Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

MARCOS PAULO RODRIGUES

BIG DATA: UM ESTUDO EXPLORATÓRIO

Assis
2013

MARCOS PAULO RODRIGUES

BIG DATA: UM ESTUDO EXPLORATÓRIO

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Ciência da Computação do Instituto Municipal de Ensino Superior de Assis – IMESA e Fundação Educacional do Município de Assis – FEMA, como requisito do Curso de Graduação.

Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto.

Área de Concentração: Informática.

Assis

2013

FICHA CATALOGRÁFICA

RODRIGUES, Marcos Paulo

BIG DATA: Um Estudo Exploratório / Marcos Paulo Rodrigues. Fundação Educacional do Município de Assis – FEMA – Assis, 2013.

69 p.

Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto

Trabalho de Conclusão de Curso – Instituto Municipal de Ensino Superior de Assis – IMESA

1. *Big Data* 2. Velocidade 3. Variedade 4. Volume

CDD: 001.6

Biblioteca da FEMA

BIG DATA: UM ESTUDO EXPLORATÓRIO

MARCOS PAULO RODRIGUES

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Ciência da Computação do Instituto Municipal de Ensino Superior de Assis – IMESA e Fundação Educacional do Município de Assis – FEMA, como requisito do Curso de Graduação.

Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto

Analisador:_____

Assis

2013

AGRADECIMENTOS

Primeiramente, a Deus, uma vez que, sem Ele, nenhum sonho ou objetivo é alcançado. Por meio de Sua infinita bondade a mim foi permitido ingressar e concluir esse ciclo maravilhoso de estudos.

Aos meus pais, Fernando e Waldeci, que sempre acreditaram no meu potencial e na minha capacidade de tornar concreto esse sonho maravilhoso, e a minha irmã, Tábata, que sempre foi minha conselheira, ouvindo e norteando as minhas decisões.

Ao meu orientador, o Professor Doutor Alex Sandro Romeo de Souza Poletto, que junto comigo abraçou uma ideia praticamente nova e se empenhou incessantemente na conclusão deste projeto. Além de sua dedicação ímpar em todas as fases deste trabalho, onde humildemente compartilhou seus conhecimentos, sanou todas as dúvidas que surgiam, tecendo conselhos e que por vários momentos se mostrando muito mais do que um orientador, um verdadeiro amigo.

Agradeço a todo o corpo docente da Fundação Educacional do Município de Assis, formado por professores que são modelos de ética e competência, que se empenharam ao máximo para formar os alunos em profissionais competentes e qualificados.

Agradeço também à minha namorada, Isabela Miranda da Costa, que mesmo durante o pouco tempo juntos já provou ser uma companheira maravilhosa que alimenta meus sonhos, que se alegra ao ver minhas conquistas e que compreendeu as minhas ausências para que eu pudesse me dedicar à conclusão desta pesquisa.

Por fim, agradeço aos meus amigos Fernando, Thiago, Valter, Lucas, Marcell, Vinícius Mello, Vinícius Santos, Frantchesco, Rodolfo, Emiliana, Camila, Carlos, José Américo, e tantos outros que conquistei no decorrer dessa longa jornada. Jamais esquecerei o que cada um agregou à minha formação acadêmica e principalmente aos meus valores enquanto ser humano. Mesmo que a distância venha nos privar do convívio, jamais os esquecerei.

Nenhum sucesso na vida

compensa o fracasso no lar.

Edson Itio Numazawa

RESUMO

Este trabalho tem por finalidade realizar uma pesquisa no sentido de gerar um embasamento conceitual teórico sobre a tecnologia de *Big Data*, bem como apresentar alguns exemplos de emprego de modo a evidenciar boas práticas em áreas como a saúde, educação, serviços financeiros e agricultura. Pretende-se, também, relacionar os desafios que essa nova era tecnológica enfrentará como a falta de profissionais capacitados, legislações que atribuem direitos e deveres para todos os envolvidos na cadeia de processo, competitividade comercial, entre outras. Espera-se que este trabalho, através do seu conteúdo teórico e das considerações obtidas através da observação dos exemplos de estudos de casos nele contido, possa servir de subsídio para a realização de novas pesquisas e no desenvolvimento de projetos que visam a colaborar com o desenvolvimento comercial das empresas.

Palavras-chave: *Big Data*; Velocidade; Variedade; Volume.

ABSTRACT

We intend to bring forth research providing conceptual and theoretical basis on Big Data technology as well as to present good examples of its application in areas such as healthcare, education, financial services and agriculture. We also plan to show the challenges presented by this new technology, like the lack of qualified professionals, legislation related to rights and duties concerning this process, commercial competitiveness, among others. Through its theoretical content and its considerations on the case studies we hope this work could be used to subsidize new research and other projects to contribute to the commercial development of companies.

Keywords: Big Data; Speed; Variety; Volume.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Uma nova era tecnológica está surgindo rapidamente: um cenário onde o volume de dados aumenta a cada instante e onde soluções podem ser encontradas em informações que, aparentemente, não possuiriam relação alguma e dados na casa dos *zetabytes* já podem ser considerados algo muito próximo de ser alcançado (WHITE, 2010, p.1-3).

De acordo com Costa *et al.* (2012), o processo de “digitalização” dos sistemas é o maior impulsionador do crescimento dos dados digitais, pois suas atividades (manipulação, transferências, armazenamento) geram informações.

O rápido crescimento no volume de dados despertou em grandes empresas como *Microsoft*, *Google*, *Facebook* e *Yahoo!*, um grande interesse, pois, através de pesquisas nessa gama de informações, seria possível descobrir os *sites* mais populares, os livros com maior número de *downloads*, os anúncios publicitários mais vistos e, desta maneira, criar formas mais atrativas e direcionadas ao senso comum, recrutando cada vez mais clientes (LAM, 2011, p.3).

Segundo White (2010, p.1):

> O aumento nas taxas de volume de dados é originário de diversas fontes, podendo ser observados algumas fontes com grande potencial de contribuição, tais como:
> A bolsa de valores norte-americana que gera 1 terabyte de novos dados por dia;
> O *Facebook*, detentor de aproximadamente 10 bilhões de imagens armazenadas, que ocupam 1 *petabyte* de armazenamento;
> O Ancestry.com, a maior empresa no mundo voltada a estudos genealógicos, possui um volume de dados calculado em 2,5 *petabytes* de dados;
> A *Internet Archive* (biblioteca digital da *Internet)*, detem 2 *petabytes* de dados, e possui aumenta em 20 *terabytes* por mês o seu volume de informações;
> O Grande Colisor de Hádrons (LHC) vai gerar aproximadamente 15 *petabytes* de informações por ano.

Estudos realizados pela *International Data Corporation* (IDC) em conjunto com a corporação EMC em junho de 2011 constataram que o aglomerado de informações

contidas na *Internet*, de 2005 para 2010, praticamente triplicou de tamanho, ao passo que, se esse aumento se mantiver na mesma frequência, em 2015 a quantidade de dados armazenados ultrapassará a casa dos 7 *zetabytes*. A pesquisa supracitada fica elucidada na Figura 1 (GANTZ; REINSEL, 2011).

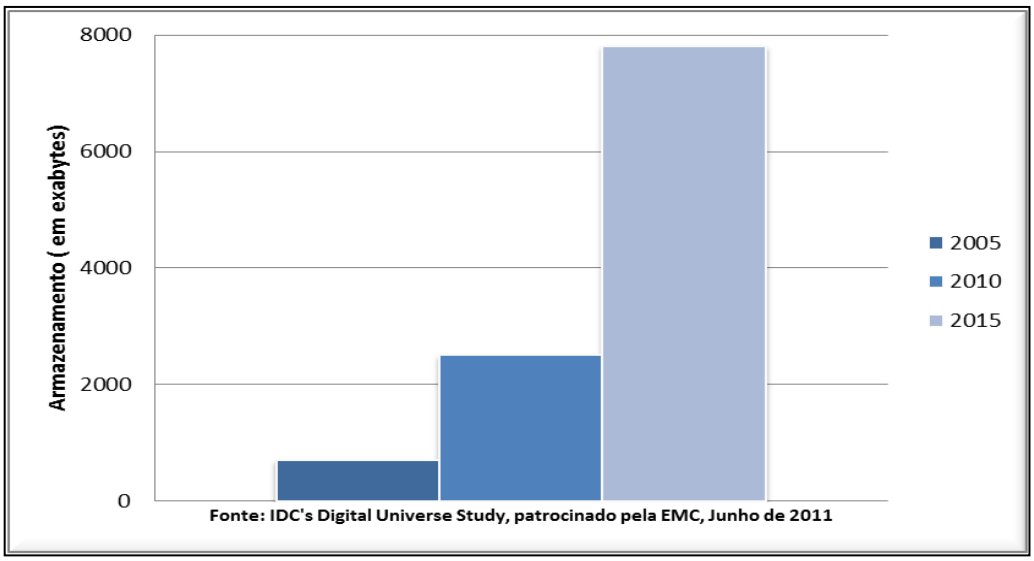

Figura 1 - Evolução da quantidade de dados armazenados na *Internet* (In: GANTZ; REINSEL, 2011)

De acordo com Goldman *et al*. (2012), surge a necessidade de um novo conceito de se pensar, pesquisar, manipular e gerar soluções para este quase imensurável montante de informações. A tecnologia *Big Data* e seus processos, então, vieram propor uma solução para o aproveitamento destes dados não relacionais, provenientes de lugares distintos, gerados por um único usuário e/ou por centenas de milhares de pessoas; essa tecnologia, em princípio, foi intitulada como uma evolução do *Data Warehouse*.

Ainda segundo Goldman *et al.* (2012), abstrair soluções para empresas e governos, mapear estilos de vida com intuito de se prever doenças e gerar padrões específicos para cada indivíduo, aumentar a segurança na geração de dados pessoais, prever catástrofes naturais, entre outras, são exemplos de utilização da tecnologia *Big Data*.

Esta nova forma de interagir com grandes quantidades de informações não se restringe apenas aos profissionais de TI (Tecnologia da Informação), mas a toda uma cadeia de setores que movimenta o mercado midiático, econômico, de lazer, saúde, entre outros. Matérias de jornais e revistas serão redigidas por jornalistas que possuam domínio na manipulação de sistemas computacionais projetados para buscas e relacionamento de informações armazenadas em grandes bancos de dados (GOLDMAN *et al.*, 2012).

Junior (2011) relata que, para trabalhar com esse novo conceito de busca informativa, será necessária a criação de um novo modelo de jornalista, um profissional que possua práticas computacionais, que consigam manipular ferramentas de pesquisas, concatenando informações diversificadas, para criar matérias fidedignas e relevantes.

1.1 OBJETIVOS

O objetivo do presente trabalho é a aquisição de um conhecimento mais específico nos processos que circundam a tecnologia *Big Data*, gerando, assim, mais conteúdo informativo, prático e didático a respeito deste novo conceito. Além da jurisprudência sobre o tema para futuras pesquisas.

1.2 JUSTIFICATIVAS

O desenvolvimento deste projeto se torna interessante pelo fato de haver pouco material para o estudo sobre *Big Data*, além de significar uma oportunidade importante de conhecer e trabalhar com uma nova tecnologia que vem ganhando interesse no mercado.

A quantidade de informações geradas a cada dia, de diversas formas e lugares distintos, por um mesmo indivíduo, possui uma imensidade de relações que passam despercebidas aos olhos do gerador. Tais informações podem ser a solução para resolver problemas em sua rotina profissional, redução de tempo em buscas por

produtos, lazer. Por isso, faz-se necessário o entendimento e o domínio de tecnologias capazes de manipular esse imenso volume de dados e nele encontrar, concatenando as informações estruturadas ou não, um padrão único e individual para aquele usuário.

1.3 MOTIVAÇÃO

O conhecimento e a manipulação de novas tecnologias, a contribuição literária de tal assunto. Além de abordar um tema que vem a cada dia se mostrando importante e fundamental na busca de novos clientes para um determinado produto, nas formas de conseguir organizar uma gama muito elevada de informações e nas soluções para problemas que, em um futuro bem próximo, possam começar a surgir com tal aumento de informações lançadas na Internet.

1.4 ESTRUTURA DO TRABALHO

O trabalho está inicialmente estruturado nos seguintes capítulos:

Capítulo 1 – **Introdução**: capítulo no qual estará exposta a ideia principal do projeto, os objetivos, justificativas, motivação e os métodos utilizados para as pesquisas e estruturação do trabalho.

Capítulo 2 – **Introdução ao *Big Data***: capítulo que contém a definição de *Big Data*, as personagens que compõem este ambiente bem como a exposição de boas práticas obtidas através do emprego da tecnologia, além dos desafios que o cenário *Big Data* encontrará.

Capítulo 3 – **Bancos de dados**: capítulo que explana o conceito de banco de dados, além de abordar as diferentes arquiteturas que são utilizadas para o armazenamento de dados, além de citar qual o tipo de armazenamento mais adequado para o *Big Data*.

Capítulo 4 – **Segurança de *Big Data***: neste capítulo serão abordadas de forma abrangente os possíveis riscos e vulnerabilidade que possa vir a surgir com a utilização dos conceitos e práticas de *Big Data*, além de expor um exemplo de utilização de *Big Data* para fins de espionagem mundial.

Capítulo 5 – **Mineração de dados**: capítulo que conceitua a mineração de dados, as tarefas que são realizadas e a harmoniosa relação que a mineração de dados possui com o *Big Data*.

Capítulo 6 – **Estudos de Casos**: neste capítulo serão apresentados alguns estudos de casos para elucidar o conhecimento proporcionado pelo referencial teórico elencado nesta pesquisa.

Capítulo 7 – **Considerações Finais**: tratará dos resultados e as conclusões sobre os assuntos que nortearam o trabalho.

Referências: parte que relaciona todas as referências consultadas durante a elaboração do projeto.

2 INTRODUÇÃO AO BIG DATA

Neste capítulo serão apresentados conceitos e noções sobre *Big Data*, as exposições de boas práticas, os desafios que a era dos grandes volumes de dados enfrentará, além de citar as personagens que compõem o ambiente *Big Data*.

2.1 CONCEITOS E NOÇÕES SOBRE *BIG DATA*

A busca por padrões comportamentais em um grande volume de dados disponíveis na rede mundial de computadores tornou-se a grande tendência que as grandes empresas adotaram para gerar lucros e aumentar seu rol de consumidores (PETRY; VILICIC, 2013, p.75).

De acordo com Petry *et al.* (2013, p.72), o *Big Data* é a forma de manipular um imenso volume de dados, oriundos das mais variadas fontes, em alta velocidade de processamento e resposta. O manuseio de um expressivo montante de informações não estruturadas, como imagens, vídeos, sms, fica inviável se para tal tarefa utilizar-se de ferramentas e técnicas que estão disponíveis no mercado e que não se norteiam nos princípios *Big Data*.

Costa *et al.* (2012), conceitua *Big Data* como um enorme volume de dados que pela inexistência, no cenário atual, de ferramentas específicas de manipulação, exige um maior esforço para o seu gerenciamento, visto que tal gerenciamento é realizado por aplicações moldadas para trabalhar com um volume menor de informações. Desta forma, para classificar um determinado bloco de dados como *Big Data*, é necessário observar se a ferramenta destinada para sua manipulação conseguirá processar a demanda de informações, de modo a alcançar satisfatoriamente o objetivo proposto.

Esta nova tecnologia veio para modificar a forma como era realizada, até então, a manipulação dos dados, que a cada instante aumenta de tamanho, substituindo a canalização para um *Data Warehouse* de um seleto conjunto de dados, pela análise na locação das informações não necessitando mais a reunião dos mesmos em um

único lugar, tarefa esta que soluciona o problema de lentidão que é gerado pela movimentação de um grande fluxo de informações. Ao adotar os preceitos *Big Data* de escalonamento de dados, a resposta é muito mais rápida e eficiente, já que eles possibilitam uma consulta às informações de maneira mais flexível, não limitando a quantidade de consultas, tampouco restringindo as formas como serão realizadas as pesquisas (COSTA *et al.*, 2012).

Algoritmos cada vez mais complexos e eficientes, capazes de encontrar padrões em acessos de músicas e fazer sugestões de outras que aquele usuário ainda não conhecia. Indicar possíveis novos amigos em redes sociais mediante uma análise minuciosa em seu rol de amigos e até mesmo relacionar em frações de segundos e gerar um relatório de um mesmo produto ou oriundo de lugares distintos para que o usuário possa realizar uma comparação, facilitando assim a compra daquele serviço ou produto (PETRY; VILICIC, 2013, p.76).

De acordo com o relatório do *World Economic Forum*, realizado em 2012, o celular vem sendo considerado a grande promessa de gerador de conteúdo para a utilização dos processos de *Big Data* na criação de novas tendências, novas formas de gerar lucros, pois o aparelho de telefonia móvel é um produto que pode ser adquirido por indivíduos de todas as classes sociais. O celular se apresenta como o produto que mais contribui para a formulação de conteúdo, ele permite que as grandes empresas e organizações consigam chegar até os níveis mais baixos de uma sociedade (WEF, 2012, p.2).

Outra ferramenta que as grandes organizações utilizam, e que médias e pequenas empresas também aderiram, é a disponibilização de produtos gratuitos nas redes sociais com o intuito de gerar, a partir de sua utilização, dados de forma voluntária gerando um perfil de consumidor, contendo suas preferências, necessidades, *hobbys*. Conteúdos esses de grande valor e fonte de matéria-prima para a geração de lucros. (WEF, 2012, p.3).

Big Data não se destina apenas às grandes empresas e multinacionais, as pequenas e médias empresas também podem se beneficiar das vantagens e facilidades que o *Big Data* proporciona aos consumidores dos seus serviços (WEF, 2012, p.2).

Serviços estes que abrangem uma gama imensurável de emprego, podendo ser utilizados na área da saúde, educação, segurança pública, qualidade de vida social, não se limitando apenas ao segmento lucrativo e comercial dos negócios e na busca por mais clientes.

Tal ideia da amplitude da aplicação dos conceitos *Big Data* fica sintetizada na página 2 do relatório do *World Economic Forum*, realizado em 2012, que elenca, entre outros, os setores da saúde, educação, serviços financeiros, agricultura, onde o emprego do *Big Data* agrega eficácia, economia e otimização de recursos.

2.2 *BIG DATA* – BOAS PRÁTICAS

Esta seção está voltada para a descrição de algumas soluções nos vários setores da sociedade onde, com o emprego das técnicas *Big Data*, ficam notórios o aumento dos resultados positivos e da otimização de recursos.

2.2.1 *Big Data* – saúde

Estudar um banco de dados alimentado com informações de saúde, onde contenha históricos de doenças em um mesmo paciente, demarcar locais aonde existam incidências de patologias, são os meios para a utilização prática do *Big Data*, com o propósito de economizar recursos e aperfeiçoar trabalhos realizados pelos agentes de saúde. Com o relatório destes levantamentos e estudos, pode-se traçar e prever surtos de doenças, elaborando campanhas de cunho educativo, conseguir relacionar enfermidades que surjam em consonância com outra que, aparentemente, não possuía relação, além da projeção de possíveis doenças que um mesmo paciente possa vir a contrair devido ao seu padrão de vida e/ou histórico médico (WEF, 2012, p.2).

2.2.2 *Big Data* – educação

Relatórios gerados a partir das técnicas de *Big Data* sobre bancos de dados com informações provenientes da educação, oriundos dos históricos de escolas públicas e privadas, universidades, históricos dos alunos, podem apontar falhas na forma de ensino, no modo como recursos estão sendo empregados, podendo ser apontado métodos mais eficientes e específicos para uma formação educacional de qualidade. O emprego das técnicas *Big Data* permite a descoberta de relações entre informações cujos vínculos entre si não poderiam ser visualizados se forem analisadas em pequena escala (WEF, 2012, p.2).

2.2.3 *Big Data* – serviços financeiros

A grande quantidade de informações digitais dos históricos de compras realizadas por meio da *Internet* permite, após analisadas, a criação de perfis dos consumidores, além de determinar hábitos financeiros, revelando formas eficientes de abordagem para a venda de linhas de crédito financeiro, mapear tipos de serviços econômicos limitando-os em regiões e zonas, possibilitando a formulação de produtos financeiros mais específicos para atender as necessidades dos clientes (WEF, 2012, p.2).

2.2.4 –*Big Data* – agricultura

Todos os processos que englobam o setor agrícola de um país, tais como aquisição de insumos e subsídios, comercialização de produtos, pagamentos, condições climáticas, entre outros, quando lançados e armazenados na área digital, possibilitam a aplicação dos processos de *Big Data*. Tal emprego pode gerar informações de grande relevância para o produtor e até mesmo para o governo, pois podem ser detectadas novas tendências na produção alimentícia e nos

investimentos. Além de gerar informações mais precisas sobre capacidade, e, consequentemente, à garantia e disponibilidade de armazenamento da produção, reduzindo os níveis de desperdício e deterioração, préstimos financeiros que se enquadram nas reais necessidades do produtor (WEF, 2012, p.2).

O acompanhamento dos resultados obtidos por meio do emprego das técnicas pode permitir ao governo realizar um rastreamento de áreas com dificuldades e com falta de incentivos. Ao detectar de forma precoce tais deficiências, torna-se mais fácil o direcionamento correto dos recursos, evitando o aumento dos índices de famílias que abandonam suas terras e, por consequência de tal ato, criam uma queda na produção agrícola (WEF, 2012, p.2).

2.3 *BIG DATA* – O CONCEITO DOS TRÊS V's

Nesta seção serão abordados os três V's que formam a conceitualização de *Big Data*: volume, variedade e velocidade.

2.3.1 *Big Data* - volume

O volume se refere à quantidade de informações digitais que são produzidas pelos usuários e/ou processos de aplicações. Esta é a característica que fica mais em evidência, pois o volume é a matéria-prima dessa nova tendência. Basicamente os processos *Big Data* são empregados com o objetivo de encontrar novos padrões e tendências em tais volumes de dados, agindo em um campo onde as ferramentas disponíveis no mercado não conseguem trabalhar satisfatoriamente com tamanha quantidade de informações (PETRY; VILICIC, 2013, p.74-75).

2.3.2 *Big Data* – variedade

Os dados que fomentam esse imenso volume de informações, em sua grande maioria não possui uma estruturação, são informações oriundas de fontes mistas de dados, como, por exemplo, fotos, músicas, mensagens de celular e eletrônicas, informações geoprocessadas, comentários em redes sociais, histórico de páginas visualizadas por meio de um navegador de *Internet*, cadeia de relacionamentos em uma rede social, ativações de leitoras de códigos de barras, entre outras. Informações geradas a partir de uma aplicação que fora executada por um usuário ou até mesmo, dados que são criados a partir de um processo automático de um *software*, essa variedade de informações podem ser alocadas em um *Data Warehouse* que será alvo do emprego das aplicações de *Big Data* e essas informações poderão possuir alguma relação entre si, gerando um novo padrão ou uma nova tendência. A variedade também é considerada um fator fundamental pela criação do *Big Data*, devido à dificuldade que ainda se enfrenta na tentativa abstrair informações de material que em sua essência não possui características semelhantes, como por exemplo, descobrir um novo padrão por meio do confronto de arquivos de musica com arquivos de imagem (PETRY; VILICIC, 2013, p.74-75).

2.3.3 *Big Data* – velocidade

A velocidade possui paridade muito grande com o grau de importância dos itens anteriores. Ela tipifica a rapidez com que as informações são criadas, selecionadas e alocadas. Atualmente, a resposta em tempo real se tornou a grande necessidade e exigência de todos os envolvidos em um processo digital, citando como exemplo o setor público, privado, as áreas de telecomunicações, os trabalhadores e clientes, entre outros. Muitos *softwares* que buscam novos padrões de informações em grandes volumes de dados não conseguem processar em tempo real as informações que possuem ciclos menores de execução por causa de sua demora na manipulação das informações. Quando se finda a execução dos procedimentos de

análise o resultado a acaba se tornando sem valor e vira descarte por se tornar um dado desatualizado (PETRY; VILICIC, 2013, p.74-75).

2.4 *BIG DATA* – DESAFIOS

Nesta seção serão expostos alguns desafios que a era dos grandes volumes de dados terá que enfrentar.

2.4.1 *Big Data* – direito do consumidor

A facilidade de aquisição de produtos e serviços digitais, somando-se com a produção quase que incontrolável e despercebida de informações via celular, *e-mail*, redes sociais, vídeo conferências, pelos usuários, torna-se necessária à criação e regulamentação de leis que defendam os direitos de privacidade e propriedade das informações particulares de cada cliente. Tal medida se pauta na preservação e no resguardo de possíveis crimes como o roubo ou utilização indevida de informações privadas. Outra medida que pode ser adotada, essa pelo poder público, é o incentivo correto e direcionado às empresas, motivando-as a fazer o uso dos dados em prol da sociedade (WEF, 2012, p.3).

2.4.2 *Big Data* – competitividade comercial

Com a imersão no mundo *Big Data*, as grandes organizações conseguirão produzir aplicações cada vez mais completas, acessíveis, eficientes e de fácil manuseio, deixando atrativa a sua compra por parte da sociedade que busca tais benefícios. Porém, a oferta de produtos tão completos cria um cenário de competitividade desnivelado, onde pequenas empresas se tornarão alvo de uma soberania por parte das grandes empresas que possuem mais recursos, criando um monopólio e prejudicando a competitividade comercial e, principalmente, o cliente. Por esse

motivo as grandes organizações que, em sua grande maioria, são do setor privado, devem ser incentivadas a trabalharem também com a produção de *softwares* com código aberto (*Open Source*), visto que, os programas *Open Source* permitem a construção de novas ideias e o aperfeiçoamento das tecnologias, além de ser uma saída para a quebra do monopólio (WEF, 2012, p.3).

2.4.3 *Big Data* – padrões de utilização dos dados

Na maioria das vezes, os dados dos consumidores são objetos de estudos de mercado sem o consentimento do mesmo e não lhe é ofertado ou informado nenhum tipo de garantia de que a sua privacidade e segurança serão preservadas.

A utilização apropriada das informações geradas pelos usuários necessitará de uma padronização análoga a de uma transação financeira eletrônica. Nos dois casos a padronização deve evoluir conforme o avanço tecnológico, com o objetivo de criar um cenário de segurança e estável ao cliente, de modo a motivar as inovações e garantir a credibilidade na venda e prestação de serviços (WEF, 2012, p.3).

2.4.4 *Big Data* – mão de obra especializada

Os processos que englobam o *Big Data* necessitam de profissionais extremamente capacitados a realizar as tarefas de mineração e análise de dados, na maioria das vezes competem tais atribuições aos cientistas e engenheiros da computação. Porém, os altos custos na contratação - e até mesmo a escassez desses trabalhadores no mercado - atrapalham os avanços na área de *Big Data*. Devido a esse empecilho, até mesmo as grandes organizações encontram dificuldades para ter acesso aos conhecimentos necessários para criar novos padrões e novas técnicas de coleta, análise, mineração de dados.

2.5 *BIG DATA* – AMBIENTE

Esta seção vai expor, segundo o relatório do *World Economic Forum*, realizado em 2012, na Genebra, Suíça, os tipos de dados, seus atores, as necessidades e os estímulos em um ambiente *Big Data*.

2.5.1 *Big Data* – Tipos de dados e personagens do ambiente *Big Data*

Realizando um estudo nos mais variados setores do ecossistema *Big Data*, podem ser observadas duas funções que possuem um papel importante para a existência desse ambiente: os mais diversificados tipos de dados e os seus personagens (WEF, 2012, p.4).

As facilidades para se adquirir e ter acesso a produtos digitais, devido ao barateamento das tecnologias, gerou um ambiente aonde um indivíduo de qualquer classe social possa exercer uma função de gerador de conteúdo. O aparelho de telefonia móvel e os serviços disponíveis pela *Internet* permitem com que os usuários colaborem para os processos de coleta de informações (WEF, 2012, p.4).

Segundo WEF (2012, p.4), outro personagem que colabora para fomentar o aumento de dados é o setor privado, pois esse setor detém um montante muito significativo de informações transacionais, além de dados coletados dos resultados de aplicações interativas com usuários, informações geradas como um subproduto de outras aplicações.

Com a utilização dos produtos criados pelas empresas privadas, o cliente, na maioria das vezes, não percebe que aquela ação está gerando conteúdo e o mesmo será utilizado pela detentora dos direitos de produção daquela tecnologia para aperfeiçoar o seu produto e construir um perfil único para aquele usuário. O setor público em vários países, da mesma forma, possui enormes volumes de dados que são extraídos de relatórios de Censos, do resultado das atividades na área da saúde, educação, economia e dos balancetes orçamentários (WEF, 2012, p.4).

A Figura 2 elucida os vários tipos de dados que compõem o ambiente *Big Data*, os interesses e necessidades dos atores neste novo cenário de dados, além dos resultados que a mineração de dados e as análises geram, quando são empregadas em um volume muito grande de dados.

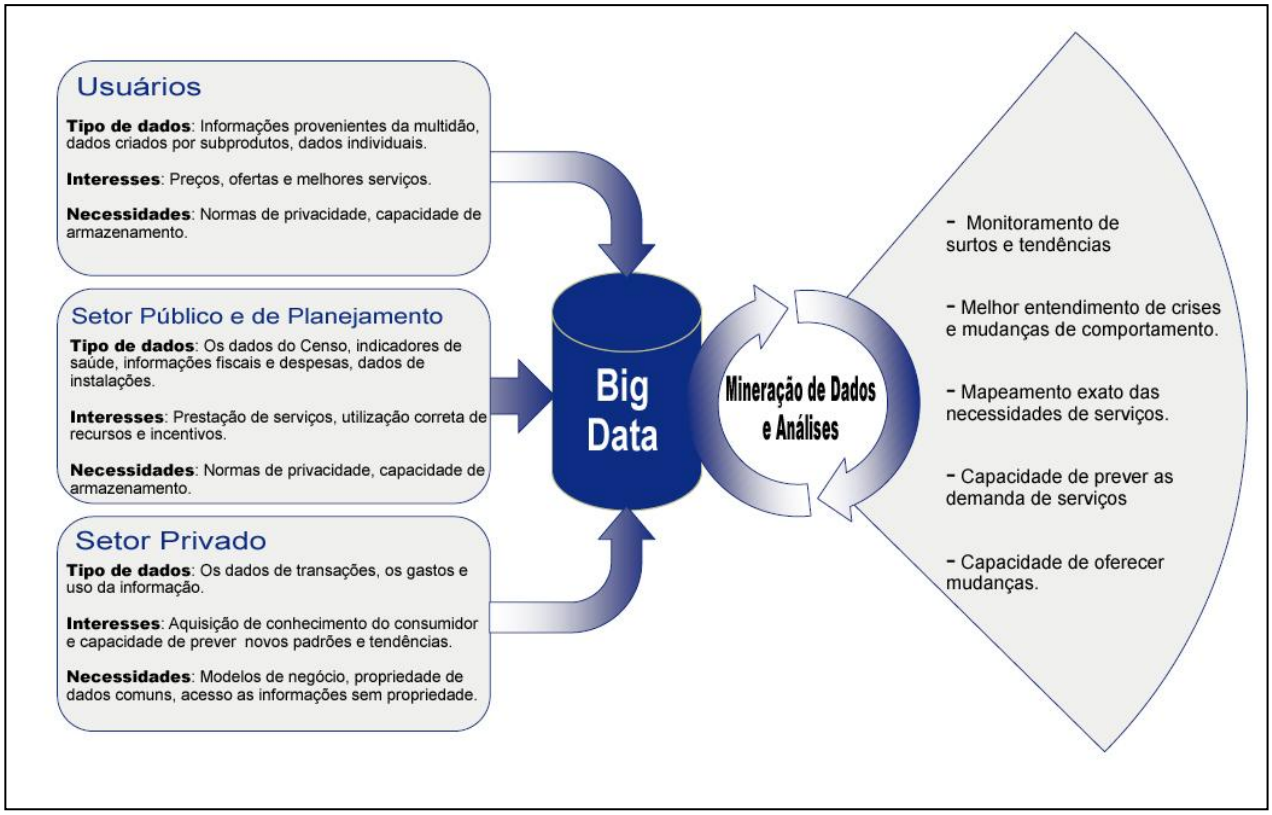

Figura 2 – Atores, interesses e necessidades do ambiente Big Data (In: WEF, 2012, p.4)

3 BANCOS DE DADOS

O termo banco de dados é utilizado para caracterizar a reunião e armazenamento de dados de forma organizada. Atualmente, quando se refere à gestão de bancos de dados, existem dois tipos distintos: bancos de dados analíticos e os bancos de dados operacionais (COSTA, 2011, p.2).

Os bancos de dados analíticos são destinados ao armazenamento de dados estáticos, informações que dificilmente serão modificadas, onde as principais tarefas serão a análise de uma tendência evolutiva das informações armazenadas permitindo a elaboração e uma possível projeção de estratégias. Geralmente este tipo de gestão de dados é consumido por organizações que se baseiam em estudos de informações coletadas por um determinado período de tempo, como laboratórios farmacêuticos, institutos de pesquisas, empresas de *marketing*, entre outros (TEOREY *et al.*, 2011).

Apesar de, habitualmente, os bancos de dados analíticos fazerem uso das informações armazenadas nos bancos operacionais, gerando um relacionamento entre ambos, a forma como as informações são processadas e estruturadas em cada um dos dois tipos de bancos são completamente específicas e incomparáveis (TEOREY *et al.*, 2011).

Bancos de dados operacionais trabalham com dados dinâmicos, informações que sofrem alterações assíduas. Este formato de banco de dados é o mais utilizado no mercado atual, onde, a frequência nas alterações e o fluxo intenso das informações são de suma importância para a manutenção e geração de lucros e expansividade dos negócios. Empresas que trabalham com transações *online*: compras, vendas, reservas e locações de produtos; são as maiores usuárias dos Bancos de Dados Operacionais.

3.1 BANCO DE DADOS – MODELOS

Atualmente existem vários modelos de banco de dados, com características e estruturas diferentes uns dos outros e nesta seção serão expostos os principais modelos de banco de dados disponíveis no mercado.

3.1.1 Modelo de banco de dados hierárquicos

Banco de dados hierárquicos, diagramado em forma de árvore, são estruturas de dados que possuem uma tabela que representa a raiz da arvore, um nó do diagrama que é caracterizado por ser uma origem comum no banco onde fora instanciada. As tabelas se relacionam entre si por meio de ligações de modo que coexista apenas a relação um-para-um e/ou um-para-muitos. Cada tabela possui uma coleção de registros, dados organizados que a tipificam. As tabelas de um banco de dados hierárquico também são classificadas como tabelas pai e tabelas filho, onde a segunda denominação representa um nó que se originou da tabela pai (SILVA, 2003, p.44-45).
Segundo Takai *et al.* (2005, p.6):

> O modelo hierárquico foi o primeiro a ser reconhecido como um modelo de dados. Seu desenvolvimento somente foi possível devido à consolidação dos discos de armazenamento endereçáveis, pois esses discos possibilitaram a exploração de sua estrutura de endereçamento físico para viabilizar a representação hierárquica das informações. Nesse modelo de dados, os dados são estruturados em hierarquias ou árvores. Os nós das hierarquias contêm ocorrências de registros, onde cada registro é uma coleção de campos (atributos), cada um contendo apenas uma informação. O registro da hierarquia que precede a outros é o registro-pai, os outros são chamados de registros-filhos.

As impossibilidades em executar relacionamentos complexos, dificulta a utilização deste modelo nos dias atuais, onde são necessários relacionamentos menos restritivos e rígidos. A Figura 3 retrata um modelo de banco de dados hierárquicos.

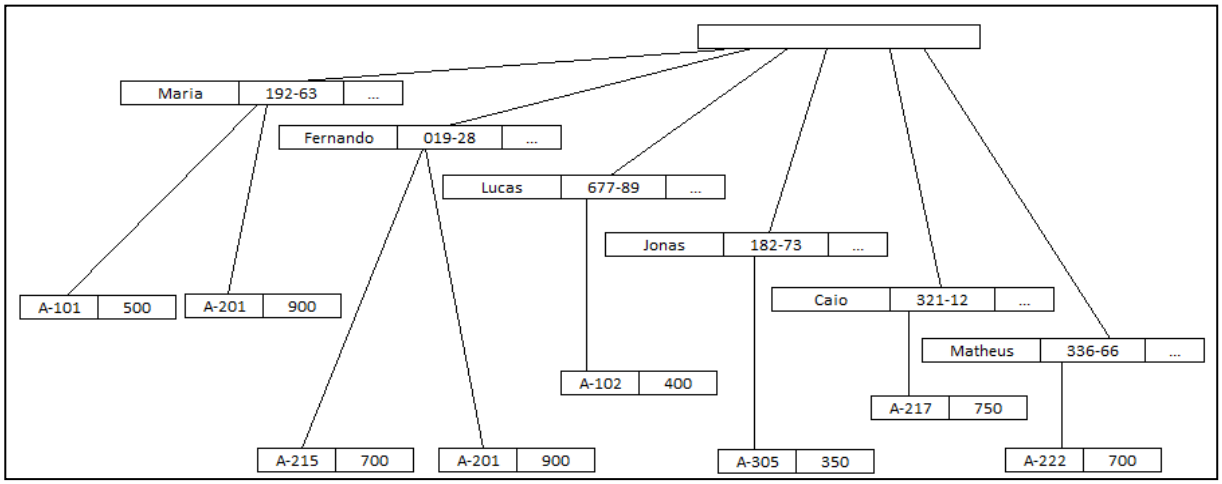

Figura 3 – Exemplo de banco de dados hierárquico (In: SILBERSCHATZ *et al.*, 2005, p.10)

3.1.2 Modelo de banco de dados de rede

O banco de dados de rede possui as mesmas características do modelo de banco de dados hierárquico, onde uma coleção de registros é associada entre si por meio de ligações. Cada registro é composto por uma associação de campos, onde cada uma contém apenas uma informação. Porém no modelo de rede, existe a possibilidade de um nó filho fazer relação com mais de um nó pai, extinguindo a ausência de relacionamentos complexos na estrutura (SILVA, 2003, p.46). Pode-se observar um exemplo de banco de dados de rede na Ilustração 4, que segue.

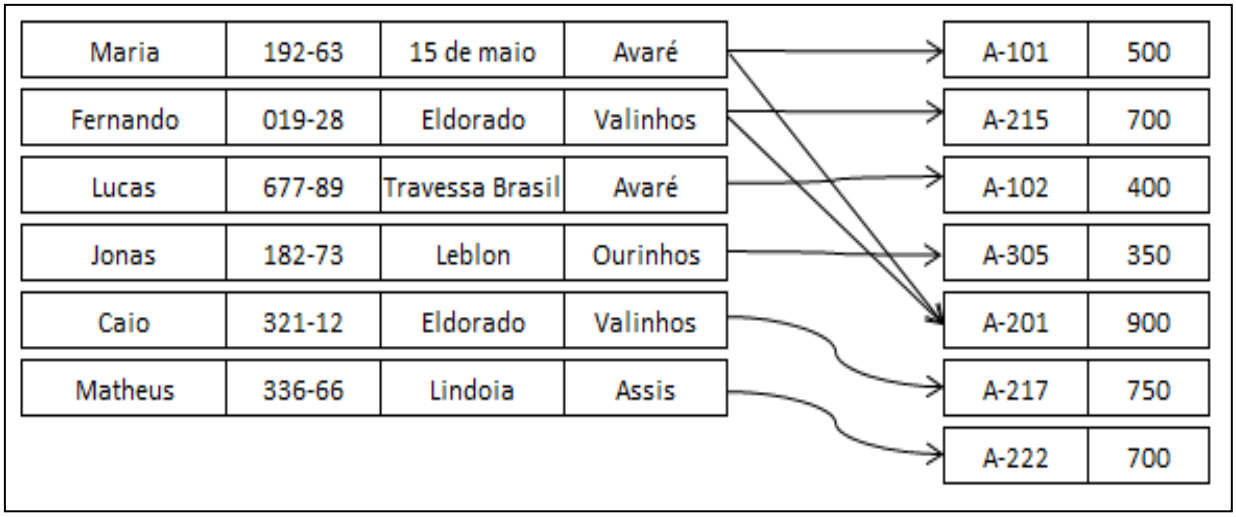

Figura 4 – Exemplo de banco de dados de rede (In: SILBERSCHATZ et al., 2005, p.10)

3.1.3 Modelo de banco de dados relacional

Este modelo de banco de dados tem sido até os dias atuais, a estrutura mais utilizada entre os principais sistemas de gerenciamento de banco de dados. A proposta de armazenamento estabelecida pelo modelo relacional dá-se pela utilização de tabelas bidimensionais, onde cada tabela é composta por linhas, que representam um dado, e colunas, que caracterizam o domínio, valores de um conjunto definido (SILVA, 2003, p.47).

Esta estrutura, também conhecida por relações está representada na Figura 5.

nome_cliente	seguro_social	rua_cliente	cidade_cliente	numero_conta
Maria	192-63	15 de maio	Avaré	A-101
Fernando	019-28	Eldorado	Valinhos	A-215
Lucas	677-89	Travessa Brasil	Avaré	A102
Jonas	182-73	Leblon	Ourinhos	A-305
Maria	192-63	15 de maio	Avaré	A-201
Caio	321-12	Eldorado	Valinhos	A-217
Matheus	336-66	Lindoia	Assis	A-222
Fernando	019-28	Eldorado	Valinhos	A-201

numero_conta	saldo
A-101	500
A-215	700
A-102	400
A-305	350
A-201	900
A-217	750
A-222	700

Figura 5 – Exemplo de banco de dados relacional (In: SILBERSCHATZ et al., 2005, p.10)

3.2 *BIG DATA* – ARMAZENAMENTO DE DADOS

De acordo com White (2010, p.3), o grande desafio na implementação de *Big Data* fica atrelado à forma de armazenamento e acesso às informações. A estagnação na evolução do processo de leitura dos dados gera um impacto negativo relevante, pois a capacidade de armazenamento evoluiu a cada dia, porém a velocidade de acesso a essas informações ainda não cresceu.

Um exemplo que elucida esse cenário se dá comparação da leitura e armazenamento de dados na década de 90 com a época atual, onde em 1991 as unidades de alocação de dados possuíam capacidade de armazenamento que girava em torno de 1 *gigabyte* com velocidade de transferência de 4,4 Mb/s, e o tempo para a leitura era de aproximadamente 5 minutos (WHITE, 2010, p.3).

Atualmente as unidades possuem capacidade de armazenar em média 1 *terabyte*, mas a velocidade de transferência não seguiu a mesma proporção evolutiva, sendo

calculada em aproximadamente 100 Mb/s, o que levaria cerca de 2 horas e meia para executar o processo de leitura de todo o *terabyte* de dados, esse tempo aumenta ainda mais quando a tarefa se torna a escrita dos dados no disco. Um desprendimento de tempo considerável diante das necessidades em tempo real que as organizações e os usuários necessitam atualmente (WHITE, 2010, p.3).

Ainda segundo White (2010, p.3), uma solução para a redução do tempo gasto nessas tarefas seria o fracionamento dessa gama de informações em unidades interligadas, assim o processo de leitura seria executado por vários discos ao mesmo tempo, trabalhando em paralelo o tempo seria reduzido para menos de 2 minutos se o trabalho de leitura dos mesmos 1 *terabyte* fosse executado por 100 unidades.

Segundo Costa *et al.* (2012), arquiteturas do tipo “nuvem computacional” (*cloud computing*), são os sistemas mais adequados para trabalhar com essa grande quantidade de dados que vem sendo gerada, pois suas estruturas são de baixo valor aquisitivo, com um grau elevado em escalabilidade, flexível e ágil, além da facilidade de execução no tocante à manipulação e ao controle de enormes quantidades de informações.

O armazenamento em nuvens, até o ano de 2020, será o método mais utilizado para alocar informações digitais. As grandes potências na era tecnológica (*Amazon*, *IBM*, *Facebook*, *Microsoft*, entre outras), vislumbrando um mercado promissor na implementação de nuvens, projetam investimentos para aumentar seus centros de dados, que operam de forma distribuída, replicando dados com o intuito de reduzir o tempo de resposta às requisições feitas pelos usuários (COSTA *et al.*, 2012).

4 SEGURANÇA DE *BIG DATA*

Segundo Petry e Vilicic (2013), atualmente cada usuário da grandiosa rede mundial de computadores possui uma vida digital, onde os eventos cotidianos do individuo, em sua grande maioria, são registrados por meio de fotografias e vídeos, depois são armazenados digitalmente, gerando um perfil digital de cada individuo, o que acarreta em uma grande vulnerabilidade e excesso de exposição da vida particular. O cuidado com os dados divulgados *online* tem que ser redobrado, já que muitas vezes estas informações estão em maior quantidade se comparadas com as informações que estão armazenadas dentro de casa (EMC, 2012).

Uma pesquisa realizada pela universidade de Cambridge conseguiu espantosos resultados na utilização de técnicas *Big Data*, relacionando informações disponibilizadas no *Facebook* com o intuito de descobrir informações omitidas. Segundo o estudo 95% dos testes, foi possível descobrir a etnia do usuário, em 88% dos resultados revelou o sexo do individuo; e a posição política e religiosa foi descoberta em 80% dos casos (PETRY; VILICIC, 2013, p.76).

A segurança em *Big Data* acaba se tornando uma vertente muito delicada, pois a ambição de conhecimento e criação de novas técnicas para manipular e até mesmo aperfeiçoar as ferramentas *Big Data* existentes atualmente no mercado causam uma vulnerabilidade acentuada, principalmente nos casos onde os códigos dos *softwares* são abertos e de fácil acesso a todos. Isso tudo resulta em um cenário de várias possibilidades para *hackers* deturparem a idéia principal do projeto, que são as oportunidades para colaboração, a troca de informação entre estudiosos no assunto, a identificação de possíveis falhas que resultam em desprendimento de tempo, fato este gerador de prejuízos, entre outras (EMC, 2012).

Por se tratar de um tema novo, as ferramentas de segurança, ainda estão em fase de aperfeiçoamento. O relatório apresentado pela *Verizon* em 2012, que aborda investigações sobre violações de dados, mostra que em mais de 90% das invasões atrasam em períodos maiores que 24 horas a evolução dos estudos sobre aquele determinado alvo do ataque, ao passo que, em quase 80% das violações levam semanas para serem detectadas. Ainda, de acordo com a pesquisa, o dinamismo no

qual estão balizados os ataques as informações dificulta as tentativas de defesas, já que elas permanecem estagnadas. Todas as peculiaridades negativas elencadas acima, unidas, retratam uma realidade de muita responsabilidade nos ambientes de TI, pois a cada instante mais processos de negócios estão compondo parte dessa montanha de dados e o fluxo de transações digitais voltadas ao comércio também vem aumentando significativamente, por causa das facilidades que a rede mundial de computadores proporciona, gerando maior rentabilidade aos negócios (EMC, 2012).

Um caso recente que elucida a falta de tecnologia voltada para a segurança de dados e a grande vulnerabilidade que as informações disponíveis na *internet* possuem, foi o caso da utilização de técnicas de *Big Data* como ferramenta de espionagem norte-americana a outros países, incluindo o Brasil.

Segundo reportagem publicada no *site* de notícias Uol, em 04 de setembro de 2013, pelo escritor Guilherme Balza, o norte-americano Edward Snowden, ex-funcionário da empresa Booz Allen que prestava serviços à Agência Nacional de Segurança dos Estados Unidos, a NSA, disponibilizou documentos de caráter secreto da agência onde revela que os Estados Unidos praticavam atos de espionagem contra outros países. Segundo os documentos apresentados, a NSA, por meio de um programa de computador, conseguia acessar *e-mails*, *chats online*, entre outras informações digitais armazenadas na *internet* e, de posse dessas informações, conseguia descobrir possíveis ataques terroristas contra os EUA.

Porém, o trabalho não se restringiu apenas em se detectar possíveis atos de terrorismo contra os EUA mas também foi revelado que, com as informações obtidas por meio do emprego do programa espião, os Estados Unidos coletavam todos os dados digitais criados por cidadãos de todas as partes do mundo, inclusive de chefes de Estado. No caso da interceptação das informações dos líderes políticos mundiais, a ação colocou os americanos em uma zona de hegemonia sobre as outras nações, pois, com os resultados obtidos da espionagem, os norte-americanos conseguiram traçar os perfis de cada líder, descobrindo como pensam, com quem se comunicam, o que planejam e suas informações empresariais. Esse conhecimento permite um controle muito grande, pois, quando se sabe o que os

outros líderes de nações estão fazendo e o que estão planejando, facilita-se o domínio sobre o espionado.

Figura 6 – Logotipo da Agência Nacional de Segurança dos Estados Unidos da América (NSA)

O repórter Mauricio Grego, em sua publicação no *site* da revista Exame, no dia 11 junho de 2013, relata que a Agência Nacional de Segurança dos Estados Unidos utiliza um programa denominado PRISM e que, por meio desse *software*, consegue ter acesso direto e irrestrito ao conteúdo dos servidores da *Verizon Communication*, uma empresa do ramo de telecomunicações que fornece seus serviços para usuários de todos os continentes do mundo. Este acesso permite o rastreamento de conversas telefônicas privadas, interceptação de mensagens eletrônicas dos clientes que possuem assinatura de serviços fornecidos pela *Verizon*.

Figura 7 – Logotipo da *Verizon Communication*

No entanto, o programa não se restringe apenas às informações contidas nos servidores da *Verizon*, mas também dá acesso a todos os dados que possam ser acessados por meio da *Internet*.

Esse controle e manipulação de tamanha quantidade de dados só são possíveis devido à criação de uma infraestrutura gigantesca de captação e processamento de dados, onde são empregadas as técnicas de *Big Data* (GREGO, 2013).

O jornal *The New York Times*, em matéria divulgada em 09 de junho de 2013 pelos repórteres James Risen e Eric Lichtblau, relata que o governo norte-americano investe bilhões de dólares na expansão da NSA, custeando inclusive a construção de um gigantesco *Data Center* na cidade de Bluffdale, no estado americano de Utah. Uma mega construção de 93 mil metros quadrados, com seu custo estimado em dois bilhões de dólares. O interesse em encontrar novos padrões em montanhas de dados não estruturados, fez com que os EUA utilizassem inclusive o supercomputador da *IBM*, o Watson, devido o seu grande poder de manipulação de informações não estruturadas, em testes para localizar possíveis padrões digitais de ataques terroristas.

Figura 8 – Watson, supercomputador da *IBM*

Existem duas principais razões que justificam a necessidade da NSA em possuir um grande poder de captação e processamento de informações, onde a primeira está pautada na interceptação indiscriminada de informações produzidas pelos usuários

convencionais da *internet*. O segundo motivo, é que grande parte das informações que circulam na rede estão criptografadas. Dados de transações financeiras, mensagens eletrônicas, trafegam cifrados na *internet*. Por isso a necessidade de computadores cada vez mais potentes em análise e processamento de informações, máquinas capazes de quebrar a criptografia e descobrir o significado dos dados que estão sendo alvo da análise (RISEN; LICHTBLAU, 2013, p.A1).

5 MINERAÇÃO DE DADOS

A Mineração de dados ou *Data Mining* é uma conceituação do ato de utilização de processos computacionais para abstrair informações, até então desconhecidas, em uma gama de grandes proporções de dados.

Existem ainda outras definições literárias sobre mineração de dados que ajudam a elucidar o conceito, tais como:

> A ação de identificar e abstrair algum tipo de conhecimento que, até então, não havia sido observado, de grandes volumes de dados é denominado mineração de dados (NAVEGA, 2002).
>
> A mineração de dados pode ser definida como um aglomerado de técnicas automáticas exploratórias em grandes volumes de dados que resulta no descobrimento de novos padrões e interações, que dificilmente seriam observados e abstraídos a olho nu devido ao enorme volume de informações (SILVA, 2006, p.16).

Fayyad *et al.* (1996) descreve a mineração de dados como uma forma de reconhecer padrões novos em grandes massas de dados, padrões estes de suma importância e altamente relevantes para gerar subsídios para investimentos novos, correções de falhas, opções estratégicas. A mineração de dados se utiliza de algoritmos que realizam o processamento dos dados e descobrem esses novos padrões.

Mineração de Dados são técnicas computacionais utilizadas no descobrimento de novos padrões e correlações entre os dados gerados por uma organização. As técnicas consistem em realizar análises em todas as informações armazenadas nos *Data Warehouse* em busca de novos padrões, novas estatísticas e regras de negócio (NIMER; SPANDRI, 1998, p.32).

Data Mining está diretamente relacionado com a metodologia empregada no processamento das informações, pois é com a aplicação desse procedimento que nos torna possível assimilar novos padrões de informações, além do surgimento de um novo conhecimento processual (GARCIA, 2008, p.20).

Alguns autores descrevem os processos que regem a mineração de dados de formas distintas, porém, suas definições compartilham da mesma essência. A seguir estão relatadas duas definições das fases que constituem o *Data Mining*:

Os propósitos iniciais de *Data Mining* estão pautados em duas etapas: a **predição**, onde é realizada uma seleção de campos específicos em um banco de dados e a partir dessa delimitação de informações, se busca prever novos padrões, valores desconhecidos ou interações com informações de interesse para a organização; a **descrição** é a fase onde se procura traçar um modelo de informação por meio dos novos padrões encontrados na etapa de predição. Para essa caracterização do novo modelo a fase de descrição utiliza algoritmos capazes de processar e descobrir as informações profícuas contidas em uma base de dados (GARCIA, 2008, p.20).

A mineração de dados pode ser dividida entre duas etapas, limpeza e seleção de dados, onde a primeira etapa consiste na garimpagem dos bancos de dados, limpeza e remoção de todos os resquícios de ruídos, as redundâncias, os dados corrompidos, gerando um repositório organizado. Após, inicia-se a próxima etapa onde os dados que compõem os repositórios organizados passam pelo processo de seleção, um refinamento no intuito de descobrir novos padrões (NAVEGA, 2002).

Ao realizar uma descrição dos processos que circundam a Mineração de Dados se destacam nas atividades preditivas a classificação e a regressão (estimativa) e nas atividades descritivas os processos que possuem um valor essencial para o *Data Mining*, os processos de associação e sumarização (*Clustering*) (GARCIA, 2008, p.21).

As tarefas de Mineração de Dados estão retratadas na Imagem 9.

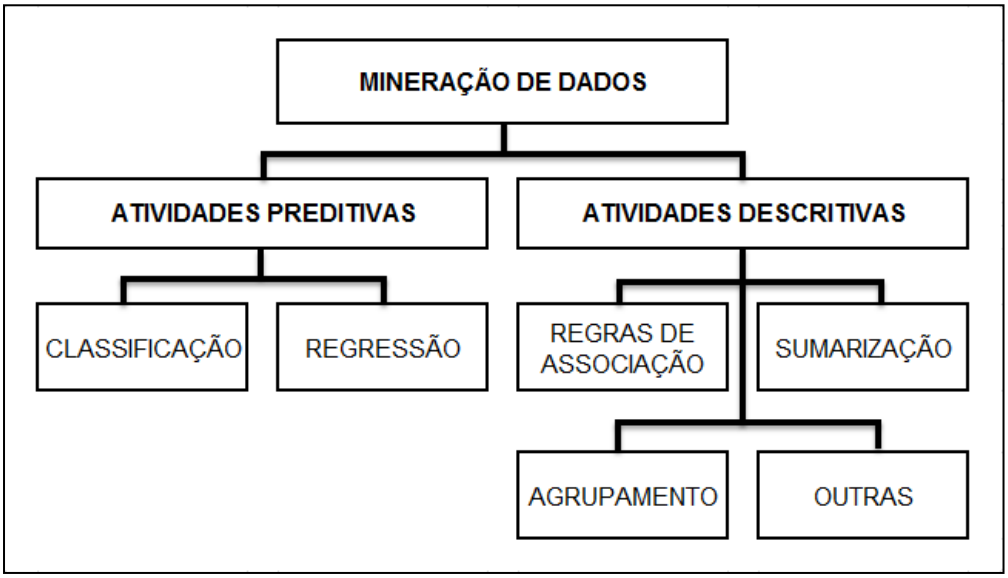

Figura 9 – Organograma das tarefas de Mineração de Dados (In: SILVA *et al.*, 2007, p.12)

5.1 MINERAÇÃO DE DADOS – TAREFAS

Nesta seção serão descritos alguns dos processos fundamentais que compõem a Mineração de Dados.

5.1.1 Mineração de dados – classificação

O processo de classificação se pauta na investigação dos dados, concatenação de informações que possuam relações e criação de um modelo de classe de dados, facilitando a análise das informações que antes se apresentavam dispersas no banco de dados e agora estão agrupadas por características em comum. A tarefa de classificação está enquadrada na classe de atividades preditivas, pois após a tipificação de um modelo de classe de dados, pode-se, de forma automática, detectar novos padrões que não se enquadram nas classes já criadas, gerando novas classes de dados. As redes neurais, árvores de decisão, algoritmo genérico,

são alguns exemplos de procedimentos que realizam o processo de classificação em uma base de dados (GARCIA, 2008, p.21).

5.1.2 Mineração de dados – associação

A tarefa de associação é um conjunto de regras da classe das atividades descritivas, que possui a incumbência de confrontar padrões de classes de dados e estabelecer uma correspondência entre as classes que já estão organizadas no *Data Warehouse*. O resultado dessa associação é a descoberta de novos padrões de dados, além da possibilidade de compreender e visionar o comportamento dos novos dados (TWO CROWS, 2005, p.10).

5.1.3 Mineração de dados – regressão

O processo de regressão se enquadra nas atividades preditivas, pois essa tarefa realiza a função de antever quais valores serão adotados pelos novos dados, tudo isso utilizando-se do conhecimento adquirido das formas já existentes de valores. A regressão trabalha com técnicas de estatísticas simples, como a regressão linear. Contudo, alguns cenários de dados não se enquadram em um padrão linear, sendo necessária a implantação de técnicas mais complexas para prever valores futuros. Redes neurais, regressão logística e árvores de decisões são alguns exemplos de técnicas que possuem um grau de dificuldade maior e que requerem mão de obra específica para o seu manuseio (TWO CROWS, 2005, p.10).

5.1.4 Mineração de dados - sumarização

A sumarização é a tarefa que realiza a divisão do banco de dados em diversos grupos de informações similares. Com o intuito de encontrar grupos que,

comparados em uma visão micro, possuam informações muito semelhantes, porém com uma notável diferença entre eles, se observados em uma visão macro. Diferentemente da classificação, a sumarização não permite prever o que será gerado ou quais atributos de dados serão agrupados. Por isso, a sumarização está relacionada nas atividades descritivas (TWO CROWS, 2005, p.6).

6 ESTUDOS DE CASOS

Neste capítulo serão apresentados alguns estudos de caso com base no sistema *Hadoop*, uma ferramenta *Big Data* para otimizar recursos e aumentar a eficiência dos serviços prestados e os lucros, iniciando com uma introdução sobre a ferramenta *Hadoop*.

6.1 INTRODUÇÃO AO *HADOOP*

Segundo White (2010, p.9), o *Hadoop* é um sistema escrito na linguagem *Java*, com seu código fonte aberto, desenvolvido para armazenar e processar grandes quantidades de dados, os *Big Data*.

Criado por Doug Cutting em 2006, inspirado no Apache *Nutch*, um motor de buscas na *internet* também de código aberto, desenvolvido para localizar informações na *web* e gerar relatórios de descoberta da informação solicitada pelo usuário por meio de uma palavra-chave, uma informação específica que o usuário do *software* necessita de informações. Porém o *Nutch* possuía uma arquitetura que não permitia uma escalabilidade para comportar os bilhões de páginas que a *internet* possuía (WHITE, 2010, p.9).

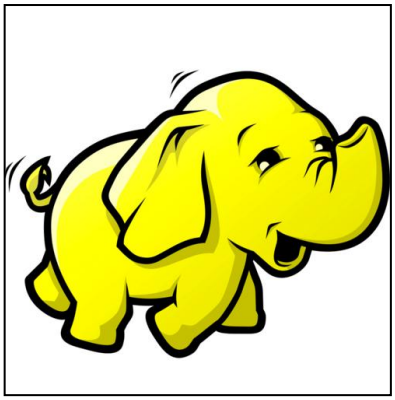

Figura 10 – Logotipo do *Hadoop*

Em 2003, foi encontrada em um artigo a solução para o problema de escalabilidade que o *Nutch* possuía. O artigo relatava a arquitetura de um sistema de arquivos distribuídos que fora desenvolvido pelo *Google*, denominado GFS (*Google File System*), destinado à utilização própria e com um dos seus objetivos pautados da redução de tempo gasto no gerenciamento dos nós de alocação de dados. O acesso à descrição da arquitetura do GFS permitiu aos idealizadores do *Nutch* desenvolver em 2004 uma nova metodologia para aprimorar os processos de armazenamento dos grandes volumes de dados que eram gerados a partir da execução do *Nutch*, que seriam os resultados dos processos de busca e indexação de informações. Essa implementação, que foi baseada no GFS, recebeu o nome de *Nutch Distributed Filesystem* (NDFS) (WHITE, 2010, p.9).

Figura 11 – Logotipo do projeto *Nutch*

No mesmo ano, a *Google* publicou um artigo se referenciando ao *MapReduce*, um novo esquema de programação, que manipula grandes volumes de dados de forma paralela, realizando uma fragmentação dos processos em grupos independentes (WHITE, 2010, p.10).

Os desenvolvedores *Nutch*, no começo de 2005, já possuíam uma implementação de *MapReduce* que era executada em ambiente *Nutch*, sendo que seus principais algoritmos já haviam sido migrados para serem executados com o *MapReduce*. A junção de NDFS e *MapReduce* alcançou resultados tão significativos, de modo a ultrapassar os limites que eram almejados pelo projeto, desta forma em 2006 foi criado um novo projeto denominado *Hadoop* (WHITE, 2010, p.10).

Esta evolução está representada na Figura 12.

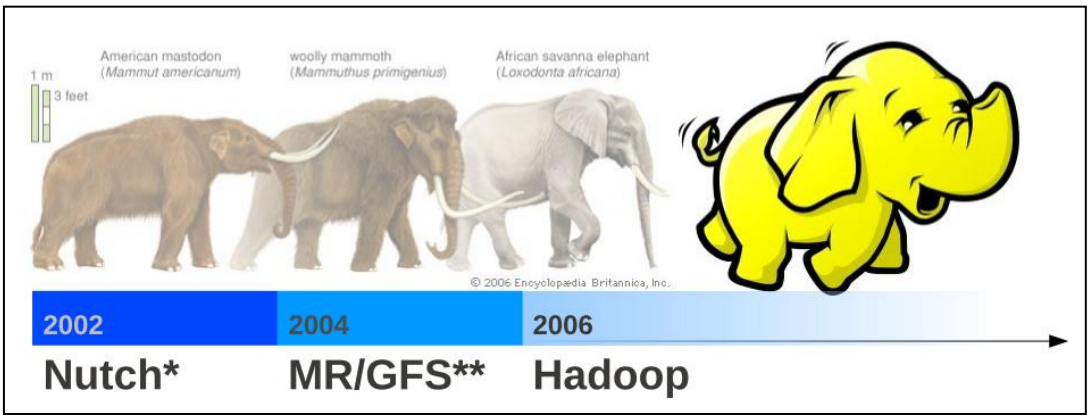

Figura 12 – Evolução do *Hadoop* (In: CORDEIRO; GOLDMAN, 2012, p.8)

Segundo White (2010, p.10), o projeto *Hadoop* ganhou força em 2006, quando Doug Cutting se juntou ao *Yahoo!*, onde foi disponibilizado para o desenvolvimento do projeto um grupo de profissionais altamente capacitados, com o objetivo de transformar o *Hadoop* em um sistema com capacidade de execução de dados a nível *web*.

O objetivo foi alcançado e o *Hadoop* começou a ser utilizado por várias organizações de grande expressividade e dos mais variadas setores de atuação na sociedade, tais como jornais, rádios, empresas de desenvolvimento de *softwares*, redes sociais, universidades, lojas virtuais, entre outros (WHITE, 2010, p.10).

A figura 13 relaciona algumas grandes empresas que, atualmente, utilizam o *Hadoop* como solução *Big Data*.

Figura 13 – Organizações que utilizam o *Hadoop*

Nas próximas subseções, serão mostrados os estudos de caso realizados pelo Jornal *The New York Times*, pela *China Mobile Communication Corporation* (CMCC) e pela *web* rádio *Last.fm* , cenários nos quais o *Hadoop* foi empregado como solução *Big Data.*

6.2 UTILIZAÇÃO DO *HADOOP* NA BASE DE INFORMAÇÕES DO JORNAL *THE NEW YORK TIMES*

A diretoria do jornal *The New York Times* decidiu, em 2007, converter todo seu arquivo de publicações para o formato PDF (*Portable Document Format*) e armazená-lo em um banco de dados do tipo *cloud computing*, onde a escalabilidade do banco fosse melhor do que a que estava em uso pela organização (LAM, 2011, p.267).

Prevendo que a realidade do ambiente em utilização não supriria as necessidades dos usuários que, com a disponibilização de toda a base de dados do jornal,

passariam a executar mais solicitações, aumentando o tráfego de dados. A primeira dificuldade a ser transpassada era a de que as publicações mais antigas, principalmente as edições publicadas entre os anos de 1851 e 1922 estavam digitalizadas e arquivadas em formato TIFF (*Tagged Image File Format*) e separadas por artigos. Desta forma, seria necessário adotar uma técnica eficiente em processamento de imagens para combinar diferentes arquivos em uma única edição do jornal e arquivá-lo em PDF (LAM, 2011, p.267).

A ideia inicial foi realizar uma conversão de todos os arquivos TIFF em formato PDF, utilizando-se de um *software* que o jornal já possuía, tratando todas as imagens como um único lote de informações e não individualmente. Concluída a etapa de conversão, a fase seguinte seria a de manipulação dos dados, atribuindo à mesma forma de manipulação que as técnicas já existentes possuíam para trabalhar com dados estáticos. Foi aí que se observou que o volume de dados era muito grande, contendo 11 milhões de artigos que totalizavam quatro *terabytes* de informações (LAM, 2011, p.267).

Diante deste cenário, um programador do *Times*, Derek Gottfrid, vislumbrou uma ótima oportunidade de utilizar dois sistemas para cumprir o determinado pela diretoria do jornal: o S3 (*Amazon Simple Storage Service*) da *Amazon*, para o armazenamento *web* e o *Hadoop* para realizar os processos de manipulação dos dados (LAM, 2011, p.267). A Figura 14 ilustra o cenário proposto pela diretoria do jornal *The New York Times*.

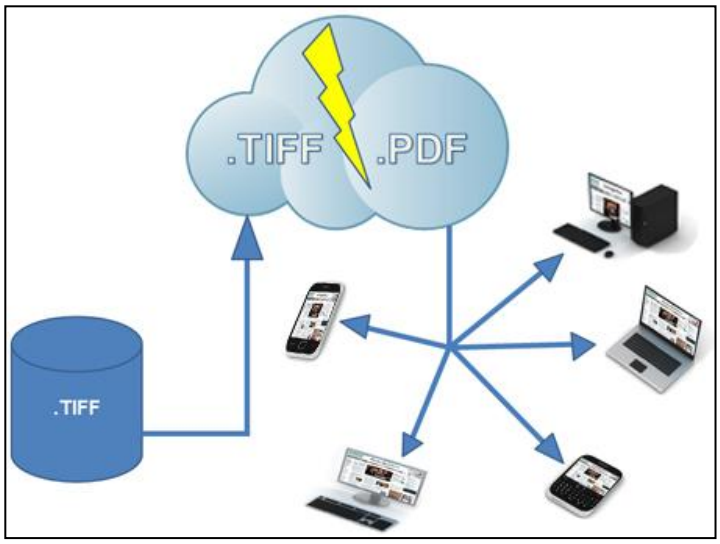

Figura 14 – Migração dos arquivos para *cloud computing*, conversão para arquivos PDF e disponibilização do conteúdo para os clientes

Derek iniciou a migração das imagens TIFF para o *Amazon S3* e logo após deu início à construção de um algoritmo para realizar a combinação das imagens, o descarte das redundâncias e gerar os arquivos no formato PDF. O algoritmo elaborado foi implementado para ser executado no ambiente *Hadoop*. A execução foi realizada em 100 nós virtuais, alocados da *Amazon* EC3 (*Elastic Compute Cloud*). A tarefa foi realizada em 24 horas, gerando 1,5 *terabyte* de arquivos do tipo PDF. Cada nó foi alocado pelo valor de U$ 0,10 gerando um gasto em computação estimado em U$ 240,00 (24 horas X 100 nós X U$ 0,10), tendo em vista que não foi computado o valor da alocação do S3, já que esta era uma das medidas a serem adotadas independente do meio utilizado para a conversão das imagens TIFF em arquivos PDF. A transferência dos dados que estavam distribuídos nos nós EC3 para o S3 também não gerou ônus, por se tratar de uma tarefa grátis, de modo que a execução do *Hadoop* não gerou gastos com aumento de banda (LAM, 2011, p.267).

A tarefa que desprendeu os serviços de apenas um funcionário possibilitou ao rol de clientes do *The New York Times* terem acessos rápido e fácil a todo o conteúdo publicado pelo jornal (LAM, 2011, p.267).

6.3 UTILIZAÇÃO DO *HADOOP* PELA EMPRESA *CHINA MOBILE COMMUNICATION CORPORATION* (CMCC)

No segmento de telefonia móvel a *China Mobile Communication Corporation* (CMCC) detém o posto de maior operadora do mundo. Possui mais de dois terços do mercado na China, disponibilizando seus serviços para mais de 500 milhões de clientes. Uma ligação telefônica gera inúmeras informações, denominadas de Registro de Dados de Chamada (CDR), que são elas: o número de identificação do cliente que está realizando a ligação, o número de quem irá receber a chamada, os horários de início e término da ligação, o tempo de duração, as informações acerca do roteamento da ligação, entre outros. Além do CDR, existem outras informações que também são geradas, as chamadas informações estruturais, dados produzidos dentro da rede, como transferências entre os diversos *switches*, terminais e nós. O tráfego de informações na rede móvel de telecomunicações chinesa é muito grande, a média de dados CDR criados por dia varia de 5 a 8 *terabytes* (LAM, 2011, p.268).

Com o intuito de descobrir novos padrões para melhorar a rede, os serviços e o direcionamento correto do *marketing*, a *China Mobile Communication Corporation* começou a investir na mineração de seus dados, realizando análises no comportamento de seus usuários, fazendo a manutenção do relacionamento empresa X cliente, além do descarte de mensagens do tipo *spam*. Porém, as ferramentas de mineração de dados que a CMCC tinham à disposição no mercado eram limitadas, pois só executavam as tarefas em um único local, obrigando a estocagem dos dados em um único banco de dados. Esta condição criava um gargalo de processos e o desempenho na obtenção dos resultados se tornava muito lento (LAM, 2011, p.268).

Foi então que a empresa decidiu realizar um projeto experimental para construir um aplicativo minerador de dados definido para funcionar em ambiente *Hadoop*. O trabalho que funcionava de forma paralela, foi intitulado de *Big Cloud – based Parallel Data Mining* (BC-PDM). Os motivos para a utilização do ambiente *Hadoop*

se pautaram na notória escalabilidade que sua arquitetura permite por ser um *software* livre, fácil de customizar, além das facilidades em sua manipulação (LAM, 2011, p.269).

O projeto implementou nove algoritmos de mineração de dados e realizou diversas operações com os dados, tais como o descarte de redundância, processamento de atributos, a realização de amostragens dos dados, e outros. Os algoritmos foram divididos em três categorias: o agrupamento, a classificação e a análise de associação. O projeto foi executado em um *cluster Hadoop* de 256 nós conectados em um único comutador (LAM, 2011, p.269).

Como os dados que subsidiaram o experimento possuíam um tamanho muito grande, tornar-se-ia difícil de observar os resultados em processos de escalas menores, desta forma, o arquivo de dados foi dividido em três escalas, conforme representação na Tabela 1.

TIPOS DE DADOS	LARGA ESCALA	MÉDIA ESCALA	PEQUENA ESCALA
Comportamento de usuários	12 Terabyte	120 Gigabyte	12 Gigabyte
Acessos de usuários	16 Terabyte	160 Gigabyte	16 Gigabyte
Associações de novos serviços	120 Gigabyte	12 Gigabyte	1,2 Gigabyte

Tabela 1 – Divisões dos dados em três escalas (In: LAM, 2011, p.269)

Os resultados do sistema atual foram comparados com um *cluster Hadoop* de apenas 16 nós. A primeira etapa de comparação foi relacionada ao tempo desprendido para realizar as operações de ETL (*Extract Transform Load*) que são: a extração, a transformação dos dados extraídos em regras de negócio e carregamento destas informações no banco de dados (comparação representada na Figura 15); e o tempo gasto nos processos de mineração de dados (representada na Figura 16) (LAM, 2011, p.270).

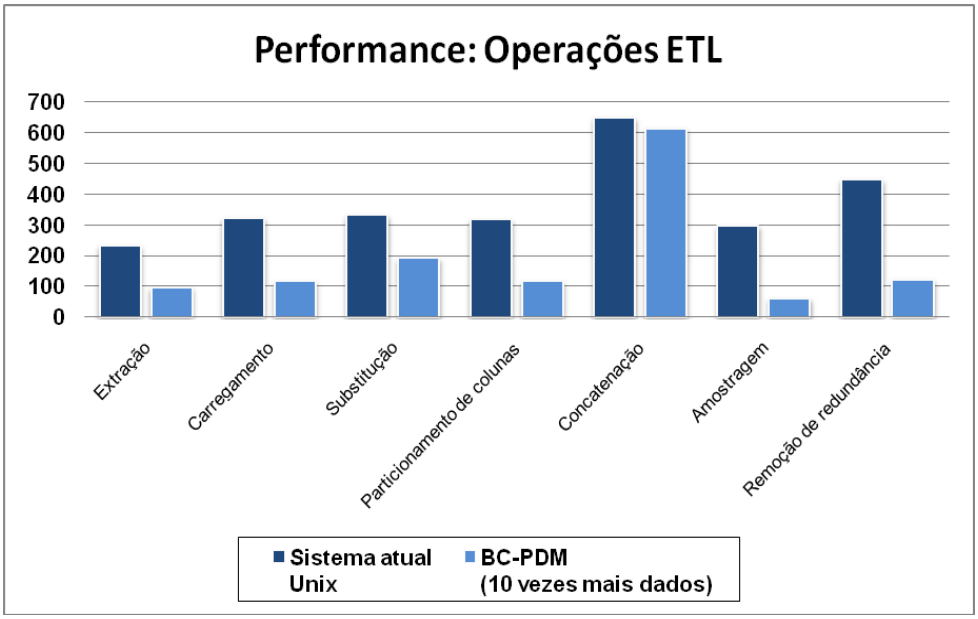

Figura 15 – Comparação de tempo gasto nas operações de ETL entre o sistema *Unix* e o sistema PC-PDM com 10 e 100 vezes mais dados que o sistema *Unix* (In: LAM, 2011, p.271)

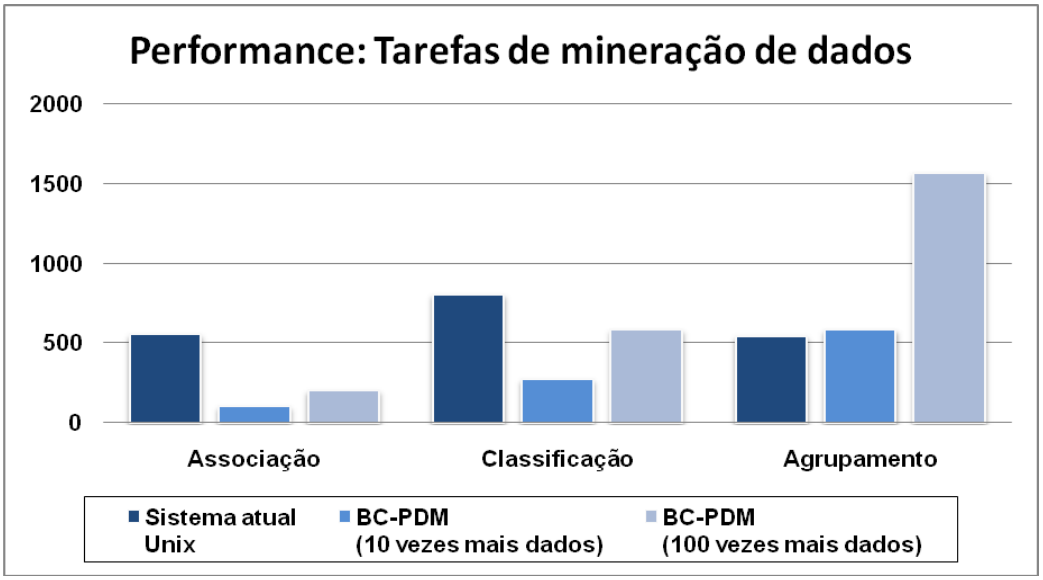

Figura 16 – Comparação de tempo gasto nas tarefas de mineração de dados entre o sistema *Unix* e o sistema PC-PDM com 10 e 100 vezes mais dados que o sistema *Unix* (In: LAM, 2011, p.271)

Para realizar o estudo comparativo o BC-PDM trabalhou uma quantidade de dados 10 vezes maior que a quantidade manipulada pelo sistema atual. Na comparação das operações ETL, o desempenho do BC-PDM foi superior em todos os processos, observando uma melhora calculada em média 15 vezes melhor do que o sistema em uso. Já na comparação dos processos de mineração de dados, o BC-PDM foi testado com 100 vezes a quantidade de dados que o ambiente atual (LAM, 2011, p.270).

Lam (2011, p.270) relata que, mesmo com a diferença notória dos volumes de dados, o BC-PDM foi mais eficiente do que o sistema atual nos processos de associação e classificação. Perdendo apenas na comparação do processo de agrupamento.

A segunda etapa de comparação abordou o custo em *Yuan* (CNY – moeda chinesa que, comparada com o dólar, possui um valor 6 vezes menor), entre o sistema atual e o *cluster Hadoop*/BC-PDM de 16 nós (LAM, 2011, p.270).

Pode ser observado na Tabela 2, que o custo da mesma estrutura BC-PDM que foi relativamente melhor na primeira etapa de comparações, também é menor que o valor gasto na infraestrutura que é empregada atualmente. A estrutura BC-PDM custa aproximadamente um quinto do valor do sistema atual, onde pode ser observado que a economia maior fica por conta da utilização dos servidores comuns, devido seu baixo valor (LAM, 2011, p.270).

		BC-PDM (16 nós)	Servidor comecial atual (Unix Server)
Custo de hardware	Capacidade de computação	CPU: 64 cores memória: 128 gigabyte	CPU: 8 cores memória: 32 gigabyte
	Capacidade armazenamento	16 terabyte	Storage array
	Custo	240,000 CNY	4,000,000 CNY
Custo de software	Banco de dados	500,000 CNY	1,000,000 CNY
	Aplicativos	300,000 CNY	500,000 CNY
	Custo de manutenção	200,000 CNY	500,000 CNY
Total		1,240,000 CNY	6,000,000 CNY

Tabela 2 – Comparação de valores do sistema em uso com o BC-PDM (In: LAM, 2011, p.271)

As comparações evidenciaram que a infraestrutura adquirida, *cluster Hadoop* BC-PDM de 256 nós, consegue trabalhar com dados no patamar dos 100 *terabytes* (LAM, 2011, p.272).

Quando o experimento foi empregado no ambiente real, um de seus objetivos iniciais foi a classificação da base de clientes em diferentes perfis digitais, com o intuito de modelar estratégias de *marketing* personalizadas para cada tipo de usuário. Foi utilizado o algoritmo paralelo de agrupamento, que executou a tarefa três vezes mais rapidamente que a solução utilizada anteriormente (LAM, 2011, p.272).

Diante de todas as informações explanadas, pode-se concluir que o mercado de telecomunicações possui níveis gigantescos de criação de dados. Tal cenário vem crescendo cada vez mais com os baixos preços dos aparelhos de telefonia móvel. Grande parte das ferramentas comerciais que estão disponíveis no mercado para manipulação de grandes massas de dados não trabalham de forma paralela, onerando o orçamento do projeto com a compra de *hardware*. O projeto ambientado

em *Hadoop* possui precisão, velocidade, um custo baixo e uma alta capacidade de escalabilidade (LAM, 2011, p.272).

6.4 UTILIZAÇÃO DO *HADOOP* PELA *LAST.FM*

A *web* rádio *Last.fm* iniciou suas atividades em 2002, com uma gama muito grande de serviços disponíveis aos seus usuários, como por exemplo, *downloads* de músicas, recomendações de *shows* e eventos musicais, entre outros. Possui atualmente 25 milhões de usuários por mês, que criam uma enorme quantidade de dados em suas interações com os serviços prestados pela rádio. A estrutura disponibilizada pela *Last.fm* gera perfis de usuários, estabelecendo gostos musicais, artistas preferidos, *shows* de interesse, serviços estes produzidos a partir do histórico de todas as interações realizadas pelo usuário no *site* (WHITE, 2010, p.497).

De acordo com White (2010, p.497), o grande aumento na quantidade de usuários em um curto espaço de tempo revelou problemas no armazenamento, processamento e gerenciamento da entrada e saída dos dados. Para solucionar os problemas, a *Last.fm* resolveu utilizar a mesma ferramenta que outras organizações aderiram para resolver os mesmos empecilhos que a *Last.fm* estava enfrentando, o *Hadoop*. Os motivos que fizeram optar pelo *Hadoop* foram*:*

- Código-fonte aberto – possibilitando modificações e criação de funcionalidades específicas.
- Escalabilidade simplificada – obtida através das combinações de *hardwares,* que trabalham de forma paralela, de baixo custo ao invés de supercomputadores.
- Sistema distribuído de arquivos – facilitando *backups e acesso* aos dados de forma mais rápida.

- *Software* livre – economizando recursos financeiros em um período onde o poder econômico da empresa era frágil.

Foi utilizada uma infraestrutura *Hadoop* composta por dois *clusters Hadoop* com mais de 500 nós, 300 núcleos e 100 *terabytes* de capacidade de armazenamento em HD. A primeira utilização do *Hadoop* pela *Last.fm* foi na confecção de gráficos que representam *rankings* de músicas mais acessadas, artistas mais populares, entre outras (WHITE, 2010, p.498).

6.4.1 *Hadoop na* produção dos gráficos da *Last.fm*

A partir do acesso realizado pelos usuários às faixas musicais contidas na página *web* da rádio, a *Last.fm* consegue produzir, com o emprego de tarefas do *Hadoop*, diversos gráficos, com informações distintas, como rankings separados por países de cantores preferidos e/ou músicas mais ouvidas (WHITE, 2010, p.498).

Segundo White (2010, p.498), esses gráficos são disponibilizados aos usuários com uma periocidade semanal e até mesmo mensal. A Figura 17 exibe um exemplo da disposição dos dados no *site* da *Last.fm* referente ao *ranking* dos artistas e músicas mais acessados em um período de avaliação semanal.

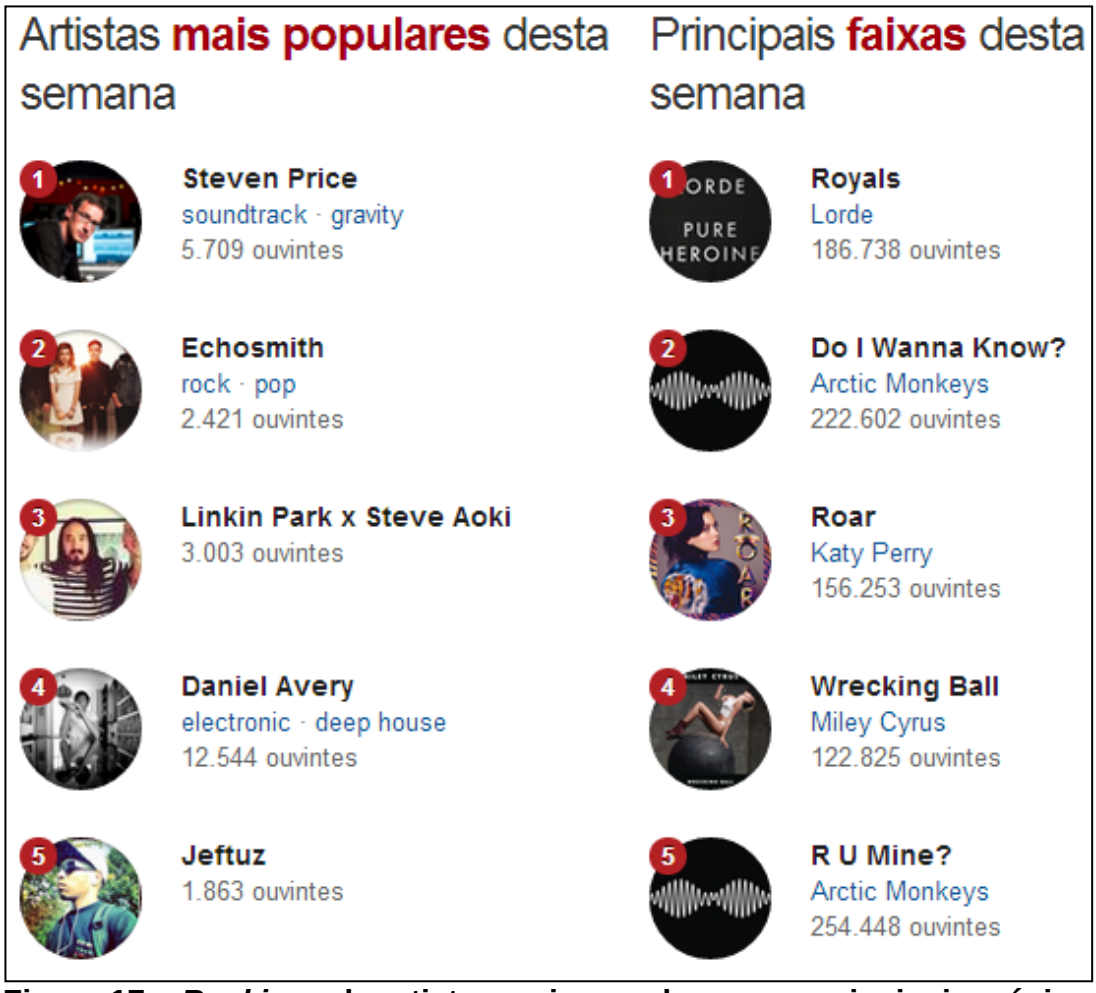

Figura 17 – *Rankings* de artistas mais populares e as principais músicas acessadas na semana (In: WHITE, 2010, p.498)

Segundo White (2010, p.499), a fomentação das informações que servem de subsídios para a execução das tarefas *Hadoop* são obtidas por dois métodos:

- Pelos aplicativos que a *Last.fm* disponibiliza para seus clientes instalarem em seus celulares, *smartphones*, *iPods* e *notebooks*. Estes aplicativos captam as informações que o usuário gera (por exemplo, o acesso de um arquivo de música no computador, celular ou outro aparelho digital com conexão à *internet*) e encaminham para a *Last.fm*. A rádio criou um termo para essa fonte de dados que foi denominada *Scrobble*,

- Pela utilização da própria rádio *web* pelo usuário. Pois, ao selecionar uma música ou realizar o *download* de um arquivo no *site* da *Last.fm* o usuário cria informações que serão utilizadas pelo *Hadoop* para determinar o perfil musical de cada cliente individualmente.

A *Last.fm* preocupou-se em separar as duas fontes de dados para que não haja uma duplicidade nas informações, permitindo com que as tarefas do *Hadoop* formulem gráficos e *rankings* fidedignos (WHITE, 2010, p.499).

Uma das funções do *Hadoop* é receber as informações, analisá-las em forma de faixas e, em seguida, criar um resumo para disponibilizar no *site* ou em outras funções *Hadoop* (WHITE, 2010, p.499).

6.4.2 Programa de análise das faixas musicais (*Track Statistics Program*)

Quando a tarefa *Hadoop* de análise inicial dos dados produzidos pelos usuários é executada, os dados são validados e convertidos em um arquivo de texto que armazena a identificação do usuário, o número de identificação da música, a quantidade de vezes que o arquivo foi acessado pelo usuário da rádio *web* e pelo *Scrobble*, além da quantidade de vezes que o arquivo foi ignorado. A tabela 3 ilustra um exemplo simplificado, tendo em vista que o cenário real possui tamanhos calculados em *gigabytes*, das informações que são geradas a partir da análise e conversão dos dados produzidos pelos usuários (WHITE, 2010, p.499).

IdUsuário	IdMúsica	Scrobble	RadioWeb	Ignorado
1111115	2222	1	0	0
1111144	2222	1	0	0
1141112	1455	0	0	1
1111125	1222	0	1	0
1111113	1149	0	1	1
1111117	2255	1	0	0
1111115	31111	0	1	0

Tabela 3 – Lista dos dados convertidos (In: WHITE, 2010, p.499)

Essa lista serve de base para a execução do programa *Track Statistics*. Seu processo de execução manipula os dados em duas formas diferentes e depois realiza uma junção dos resultados, produzindo um resultado final de todo o processo de execução do programa. A Figura 18 demonstra os processos do *Track Statistics* (WHITE, 2010, p.500).

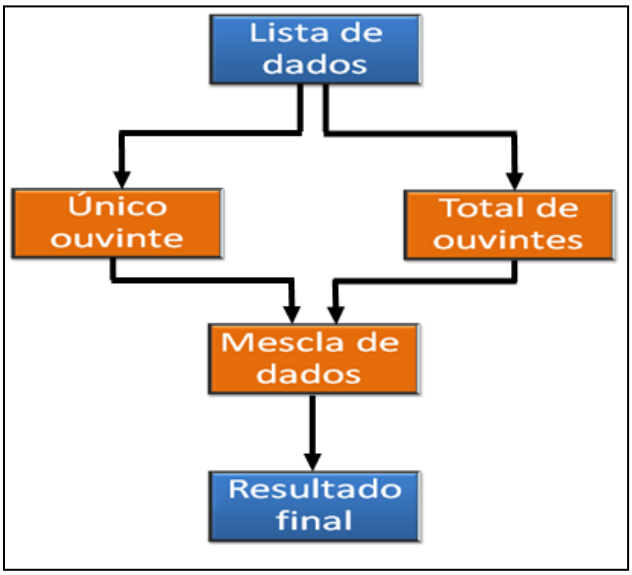

Figura 18 – Etapas da execução do *Track Statistics Program* (In: WHITE, 2010, p.500)

Na etapa de intitulada como Único Ouvinte, são considerados apenas as primeiras músicas acessadas por cada usuário, sendo descartados os acessos seguintes realizados pelo cliente. Esse processo gera a quantidade total de usuários que fez acesso a um mesmo arquivo (WHITE, 2010, p.500).

Na tarefa Total de Ouvintes, é computado a quantidade de vezes em que cada arquivo é acessado tanto pelos usuários da rádio *web Last.fm* quanto pelos S*crobbles*. Mesmo que a origem dos dados seja a mesma, cada uma das operações produz resultados distintos: a primeira tarefa descrita acima gera informações relacionadas aos acessos dos arquivos primariamente pelo usuário e a segunda tarefa resulta em informações quanto ao número de vezes que cada faixa de música foi acessada (WHITE, 2010, p.500).

A tarefa de Mescla de Dados reúne os resultados obtidos através do emprego dos processos de Único Ouvinte e Total de Ouvintes, gerando os resultados finais do processo, que são (WHITE, 2010, p.500):

- Quantidade de usuários que ouviram uma mesma música como primeira escolha.
- Quantidade de acessos de um arquivo realizado pelos usuários da rádio *web*.
- Quantidade de acessos de uma música realizado pelos *Scrobbles*.
- Quantidade total de acessos que o arquivo foi visualizado.
- Quantidade total de vezes que o arquivo foi ignorado.

Os resultados finais são armazenados em um servidor e através de um serviço *web* são disponibilizados no *site* da *Last.fm* para os usuários realizarem consultas. A Imagem 19, ilustra a visualização de um arquivo de mídia na página *web* da rádio (WHITE, 2010, p.505).

Figura 19 – Visualização de uma faixa de áudio no *site* da rádio *Last.fm* (In: WHITE, 2010, p.500)

A *Last.fm* encontrou no ambiente *Hadoop* uma solução para a manipulação dos mais variados tipos de dados que circulam em suas aplicações, extraindo funcionalidades e criando novas formas de satisfazer os desejos de seus clientes, criando perfis individuais para cada usuário (WHITE, 2010, p.506).

Com base nos resultados obtidos através do emprego da ferramenta *Hadoop* na *Last.fm,* White (2010, p.506) conclui que, por ser uma comunidade ativa no desenvolvimento de novas funcionalidades e melhorias nas aplicações já existentes, o *Hadoop* se torna uma ferramenta essencial para a nova era de *Big Data*, pois sua execução baseada no paralelismo permite uma resposta de execução (análises de dados, inserções, exclusões, atualizações) e produção de resultados, muito superior e em menor tempo quando comparado com outros sistemas disponíveis no mercado (WHITE, 2010, p.506).

7 CONSIDERAÇÕES FINAIS

O tema, no campo tecnológico, ainda carece de estudos e conteúdo documental, uma vez que as poucas informações sobre o assunto estão expostas de forma dispersa e muito sucinta, somada a dificuldade que é reproduzir um ambiente *Big Data*. O presente estudo exploratório sobre a tecnologia de *Big Data* apresentou de forma generalizada as várias áreas que englobam o ambiente de grandes dados. Também foram relatadas as necessidades que surgiram com este advento tecnológico, que vão desde as ponderações que decidem ou não pelo emprego da tecnologia, até a geração do produto final, para as quais os processos de *Big Data* são interpostos. Porém as inovações e os desafios da era *Big Data* estão muito além do que foi apresentado neste trabalho. Cada quesito documentado nesta monografia possui totais condições de formar um tema de pesquisa, sendo assim, o objetivo proposto se limitou a situar o leitor deste projeto em um contexto geral, pois se cada tema fosse descrito com todos os seus pormenores, os prazos estabelecidos não seriam suficientes para tal explanação.

Trabalhar com uma tecnologia inédita e que, ao mesmo tempo se tornou um assunto viral no campo da informática sobre o qual os estudos de caso estão disponibilizados de forma singela e empobrecida de informações relevantes, torna-se uma árdua tarefa que necessita do empenho de vários grupos de pessoas, pois o assunto mostra a cada dia que será o propulsor de uma nova revolução mundial na área de Tecnologia da Informação.

As grandes inovações de *marketing*, as descobertas de novos padrões de informações, o aumento dos lucros das empresas, a necessidade de se criar leis digitais, a predição de catástrofes, entre outras, serão criadas, analisadas e manipuladas em ambientes *Big Data* em um futuro muito próximo.

Os escassos resultados do emprego de *Big Data* que estão disponíveis em forma de material literário mostram que as técnicas de manipulação *Big Data* são as soluções com o menor custo, o menor desprendimento de tempo e conseguem satisfazer todos os objetivos propostos a elas.

7.1 TRABALHOS FUTUROS

Uma primeira sugestão de trabalhos futuros refere-se à possibilidade de que sejam desenvolvidas ferramentas computacionais, com base em algoritmos especificados, reproduzindo um ambiente de *Big Data* para subsidiar novas pesquisas e auxiliar a descoberta de conhecimento e novos padrões.

Outra sugestão de trabalhos futuros pauta-se na possibilidade de estudar e analisar as ferramentas que manipulam *Big Data*, no sentido de apontar as vantagens e desvantagens, auxiliando as empresas na escolha de uma ferramenta que atenda as necessidades da organização, levando em consideração seu porte e área de atuação no mercado.

REFERÊNCIAS

BALZA, Guilherme. **Brasil é o grande alvo dos EUA", diz jornalista que obteve documentos de Snowden**. Universo Online, UOL. Disponível em: <http://noticias.uol.com.br/internacional/ultimas-noticias/2013/09/04/brasil-e-o-grande-alvo-dos-eua-diz-jornalista-que-obteve-documentos-de-snowden.htm>. Acesso em: 17 set. 2013.

CORDEIRO, Daniel; GOLDMAN, Alfredo. ***Apache Hadoop*: Conceitos teóricos e práticos, evolução e novas possibilidades**. 2012. 88p. Mini-curso avançado realizado no III Escola Regional de Alto Desempenho de São Paulo – ERAD SP, 2012, Campinas, São Paulo, julho, 2012.

COSTA, Carlos Eduardo R. da. **Armazenamento de dados em Sistemas Gerenciadores de Banco de Dados Relacionais (SGDBR's)**. 2011. 47p. Trabalho de Conclusão de Curso (Tecnologia em Processamento de Dados) – Faculdade de Tecnologia de São Paulo – FATEC/SP.

COSTA, Luís Henrique M. K.; AMORIM, Marcelo D. de; CAMPISTA, Miguel Elias M.; RUBINSTEIN, Marcelo G.; FLORISSI, Patrícia; DUARTE, Otto Carlos M. B. **Grandes massas de dados na nuvem:** desafios e técnicas para inovação. In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES – SBRC, 2012, Ouro Preto, Minas Gerais, maio, 2012.

EMC. **A realidade sobre gerenciamento de segurança e *big data***. *White Paper*, 2012. Disponível em: <http://brazil.emc.com/collateral/white-papers/h0812-getting-real-secuirty-management-big-data-wp.pdf>. Acesso em 18 set. 2013.

FAYYAD, Usama; PIATETSKI-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Communications of the ACM**, v. 39, nov., 1996. p. 27-34.

GANTZ, John; REINSEL, David. **Extracting value from chaos**. Disponível em: <http://brazil.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>. Acesso em: 25 jun. 2013.

GARCIA, Edi Wilson. **Pesquisar e avaliar técnicas de mineração de dados com o uso da ferramenta Oracle Data Mining**. 2008. 66p. Trabalho de Conclusão de Curso (Tecnologia em Processamento de Dados) - Fundação Educacional do Município de Assis – FEMA/Instituto Municipal de Ensino Superior de Assis - IMESA.

GOLDMAN, Alfredo; KON, Fabio; PEREIRA, Francisco Jr.; POLATO, Ivanilton; PEREIRA, Rosangela de Fátima. Apache Hadoop**:** conceitos teóricos e práticos, evolução e novas possibilidades. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 2012, Curitiba. **Anais do XXXII Congresso da Sociedade Brasileira de Computação**, julho, 2012.

GREGO, Maurício. **Como a NSA consegue espionar milhões de pessoas nos EUA. Revista.** Exame *on-line*. Disponível em: <http://exame.abril.com.br/tecnologia/noticias/como-a-nsa-consegue-espionar-milhoes-de-pessoas-nos-eua>. Acesso em 22 set. 2013.

JUNIOR, Walter L. T. **Jornalismo Computacional em função da Era do Big Data**. Rio de Janeiro: ECO – Universidade Federal do Rio de Janeiro, 2011.

LAM, Chuck. **Handoop in Action**. Stamford: Manning, 2011.

NAVEGA, Sérgio. Princípios Essenciais do *Data Mining*. In: INFOIMAGEM'2002 – DOCUMENT MANAGEMENT CONFERENCE & EXPOSITION, 2002, São Paulo, Brasil. **Anais da INFOIMAGEM'2002**, novembro, 2002.

NIMER, F; SPANDRI, L. C. *Data Mining*. **Revista Developers**. v.7, fevereiro, 1998.

PETRY, André; VILICIC, Filipe. A era do Big Data e dos algoritmos está mudando o mundo. **Revista Veja**. n.20, maio, 2013. p.70-81.

RISEN, James; LICHTBLAU, Eric. How the U.S. Uses Technology To Mine More Data More Quickly. **The New York Times**, Washington, 09 junho, 2013. p.A1.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. **Sistema de Banco de Dados**. 3. ed. Tradução de Marília Guimarães Pinheiro e Cláudio César Canhette. São Paulo: Makron Books, 2005.

SILVA, Anderson Evandro Simeão. **Uma Análise Comparativa das Tecnologias de Banco de Dados Relacional e de Banco de Dados Nativamente Orientado a Objetos**, 2003. 202p. Tese (Mestrado) – Instituto Tecnológico de Aeronáutica – Campo Montenegro, São Paulo, São José dos Campos, 2003.

SILVA, Fábio Alves; GOES, Leandro. **Estudo e Avaliação de Técnicas de Mineração de Dados no SGBD Comercial Oracle**. Centro Universitário Padre Anchieta Jundiaí: São Paulo, 2007.

SILVA, Thiago Miranda Amorim. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados**.

2006. 50p. Trabalho de Conclusão de Curso (Ciência da Computação) – Centro de Informática, Universidade Federal de Pernambuco. Pernambuco, Recife, 2006.

TAKAI, Osvaldo Kotaro; ITALIANO, Isabel Cristina; FERREIRA, João Eduardo. **Introdução a Banco de Dados**. São Paulo: Departamento de Ciência da Computação; Instituto de Matemática e Estatística; Universidade de São Paulo, 2005.

TOREY, Toby; LIGHTSTONE, Sam; NADEAU, Tom; JAGADISH, H. V. **Database Modeling and Design: Logical Design**. 5 ed. USA: Elsevier, 2011.

TWO CROWS. **Introduction to Data Mining and Knowledge Discovery**. 3 ed. Disponível em: <http://www.twocrows.com>. Acesso em 19 set. 2013.

WHITE, Tom. **Hadoop: The Definitive Guide**. 2 ed. Cambridge: O'Reilly, 2010.

WEF (World Economic Forum). **Lista mantida pelo *Committed to Improving The State of The World***. Genebra: World Economic Forum, 2012.