youtube.com/watch?v=-a2pyU0uhww

SOFTWARE INTE. Extract, Transformation and Load Big Data with Apache Hadoop,2016
Disponível em: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

TAURION, C. Big Data. Brasport.2013

YARN – Hadoop beyond MapReduce; 2016.
Disponível em : https://www.youtube.com/watch?v=HHv2pkIJjR0http://spark.apache.org/