Big Data
Desafios e Oportunidades na Análise Massiva de Dados

Eduardo Martins Barros
de Albuquerque Tenório

Recife, 23 de janeiro de 2014

0 — Índice

1 — Motivação

O interesse e a execução de estudos com grandes quantidades de dados não é algo novo. Desde o surgimento dos computadores modernos, diversos tipos de aplicações exigiram o processamento de dados diversos, coletados em grande escala e cuja saída poderia influenciar drasticamente uma decisão (financeira, científica, militar e etc.).

Nos últimos anos estas aplicações puderam aproveitar a popularização dos supercomputadores graças à Lei de Moore, que proporcionava mais poder computacional a custos cada vez mais baixos. Contudo, a evolução desacelerou, tornando o investimento em computadores mais potentes inviável (Lei de Moore torna-se obsoleta). Logo, novas alternativas foram pensadas, dentre elas o uso de sistemas distribuídos compostos de unidades de processamento mais simples operando em paralelo.

Houve também um aumento da informática em nosso cotidiano, tendo como efeito "colateral" uma explosão de rastros digitais deixados em qualquer sítio virtual ou físico. Seja procurando um livro na Amazon ou pagando um jantar com o cartão de crédito, a quantidade de informações produzidas permite entender além de qual livro foi pesquisado ou o quanto se gastou no jantar. Pode-se saber quais gêneros literários o usuário prefere (histórico), próximas compras (listas de desejos) e sua maneira de pesquisar (navegação pelo site). Também é possível saber se a pessoa jantou sozinha (nota fiscal), se foi um jantar romântico ou de negócios (vinho, champanhe e demais pedidos) e até se quem pagou a conta distribui gorjetas (10% opcionais).

Com este cenário, decisões que antes eram baseadas em achismos (profissionais ou não) podem ser tomadas com o auxílio de uma análise quantitativa e qualitativa, agregando valor aos negócios e cortando despesas difíceis de detectar "a olho nu". O surgimento do Big Data foi apenas consequência deste novo paradigma (computação distribuída e paralela) aplicado à crescente proliferação de dados digitais.

2 – Introdução

Big Data não é algo inteiramente novo. A análise estatística de grandes quantidades de dados é estudada desde os primórdios da computação. A novidade decorre do fato de um alto poder computacional estar ao alcance de virtualmente qualquer pessoa, tornando a infraestrutura de TI praticamente uma commodity. Assim, foram desenvolvidas técnicas de análise estatística que conseguem trabalhar com grandes quantidades de dados variados, utilizando sistemas distribuídos e paralelos e com tempo de resposta praticamente instantâneo.

Existem algumas definições para o que atualmente é entendido por Big Data. Segundo o SAS Institute (Statistical Analysis System), Big Data é *"um termo popular usado para descrever o crescimento, a disponibilidade e o uso exponencial de informações estruturadas e não estruturadas"* [1]. O professor Christopher Barnatt, da Universidade de Nottingham considera Big Data *"a próxima grande coisa em computação, e gera valor a partir de grandes coleções de dados que não podem ser analisadas com técnicas tradicionais de computação"* [2].

Embora ainda não haja uma definição muita precisa, existem alguns consensos. Basicamente, Big Data é a análise estatística de grandes coleções de dados (**volume**), dispostas em formatos diversos (**variedade**), geradas e analisadas com rapidez (**velocidade**) [2][3]. Estes são os tradicionais 3 V's do Big Data.

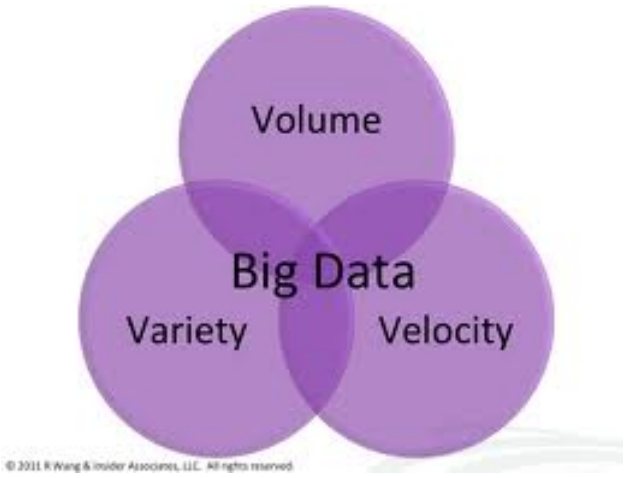

Figura 1. OS 3 V's do Big Data.

3 — Exemplos conhecidos

Atualmente existem diversos exemplos de soluções de Big Data, abrangendo todas as principais definições possíveis. Contudo alguns são mais emblemáticos, seja pela criatividade da aplicação, seja por quem está desenvolvendo-a.

3.1 — Moneyball

Moneyball é um livro (e posteriormente um filme de 2011) que tem como pano de fundo o uso de análise estatística para melhorar a performance de uma equipe. Foi uma das primeiras abordagens do Big Data para o público leigo. Na história (real) um dirigente de um time de baseball mediano passa a usar os serviços de um matemático para ajudá-lo a montar um time mais competitivo [4], baseando suas contratações em critérios diversos, desde qual jogador faz mais *home runs* até quem vende mais camisas (e assim conseguir caixa para contratar mais *home runners*). Em seguida, times maiores passaram a empregar técnicas semelhantes.

Embora seja um exemplo popular do uso de Big Data, a análise estatística efetuada carece de uma variedade de dados. O que existe são estatísticas bem estruturadas dos jogadores da liga de baseball. Há também o fato do tempo de resposta não precisar ser quase instantâneo.

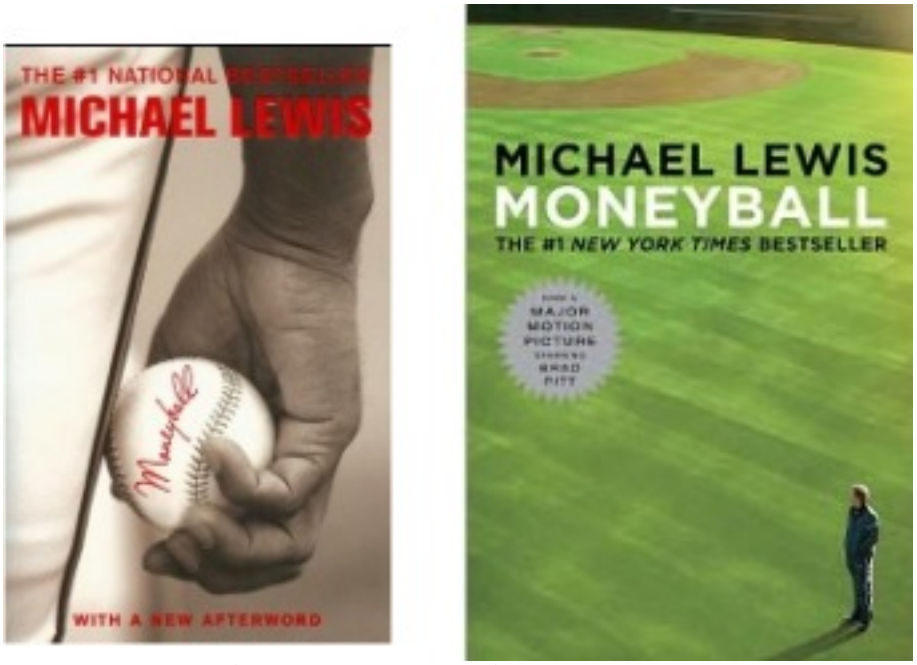

Figura 2. Capa do livro Moneyball.

3.2 — Walmart

A rede de supermercados americana Walmart foi uma das primeiras empresas a investir no Big Data (antes mesmo do termo tornar-se popular) [5]. A rede opera mais de 1 milhão de transações de clientes a cada hora, além de analisar 100 milhões de palavras-chave para otimizar ofertas, utilizando uma base diária.

O Walmart está tornando o Big Data parte de seu DNA. Somente no ano de 2012 o número de nós Hadoop de seu cluster saltou de 10 para 250 [5]. A empresa também possui um database enorme, contendo 2,5 petabytes.

Figura 3. Unidade do Walmart.

3.3 — Amazon

Depois de criar um serviço de computação em nuvem, a Amazon decidiu implementar soluções de Big Data, além de comercializá-las *as a Service*. Os motivos são semelhantes aos do Walmart: milhões de operações por dia, ½ milhão de consultas sobre produtos comercializados por terceiros, etc.

Assim como o Walmart, a Amazon utiliza Hadoop como base para seu serviço, implementando o Elastic MapReduce [6], uma virtualização com objetivo similar ao serviço de computação em nuvem (Big Data *on demand*), além de outros serviços.

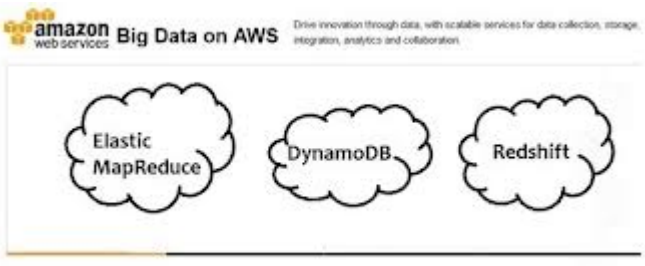

Figura 4. Big Data na AWS.

4 — Os 5 V's

O Big Data se apóia em 3 dimensões básicas: **volume, variedade** e **velocidade.** Se alguma delas estiver ausente, a solução implantada não pode ser chamada de "solução de Big Data". Além destas, algumas definições incorporam dimensões adicionais. Dependendo de quem define pode ser mais de 1 dimensão, em alguns casos chegando a mais de 3. Neste trabalho serão adicionadas as dimensões **veracidade** e **valor** [7].

4.1 — Volume

A dimensão volume é bastante clara. Todo dia são gerados petabytes de dados a uma taxa crescente. Na década passada falar em terabytes de dados pessoais era algo futurista. Hoje qualquer notebook mediano apresenta uma capacidade de armazenamento de 1 terabyte. Trabalhar com petabytes é algo normal em soluções de Big Data e já se fala em coleções contendo exabytes. Os experimentos no Large Hadron Collider representam em torno de 150 milhões de sensores, entregando dados 40 milhões de vezes por segundo [8].

4.2 — Variedade

Variedade também é uma dimensão fácil de entender. A maioria dos dados existentes no mundo vem de fontes não estruturadas (emails, redes sociais, vídeo e áudio, documentos eletrônicos, sensores e etiquetas RFID e etc.) [8]. Poucos estão estruturados em bancos de dados relacionais, com tabelas e regras bem definidas. Logo, é necessário entender como extrair e usar todos esses dados sem prejudicar a performance da aplicação.

4.3 — Velocidade

A rapidez com que dados são gerados é enorme. Com a inclusão digital e a Internet das Coisas [2], a quantidade de informações que passam por canais digitais é cada vez maior. Em muitos casos é necessário que haja resposta praticamente em tempo real [8]. Uma análise mais demorada pode não ser mais necessária ao ser entregue.

4.4 – Veracidade

É preciso que os dados façam sentido dentro do escopo do problema e que sejam autênticos [8]. Numa análise contendo grandes volumes de dados dos mais diversos tipos, não saber separar informação correta de informação errada é desastroso para uma decisão estratégica. Garbage In, Garbage Out (GIGO).

4.5 – Valor

É absolutamente necessário que a organização que esteja implementando projetos de Big Data obtenha um retorno de seu investimento. Um bom exemplo é uma companhia de seguros que minimiza os riscos de fraude ao extrapolar sua análise além da sua base de dados tradicional, minerando dados em redes sociais [8].

5 — Mitos sobre Big Data

Como o termo Big Data está popular, muitas organizações alegam trabalhar com projetos de "Análise de Big Data" apenas por inserirem um pouco de inteligência computacional em suas análises. Contudo Big Data não é apenas varrer um banco de dados de alguns terabytes e relatar que *"20% dos seus clientes são responsáveis por 80% de seus ganhos" (Princípio de Pareto)*.

A seguir serão mostrados alguns mitos comuns de mitos sobre Big Data.

5.1 — Big Data é apenas volume

Conforme dito anteriormente, o Big Data apóia-se em 3 dimensões: volume, variedade e velocidade. Ter apenas um grande volume de dados não é imperativo para um projeto ser de Big Data. Além disso, o conceito de "Big" é relativo [9]. O caso do livro Moneyball é emblemático. Existe uma quantidade de dados grande, contudo carece de variedade e velocidade.

5.2 — O enfoque é tecnológico

O enfoque é de negócios. Existe a barreira tecnológica, mas ainda assim é necessário possuir analistas especializados. É sobre agir de maneira inteligente, ter uma vantagem competitiva com um melhor entendimento das necessidades do cliente [9].

Muitos departamentos de RH em grandes empresas como Google e Facebook implementam soluções de Big Data para filtrar milhões de currículos. Fazem buscas em sites de recrutamento por termos específicos, além de analisar os perfis dos candidatos nas redes sociais. Sem o especialista, todo o aparato tecnológico é irrelevante.

5.3 — Big Data é "social data"

Big Data é "social data" para organizações cujo enfoque seja em redes sociais (Facebook, Instagram e similares) [9]. Contudo, cientistas mapeando o genoma humano ou estudando o universo não lidam com dados provenientes de redes sociais. Ainda assim, seus dados são enormes, de vários tipos e com a necessidade de responder em tempo hábil.

6 – Arquitetura

Para executar a análise de dados com alta performance, é necessário que o projeto de Big Data seja implementado utilizando uma arquitetura de computadores em paralelo e distribuídos em clusters.

A idéia por trás da quebra de um problema em vários pedaços que possam ser analisados em paralelo e depois reorganizados tem a ver tanto com requisitos temporais quanto tolerância a falhas e escalabilidade. Se várias partes do problema são independentes, podem ser computadas separadamente. Logo é melhor ter várias unidades de processamento mais simples do que uma poderosa demais, cujo trabalho é feito de forma (quase) linear. Além disso, se uma das pequenas unidades ficar indisponível, as demais conseguem redistribuir a carga de trabalho. Também fica mais fácil adicionar novos nós ao cluster do que trocar um supercomputador por outro mais potente.

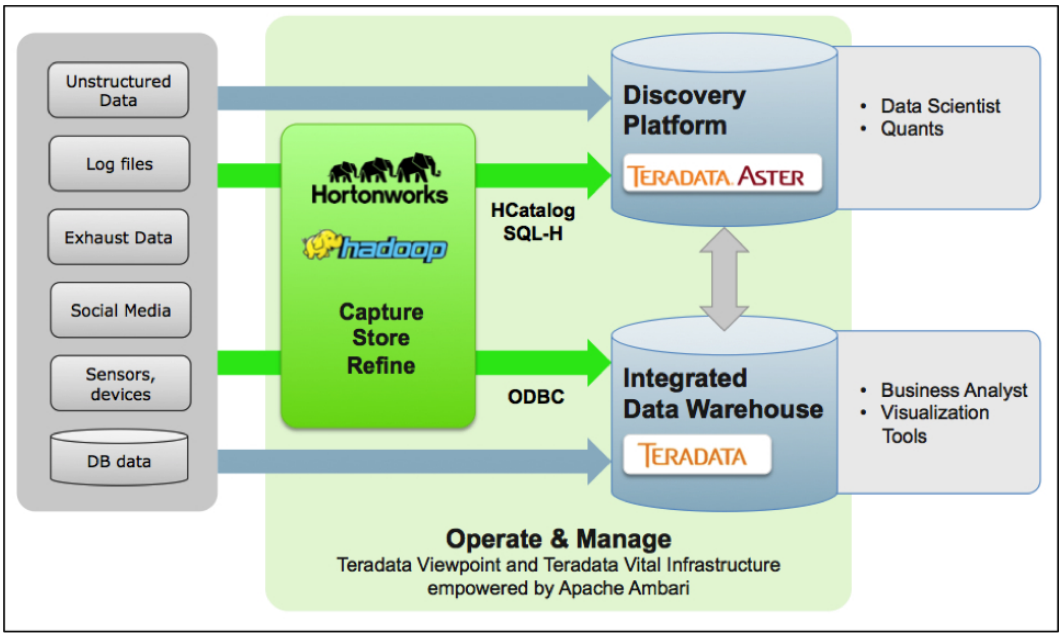

Figura 5. Visão global de uma arquitetura de Big Data.

Para poder trabalhar desta forma, é necessário que a arquitetura de uma solução de Big Data execute um algoritmo capaz de mapear as partes separáveis do problema e reduzí-lo em problemas menores. Para tanto, é necessário armazenar os dados de uma maneira cujo acesso seja menos custoso para uma arquitetura em grafo. Logo, MapReduce e NoSQL possuem um papel fundamental.

6.1 — MapReduce e Hadoop

MapReduce é um paradigma de programação para processar grandes coleções de dados utilizando um algoritmo paralelo, distribuído em um cluster [10].

É composto por 2 funções inspiradas na programação funcional: Map() filtra e separa os dados e Reduce() reorganiza os dados processados para gerar uma saída única [10].

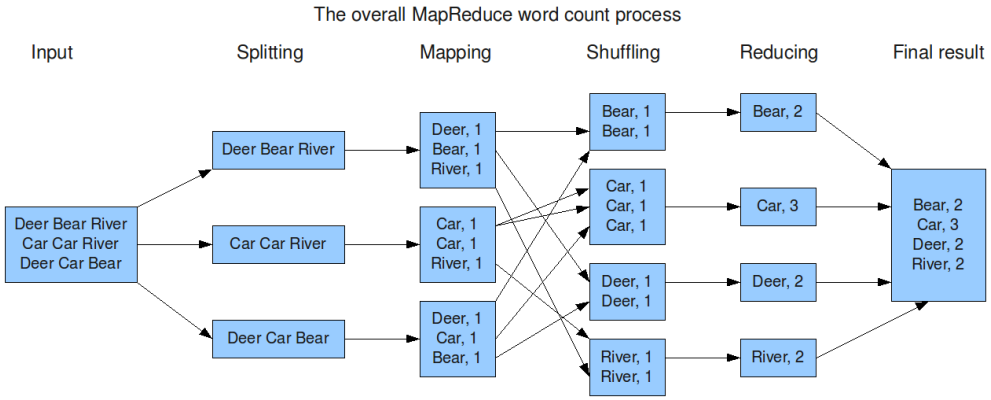

Figura 6. Visão geral do MapReduce.

O nome MapReduce originalmente referencia uma tecnologia proprietária da Google, mas com o tempo foi generalizada. Uma implementação open source bastante conhecida e popular é o Hadoop da Apache Software Foundation.

Hadoop é um framework open source para computação distribuída, escalável e e confiável [11] e é a tecnologia líder na área [2]. É baseado no MapReduce, mas esta é apenas uma parte da plataforma. O Hadoop também é composto de um sistema de arquivos distribuído (Hadoop Distributed File System — HDFS) que proporciona acesso rápido aos dados da aplicação, Hadoop YARN, um framework para agendamento de tarefas e gerenciamento de cluster, além de uma biblioteca padrão [11]. É escrito em Java, sendo multiplataforma.

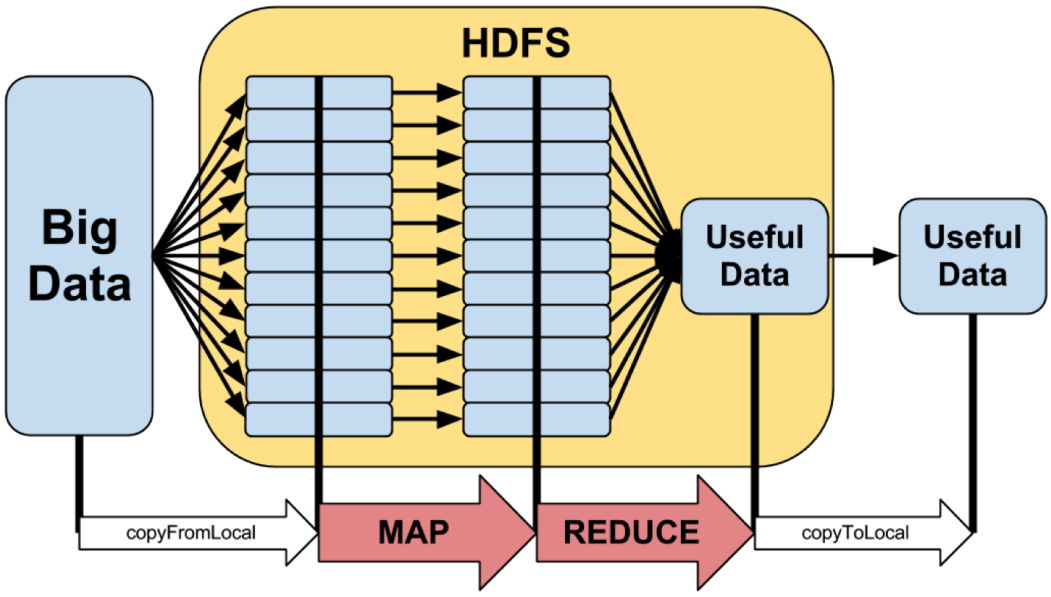

Figura 7. Fluxo simplificado do Hadoop.

6.2 – Data Warehouse e NoSQL

Para Big Data, o ecossistema gira em torno de alguns princípios básicos na formulação de sua arquitetura, incluindo armazenamento escalável, computação paralela e gerenciamento de dados [12]. Foi visto na seção anterior a tecnologia MapReduce e sua evolução, Hadoop. Ambas permitem escalabilidade ao paralelizar a computação, distribuindo-a em clusters. Foi mostrado o HDFS, o sistema de arquivos do Hadoop, que permite um acesso mais rápido e direto. Contudo, outras alternativas encontram-se no uso de Data Warehouses e bases de dados NoSQL.

Data Warehouse é uma coleção de repositórios de dados (Data Marts) contendo informações com aspecto temporal, de fontes e formatos diversos. É um auxílio para análise estatística.

NoSQL são bases de dados que vão além das tradicionais tabelas dos bancos relacionais. O nome NoSQL é um acrônimo para “Not only SQL”, significando que bancos de dados NoSQL não são excludentes a bancos de dados relacionais, mas uma expansão [13]. Existem diversos tipo de bancos NoSQL, sejam orientados a grafos, documentos ou outro paradigma que possa surgir. A vantagem de utilizar bancos de dados não-relacionais encontra-se na agilidade com a qual as informções podem ser arrumadas, além de não ter a rigidez necessária de um BD relacional. Enquanto isso facilita o acesso e a leitura dos dados, sacrifica a consistência dos mesmos [13].

7 — Desafios e Oportunidades

Por ser um paradigma novo na área de análise estatística, Big Data apresenta alguns desafios tanto de ordem tecnológica quanto de formação. Contudo, estes desafios estão atrelados a novas oportunidades que surgem para profissionais que se especializam em resolver estes novos tipos de problemas.

7.1 — Desafios

Dentre os desafios existentes para seguir carreira como um Big Data Analyst está a falta de formação especializada. Por ainda ser um termo novo, não existe um perfil bem definido. Além disso, muitas empresas que são novas no Big Data não acreditam que possam retreinar seus funcionários, pois *"a dificuldade é tamanha porque o conceito vai além dos dados armazenados na TI tradicional"* [14].

Também existe um problema físico. A quantidade de dados criados está aumentando exponencialmente, enquanto as tecnologias de armazenamento não evoluem de acordo com a mesma função [15].

7.2 — Oportunidades

Segundo a consultaria McKinsey, Big Data é *"a próxima fronteira para inovação, competição e produtividade"* [16]. Ao usá-la, uma empresa pode *"aumentar seu retorno sobre investimento de 10% a 20%, ou o equivalente a US$ 200 bilhões em todo o mundo"* [17].

Além de auxiliar executivos a melhorarem seus lucros, soluções de Big Data podem ajudar em questões de saúde pública, como por exemplo vacinação. O projeto Tycho da Universidade de Pittsburgh *"digitalizou e padronizou mais de 1 século de relatórios semanais colhidos em diversas cidades dos Estados Unidos... dando ao público e aos pesquisadores acesso a registros históricos sobre mais de 56 doenças..."* [18].

Em suma, Big Data dá a oportunidade àqueles que a usam de entender melhor problemas em larga escala. Isto não é bom somente para a economia, mas também para a ciência, saúde e educação.

8 – Críticas

Embora Big Data seja a “bola da vez”, existem algumas críticas ao seu uso. Muitas pessoas acham que não somente ter tantos dados disponíveis, mas também uma grande capacidade de análisá-los pode ser invasivo. Muitos já acham uma violação de privacidade os anúncios personalizados da Google. Além disso, torna as pessoas mais previsíveis a talvez atpe mais manipuláveis. Um político em época de campanha poderia facilmente entender o que seu eleitorado deseja e atuar de acordo [19]. Ainda na questão da privacidade, é preciso efetuar um maior controle de acesso aos dados coletados, de forma a evitar que pessoas mal intencionadas causem grandes estragos ao usá-los.

9 — Conclusão

Big Data é mais do que um conceito tecnológico, é um novo paradigma em análise estatística. Seu uso está sendo cada vez mais disseminado e tornando-se comum. Em poucas décadas técnicas de Big Data Analysis serão ensinadas em qualquer curso de Ciências da Computação, assim como Sistemas de Comunicação e Compiladores atualmente.

Entender o que significa Big Data é imperativo para qualquer pessoa que esteja iniciando uma carreira na área de tecnologia, seja informática ou não, pois este paradigma irá transformar como entendemos e afetamos a sociedade de uma maneira ainda muito bem entendida.

10 — Referências

[1] www.sas.com/offices/latinamerica/brazil/solucoes/bigdata/

[2] www.youtube.com/watch?v=7D1CQ_LOizA

[3] www.gartner.com/technology/topics/big-data.jsp

[4] en.wikipedia.org/wiki/Moneyball

[5] www.bigdata-startups.com/BigData-startup/walmart-making-big-data-part-dna/

[6] aws.amazon.com/pt/big-data/

[7] cio.com.br/opiniao/2012/05/11/o-caos-conceitual-e-os-5-vs-do-big-data/

[8] en.wikipedia.org/wiki/Big_data#Big_science

[9] www.thedatacreatives.com/2013/12/5-big-data-myths-debunked.html

[10] research.google.com/archive/mapreduce-osdi04-slides/index.html

[11] hadoop.apache.org/

[12] data-informed.com/introduction-nosql-data-management-big-data/

[13] en.wikipedia.org/wiki/NoSQL

[14] convergenciadigital.uol.com.br/cgi/cgilua.exe/sys/start.htm?infoid=34475&sid=97#.UuQwLRLNjrc

[15] www.contegix.com/big-data-comes-with-big-problems/

[16] www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[17] www.businessinsider.com/big-data-can-boost-marketing-roi-2013-11

[18] www.fool.com/investing/general/2013/12/18/3-ways-big-data-can-boost-vaccine-effectiveness.aspx

[19] info.abril.com.br/noticias/ti/big-data-ajudou-obama-a-ganhar-eleicoes-15012013-25.shl