UNIVERSIDADE VEIGA DE ALMEIDA – UVA

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

BIG DATA: ANÁLISE E MINERAÇÃO DE DADOS

LETÍCIA DE OLIVEIRA MARQUES

RIO DE JANEIRO

2016

UNIVERSIDADE VEIGA DE ALMEIDA - UVA

LETÍCIA DE OLIVEIRA MARQUES

Monografia apresentada ao curso de Ciência da Computação da Universidade Veiga de Almeida, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Alfredo Nazareno Pereira Boente

BIG DATA: ANÁLISE E MINERAÇÃO DE DADOS

RIO DE JANEIRO

2016

UNIVERSIDADE VEIGA DE ALMEIDA - UVA

BACHARELADO EM ENGENHARIA DA COMPUTAÇÃO

LETÍCIA DE OLIVEIRA MARQUES

BIG DATA: ANÁLISE E MINERAÇÃO DE DADOS

Monografia apresentada como requisito parcial à conclusão do curso em Bacharel em Ciência da Computação.

APROVADA EM:

CONCEITO: _____

BANCA EXAMINADORA:

_____

PROF. ALFREDO NAZARENO PEREIRA BOENTE, PHD.

ORIENTADOR

_____

PROF. LUIS OTAVIO DE MARINS RIBEIRO, DSC.

_____

PROFª . RENATA MIRANDA PIRES BOENTE, MSC.

Coordenação de Ciência da Computação

Rio de Janeiro, 18 de Junho de 2016.

Dedico este trabalho a minha mãe, que sempre acreditou no meu potencial me impulsionando sempre a ir adiante, fazendo desta caminhada, um grande aprendizado e tornando-a possível.

“Um projeto *Big Data* requer uma transformação sincronizada entre pessoas, processos e tecnologias. Todas as três devem marchar em sincronia, caso contrário o projeto falhará.”

Fonte: Minelli, Chamber, Dhira; Wiley

RESUMO

Com a importância dos dados nos diversos fatores e áreas da sociedade, o armazenamento destes foi se tornando cada vez mais importante, onde ao passar das épocas foi-se tornando visível o avanço quanto a necessidade de dados e consequentemente a esta visualização, se foi percebendo um acumulo cada vez maior de informações. Podemos dizer que o mundo muda rapidamente e a quantidade de informação armazenada cresce *n* vezes mais rápido a cada ano e todos são afetados pelas mudanças que este crescimento desencadeia, o termo *Big Data* est diretamente associado a estas mudanças e ao futuro do manuseio destes dados, e este está fadado a abalar tudo, dos negócios a ciência e saúde, governo, educação, economia, ciências humanas e todos os demais aspectos da sociedade.

No decorrer deste trabalho é mostrado de forma simplificada o que está por traz dos princípios de análise e mineração que se encontram em projetos de *Big Data*, identificando as fases e métodos utilizados no decorrer destes processos, com o intuito de abrir um pouco o universo tão grande que o *Big Data* está se tornando.

Palavras-Chave: Big Data, Mineração, Análise, Dados

ABSTRACT

With the importance of data on various factors and areas of society, the storage of these was becoming increasingly important, where eith the passing of times was becoming noticeable the progress as the need of data and therefore this visualization of it, then it was realized that an information accumulation had exponentily growing. We can say that the world changes quickly and the amount of information stored grows n times faster each year and everyone is affected by the changes that this growth triggers, the term of Big Data is directly associated with these changes and the future of the handling of this data, and it is destined to shake everything from business to science and health, government, education, economics, social sciences and all the other aspects of society.

Throughout this work it is shown, in a simple way, what is behind the principles of analysis and mining that are in Big Data projects, identifying the phases and methods used in the course of these processes, in order to open a little the large universe that Big Data is becoming.

Keywords: Big Data, Analysis, Mining, Data

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ABNT – Associação Brasileira de Normas Técnicas

TCC – Trabalho de Conclusão de Curso

IBM – International Business Machines

CRISP-DM – Cross-Industry Standard Processo of Data Mining

DBScan – Density-based Spatial Clustering of Application with Noise

DHT – Distributed Hash Table

UFC – Universidade Federal do Ceará

RAM – Random Access Memory

GB – Gigabyte

CPU – Central Processing Unit

SUMÁRIO

1 INTRODUÇÃO

Com o decorrer do tempo, ficou cada vez mais clara a importância e a dependência que as informações possuem para o bom desenvolvimento e crescimento das empresas, não importando qual seja sua estrutura ou seguimento.

Com a importância dos dados nos diversos fatores e áreas da sociedade, o armazenamento destes foi se tornando cada vez mais comum, o que resultou em um Bases de dados em formatos diversificados e não controlados e com o decorrer deste armazenamento, foi-se notando um acumulo cada vez maior dos mesmos.

Ao contrário do que era visto a alguns anos atrás, *terabytes* já são encontrados em nossas casas e estima-se que uma empresa com mil funcionários gera anualmente 1000 *terabytes*, sem falar que essa quantidade tende a aumentar cinquenta vezes até 2020, então notamos que já falamos comumente em *petabytes* e *zetabytes* começam a ser uma escala real e não mais imaginária e futurista. Este crescimento de dados não para e não tende a diminuir, possuindo cerca de 60% de crescimento anual das informações das empresas.

A forma de armazenamento destas informações vem sendo modificadas de acordo com as tecnologias lançadas, por exemplo, quando não havia TI, os diretores das empresas eram obrigados a criar estratégias baseadas em pouca informação, não porque eles queriam assim, mas porque simplesmente era inviável coletar muitas informações, neste processo diversos dados relevantes eram perdidos. Com a chegada da tecnologia da informação e a adaptação do ser humano a ela a perda destas informações vem sendo diminuída cada vez mais, levando estas organizações a necessidade da utilização destas informações de maneira objetiva, para assim se tornarem mais competitivas. Surgiu então uma corrida para armazenar o máximo possível de informações sobre o negócio e analisa-las da maneira mais rápida possível, fazendo com que dados de diversas áreas e de diferentes formas sejam armazenadas de forma não estruturada.

Após todo este processo que percebemos que a questão não é que você está coletando grandes quantidades de dados, mas sim o que você faz com eles, é então que é apresentado o conceito de *Big Data*, que tem como grande desafio a administração de um grande volume de dados, minerando as informações não somente com rapidez, mas também de forma concisa da maneira necessária para as organizações, possibilitando então a disponibilizadas aos usuários.

Teoricamente não existe uma unanimidade quanto a definição simples do que é *Big Data*, apesar de que para diversos autores existe um consenso quanto a sua força modificadora no contexto da utilização dos dados.

Neste trabalho temos como objetivo mostrar o que está por traz dos projetos de *Big*

Data, identificando as dificuldades encontradas no caminho das análises e os processos utilizados para auxiliar aos analistas e programados, para assim, direciona-los a encontrar quais as medidas e caminhos se deve tomar durante o decorrer dos projetos.

Será mostrado no decorrer do trabalho técnicas utilizadas para auxilias na separação e extração dos dados relevantes para os usuários, mostrando formas de como estes dados devem ser vistos e tratados pelo ponto de vista dos usuários e dos responsáveis pelo projeto.

Para finalizar, serão passadas algumas das ferramentas utilizadas para armazenar e utilizar estas massas de dados adquiridas, citando como é realizada a utilização na prática destes dados, mostrando ao final modelos reais da utilização de *Big Data* e seus recursos no mundo.

2 REFERENCIAL TEÓRICO

Neste capítulo serão abordados de forma simples as técnicas e modelos de análise, mineração e extração de uma grande massa de dados para seu uso, mostrando inicialmente como o termo *Big Data* se originou, além de sua utilidade e funcionalidade no montante de dados existentes. Apresentamos também as técnicas para a utilização destes dados, assim como Kenneth Cukier, Chun-Yang Zhang, Min Chen, Jo Ho-Kim, Thomas Davenport, entre outros apresentaram em seus estudos publicados, onde terão seus nomes mencionados ao decorrer deste trabalho.

2.1 Big Data

Nos últimos 20 anos, os dados têm aumentado de forma surpreendente, de formas diferentes e em vários campos da sociedade. De acordo com vários campos de pesquisa a massa de dados geradas, ao redor do mundo, teve um aumento de aproximadamente nove vezes no decorrer dos últimos dez anos, possuindo a tendência de crescer cada vez mais, neste quesito, podemos citar ainda um estudo realizado pela IBM, diz que até o ano de 2008 já tinham sido produzidos mais de 2,5 quintilhões de bytes, sendo que aproximadamente 90% das informações armazenadas em centrais, foram produzidas nos últimos dois anos, fato este devido a inserção das empresas no meio online (Internet), além da difusão de dispositivos móveis por exemplo.

A partir do momento em que a sociedade começou a acumular esta grande massa de dados e com a falta de mecanismos para suportar os mesmos, a sociedade começou a perder informações relevantes, foi então que surgiu diversas formas de tratamento para o gerenciamento destes monte de dados, é neste universo que surgiu o termo *Big Data*, que refere-se às tecnologias e iniciativas que envolvem conjuntos gigantescos de dados que são diversificados, que possuem grande complexidade quanto ao armazenamento, ao modo de transportamento, assim como analisar os mesmos a partir do uso de ferramentas ou métodos de processamento de dados tradicionais, além disto, *big data* também traz novas oportunidades para a descoberta de novos valores, nos ajudando a ter uma melhor compreensão dos dados que não vemos, assim como novos desafios.

Na verdade, o advento do *Big Data* é o espelho da evolução tecnológica social. Trata-se da nossa grande capacidade de captar montantes de informações, analisá-las de imediato e tirar conclusões, por vezes, profundamente surpreendentes. Um fenômeno emergente, de amplitude crescente e ações tão distintas que atendem desde uma busca por melhores tarifas de

bilhetes de avião até a dataficação de textos contidos em milhões de livros. Nosso crescente poder na computação entra em ação para descobrir epifanias nunca antes imaginadas. Estamos diante de uma revolução emparelhada à Internet. [Temática, 2010]

Segundo a IBM (International Business Machines) uma das funções e atuações fundamentais do *Big Data* é justamente de ser capaz de trabalhar com muitas variáveis simultaneamente, além de leitura e reinderização de imagens, em tempo mínimo e muita eficácia.

Ao falarmos, lermos ou ouvirmos falar sobre o termo *Big Data* seus 3 V's estão sempre sendo atribuídos a ele. Estes V's surgiram em 2001, através do analista do setor de Doug Laney (Gartner) que articulou a definição do termo *Big Data* em três Vs na origem: Volume, Velocidade e Variedade, porém diversos outros autores como Francisco Oliveira, citam que estas características não deveriam ficar somente ligadas a estes citados, mantendo a ideia de que não seriam apenas estas três características, acrescentando a estas Variabilidade, Complexidade, Veracidade, Valor. Porém qual seria o significado de cada característica, estes elementos são descritos abaixo:

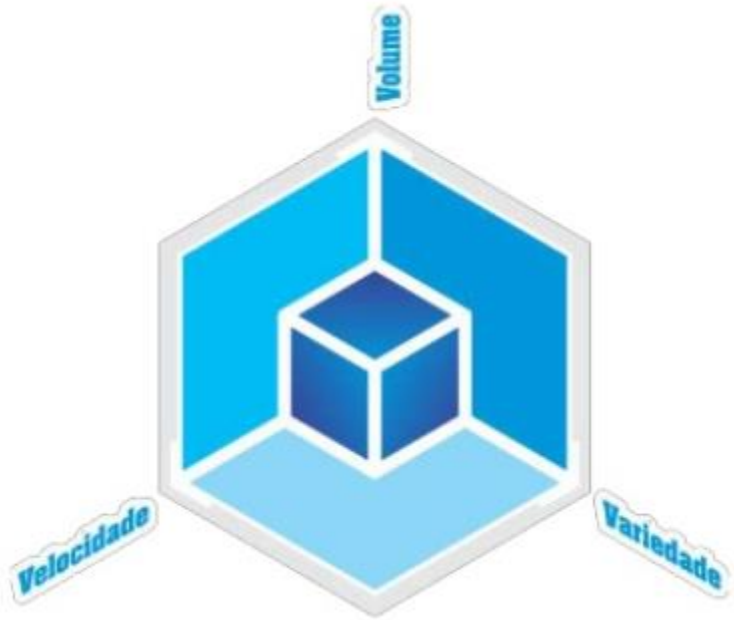

Figura 1: Os três V's iniciais
Fonte: Francisco Oliveira (2013)

- Volume – Este está relacionado ao volume dos dados armazenados/produzidos. Neste tema, atualmente surgem questões como, determinar a relevância entre os grandes volumes de dados e como criar valor a partir dessa relevância.
- Velocidade – Este está relacionado velocidade, que significa o quão rápido os

dados são produzidos como o quão rápido são tratados. Reagir rápido o suficiente para lidar com a velocidade é um desafio para a maioria das organizações.

- Variedade – Este está relacionado ao formato dos dados trabalhados, a diversidade de como os dados são armazenados é enorme o que dificulta na hora da tratativa dos mesmos.

Estas características estarão sempre ligadas as necessidades dos usuários e das aplicações, onde os dados serão mostrados de acordo com a forma com que serão coletados e extraídos.

O processo de desenvolvimento para uso desta massa de dados é longo e trabalhoso, envolvendo diversas fases, que unidas e trabalhadas tornem o processamento, a visualização, a distribuição e a análise dos dados possível; porém de acordo com Jonathan Stuart Ward e Adam Barker apesar de existirem tantas fases, é possível dizer que *Big Data* é predominantemente associada a duas ideias, sendo elas o armazenamento e a análise dos dados. Iremos analisar e visualizar alguns destes pontos no decorrer deste trabalho.

2.2 *Big Data*: Métodos e Técnicas de Análise

Com uma grande massa de informação a ser trabalhada, a extração dos dados relevantes se tornou algo de valor fundamental, porém os modelos e técnicas de análise e extração existentes não são aplicados a estes, então como capturar o valor dos mesmos? Foi desta forma que se viu a necessidade de desenvolver novas técnicas e tecnologias para fazer esta análise. Até o momento foram desenvolvidas diversas técnicas para o tratamento, análise e visualização de *Big Data*.

De acordo com alguns autores, a análise dos dados é a fase mais importante no tratamento da retirada de valor quando trabalhamos com *Big Data*, este fato se deriva a partir do ponto em que necessitamos que os dados sejam concretos e possuam valor.

Técnicas de *Big Data* envolvem diversas disciplinas, assim como métodos analíticos para dados tradicionais e *Big Data*, arquitetura analítica para *Big Data* e produtos de *software* usados para mineração e análise.

Existem diversos métodos e técnicas, algumas sendo mais evidenciadas que outras, pode-se citar como exemplo as seguintes técnicas:

- Análise de Cluster – é um método estatístico utilizado para agrupar objetos, sendo mais especificamente, classificando os objetos de acordo com algumas características.
- Análise Fatorial – visa descrever a relação entre muitos elementos utilizando somente alguns fatores, agrupando diversas variáveis relacionadas em um fatorial e então este fatorial, utilizando-o para revelar as informações mais valiosas dos dados originais.
- Análise de Correlação – é um método analítico utilizado para determinar as leis de relacionamento, correlação, dependências e restrição mútua. Tais relacionamentos podem ser classificados de dois modos:
- Função: reflete a estrita relação de dependência entre os fenômenos, que também é chamada de definitiva relação de dependência;
- Correlação: reflete as algumas relações de dependência indeterminadas ou inexatas.
- Análise de Regressão – é uma ferramenta matemática que revela as correlações entre uma variáveis e diversas outras. Baseada em um grupo de experimentos ou observações de dados, análise de regressão identifica os relacionamentos das dependências entre as variáveis ocultas por aleatoriedade.
- Teste A/B - também chamado de teste de bucket. É uma tecnologia para determinar como melhorar as variáveis alvo através de grupos de comparação. *Big Data* necessitará da execução de diversos testes e análises.
- Análise Estatística – é baseada em teoria estatística, um a vasta área para matemática aplicada. Na teoria estatística, aleatoriedade e incerteza são modeladas com a Teoria das Probabilidades. Esse modelo pode fornecer descrição e inferência para *Big Data*, podendo a análise estatística descritiva resumir e descrever conjuntos de dados, enquanto a análise estatística inferencial pode desenhar a conclusão de dados sujeitos a variações aleatórias.
- *Map/Reduce* - é um modelo de programação desenhado para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes. Programas *MapReduce* são escritos em um determinado estilo influenciado por construções de programação funcionais, especificamente expressões idiomáticas para listas de processamento de dados.

- A compreensão do funcionamento da técnica *Map/Reduce* ajuda no desenvolvimento de aplicações críticas que trabalham com *Big Data*, que são bancos de dados difíceis de serem manipulados por tecnologias convencionais. Entretanto, a adoção de uma implementação *Map/Reduce*, requer habilidades específicas sobre o modelo de programação *Map/Reduce*, que introduz novas exigências na forma de modelar a abstração sobre os dados e processar *Big Data*.

Acima são citadas algumas das diversas técnicas, métodos e ferramentas existentes para que a fase de análise do montante de dados seja executada. Mesmo a análise sendo uma fase de grande importância ela não está sozinha, para que esse processo possa ser bem desenvolvido e trabalhado a parte de data mining, ou traduzido livremente mineração de dados, acompanha lado a lado todo o processo de retirada dos dados com valor e este é o assunto apresentado a seguir.

2.3 *Big Data*: Data Mining (Mineração de Dados)

Data mining é uma expressão originada na Inglaterra, que está ligada à informática, podemos traduzir de forma simples como mineração de dados. A mineração de dados consiste em uma funcionalidade que agrega valor e organiza dados, encontrando nos mesmos padrões, associações, mudanças e anomalias relevantes. [DevMedia - 2009].

Podemos definir de forma simplificada como processo de extração de dados escondidos e desconhecidos que tenham um valor em potencial e conhecimento de dados massivos, incompletos, barulhentos, nebulosos (fuzzy) e aleatórios.

Este assunto será melhor desenvolvido a frente, onde seu funcionamento será melhor detalhado.

3 ARMAZENAMENTO DE DADOS NO DECORRER DOS ANOS

Desde a antiguidade o homem sente a necessidade de evoluir, onde a mesma vem se modificado com o decorrer do tempo; estando diretamente ligada as preocupações da época. Tudo se iniciou com o homem pré-histórico, que possuía basicamente duas preocupações: manter-se aquecido durante o frio, vento e chuva; caçar seu alimento para se manter vivo. Após sua evolução, com o passar dos milênios, o homem saiu de sua maneira pré-histórica e foi-se verificada uma nova necessidade, maneiras para facilitar a coleta de alimentos e formas de armazena-los. Registros históricos comprovam que as primeiras civilizações no antigo Egito já possuíam técnicas de armazenamento de alguns tipos de alimento. Passando os anos e chegando ao século XIV, foi registrada a invenção da primeira geladeira industrial e, posteriormente, um grande número delas ocupavam também os espaços residenciais. [MG (DINO) - 2015]

Nos dias atuais não nos preocupamos com a forma de coletarmos nossos alimentos ou mesmo armazena-los, pois, a sociedade se desenvolveu para que nossas necessidades básicas fossem supridas, sendo assim, da mesma forma que o tempo passou e mudou as necessidades atuais mudaram. [MG (DINO) - 2015]

As necessidades se adaptaram ao tempo, e na era da informática em que vivemos os problemas são outros. Não nos preocupamos mais em como guardar nossos alimentos ou até mesmo as imagens da câmera digital ou onde guardar aqueles arquivos de vídeo e música baixados na internet. Por sorte, os meios de armazenamento também evoluíram com o decorrer das épocas. Já verificamos a criação de diversos métodos de armazenamento como: Cartão Perfurado, Cassete, Disquete, DAT, CD-R, Pen drive e até chegarmos na Cloud, mas se os meios de armazenamento estão evoluindo, qual a necessidade que devemos suprir no momento? O desafio em nosso tempo, está relacionado com a forma como utilizamos as informações que temos armazenadas, isto é, como aproveitamos todos os dados que temos espalhados em diversos lugares.

Se pensarmos na quantidade de informação que armazenamos e na quantidade de lugares onde estas estão armazenadas, não conseguimos identificar de forma fácil, como podemos aproveitar os dados e a melhor forma de fazê-lo, agora imagine uma empresa, independentemente de seu tamanho, e todos os dados que a mesma tem que armazenar e ao mesmo tempo utilizar. Se esta é uma tarefa dura para pessoas físicas imagina para empresas que necessitam armazenar megabytes, gigabytes e até mesmo *terabytes* de informações todos os dias. [Jornal Conexão Sebrae/MS]

Muitas pessoas tentaram desvendar a quantidade de informação e dados que estão

dispostos a nosso redor, tentando mensurar seu crescimento, onde cada, atingiu graus de sucesso mensurado, devido ao fato de medirem elementos distintos. Um dos estudos mais abrangentes foi realizado por Martin Hilbert, da *Annenberg School for Communication and Journalism*, da *University of California*. Ele tentou quantificar tudo que já foi produzido, armazenado e comunicado, o que incluiu dados de diversos tipos como, livros, imagens, e-mails, fotografia, música e vídeo (analógico e digital), vídeo games, ligações telefônicas e até mesmo cartas e sistemas de navegação para carros. [Schonberger, Viktor Mayer – *Big Data*]

De acordo com Hilbert, existiam em 2007 mis de 300 exabytes de dados armazenados. Para compreender melhor o que isto significa podemos colocar como exemplo a seguinte dimensão: pense que um filme digital pode ser comprimido em 1 gigabyte. Um exabytes seria o equivalente a um bilhão de gigabytes, resumindo em poucas palavras, isto é muito. [Schonberger, Viktor Mayer – *Big Data*]

O interessante desta pesquisa de Hilbert é que em 2007, apenas 7% dos dados eram analógicos (papel, livro, fotografia...) e apenas o restante era digital. Porém a pouco tempo o cenário mudou. Apesar de termos como ‘Revolução da informação’ e ‘era digital’ estarem presente desde os anos 1960, elas apenas se tornaram parte da realidade, sendo deixadas de lado. [Schonberger, Viktor Mayer – *Big Data*]

Evoluindo um pouco mais o tempo, no ano 2000, apenas um quarto da informação armazenada no mundo era digital. Mas como os dados digitais e expandiram rapidamente, de acordo com Hilbert, os mesmos se dobrariam em pouco mais de três anos. E assim em 2013, podemos verificar que a quantidade de informação armazenada no mundo é estimada em 1200 exabytes dos quais menos de 2% se encontram em seu estado analógico. [Schonberger, Viktor Mayer – Big Data]

Na era da comunicação digital, onde os dados podem ser compartilhados e distribuídos, de uma ponta a outra do planeta em rápidas frações de tempo, a posse da informação tornou-se o ativo mais importante das empresas. Em alguns casos as mesmas se tornam o bem mais valioso do empreendimento, sendo até mesmo mais valorizado do que o patrimônio físico, como imóveis e veículos. [Jornal Conexão Sebrae/MS]

Dificilmente uma empresa consegue sobreviver no mercado competitivo e globalizado sem coletar e gerenciar as informações relativas aos seus clientes, fornecedores, concorrentes, indicadores econômicos e sociais, produtos e serviços, exportações e outras que impactam direta ou indiretamente em seu negócio. Mas e quando as informações tomam dimensões gigantescas e as empresas não possuem, ou mesmo não sabem, que estão perdendo dados vitais para seus negócios ou então, que seus bancos poderiam ser muito melhor aproveitados se

vissem os dados de outra forma. Foi analisando e estudando estes casos que o termo *Big data* surgiu com o intuito de fazer com que as informações armazenadas sejam melhor aplicadas e utilizadas no dia a dia das empresas e dos indivíduos. [Jornal Conexão Sebrae/MS].

4 O QUE HÁ POR TRAZ DO TERMO BIG DATA

Como apresentado nos capítulos anteriores, o mundo em que vivemos esta inundado de dados, de diversas formas em diversos lugares. Onde a cada momento eles se proliferam em uma velocidade gigantesca. Sendo apresentados a cada vez mais dados, dispomos de cada vez mais dos mesmos, para utilizar com fins de melhorar a tomada de decisões em diversos setores da sociedade. Se não for possível explorar estes dados para uma melhor tomada de decisão, é possível dizer que as informações coletadas foram desperdiçadas, tornando o desempenho aquém de ótimo. Por esta razão utilizamos os paradigmas que veem junto ao *Big Data* para que obtenhamos o melhor proveito junto aos dados trabalhados. [Thomas H. Davenport, Dados Demais!]

4.1 Definição

O conceito de *Big Data* começou a ser discutido a cerca de 70 anos atrás, quando a produção escrita da humanidade crescia a taxas exponenciais, entretanto o termo só foi criado há 17 anos, onde se referia a impossibilidade de armazenamento de grandes volumes de informação. [Marcos Vieira – 2014]

Existem diversas definições do que seja realmente *Big Data*, podendo ser mais ampla ou mais detalhada. Fazendo uma junção de ideias com o que diversos autores falaram, os autores Stuart Ward e Barker puderam traduzir seu significado a “Uma das ideias mais importantes da informática, um termo que descreve o armazenamento e análise de data sets grandes e/ou complexos usando uma série de técnicas para sua aplicação e funcionamento”.

De uma forma mais simples, é possível definir este termo como uma análise de uma grande quantidade de dados para a geração de resultado importantes que, em volumes menores, dificilmente seriam alcançados. Por se trabalha com uma quantidade de dados muito ampla, é necessária a utilização de ferramentas especiais, feitas com o intuito de serem utilizadas para manejar estes montantes de dados, sempre com o intuito de se retirar todo o tipo de informação que possa ser encontrada, para que assim se possa analisar a mesma e aproveita-la em tempo real. [Emerson Alecrim – 2013/2015]

Porém, apesar de toda esta teoria, sobre qual a verdadeira definição do que é *Big Data*, o autor Marcos Vieira fala que na prática, ‘*Big Data*’ apenas define informações de uma natureza específica, ‘*Big Data*’ não fala sobre como utilizar essa informação, com que agilidade

ela deve ser manipulada ou que tratamentos estatísticos ela deve receber. O termo define uma situação no mundo real onde existe um problema, mas não uma solução. Por isso não faz sentido um projeto de '*Big Data*', mas sim projetos que resolvam pontos específicos desse universo, utilizando os conceitos trazidos pelo *Big Data*. [Emerson Alecrim – 2013/2015]

Ao final, podemos dizer que o mundo muda rapidamente. A quantidade de informação armazenada cresce quatro vezes mais rápido que a economia mundial, enquanto a capacidade de processamento dos computadores cresce quatro vezes mais rápido. Todos são afetados pelas mudanças, e o *Big Data* esta fadado a balar tudo, dos negócios a ciência e saúde, governo, educação, economia, ciências humanas e todos os demais aspectos da sociedade. [Schonberger, Viktor Mayer – *Big Data*]

4.2 Os V'S do Big Data

Assim como mencionado o início deste trabalho, foi com intuito de deixar a ideia de *Big Data* mais simples e clara de ser compreendida que *Gartner* e diversos estudiosos, ao se dirigirem ao termo o relacionaram aos tão conhecidos 'Vs'; onde a partir deste pequeno resumo foi satisfatoriamente possível descrever este termo tão extenso.

A quantidade destas palavras descritivas vem sofrendo diversas alterações com o decorrer do tempo. Onde inicialmente o termo de *big data* podia ser descrito utilizando os originais três Vs, podemos nos dias atuais encontrar autores mencionado quatro ou até mesmo cinco Vs. O mais comumente falado é na utilização de quatro destas palavras, sendo elas - volume, velocidade e variedade e valor aparecendo posteriormente.

Figura 2: Os quatro V's atuais
Fonte: Microsoft

4.2.1 Volume

O aspecto de volume já nos é familiar, ele está relacionado a gigas, teras, petabytes de informações. Neste campo estamos abrangendo x quantidade de dados, os mesmos que são gerados por funcionários, clientes, fornecedores e máquinas em qualquer empresa todos os dias.

Porém neste aspecto é necessário se tomar muito cuidado, pois o volume é muito abrangente, onde para cada caso é necessário todo um trabalho de reconhecimento, pois cada projeto de *Big Data* terá características especificas, podendo este tema de volume ser sofrer alterações gigantescas de um caso para outro.

Através do livro Dados Demais de Thomas H. Davenport, podemos verificar está singularidade através de alguns simples exemplos:

- Trinta bilhões de unidades de conteúdos foram acrescentadas ao facebook este mês, por mais de 600 milhões de usuários;
- Mais de 2.5 bilhões de mensagens de texto foram enviadas por dia em 2009;
- Os usuários do YouTube veem mais de 2 bilhões de vídeos por dia;
- A Cisco Systems estimou que, no fim de 2011, 20 domicílios típicos geraram mais tráfego na internet que todos os usuários de internet em 2008, entre outros.

4.2.2 Velocidade

Este aspecto está relacionado diretamente com um dos intuitos básicos de *Big Data*. Sempre lembramos que cada projeto possui suas características, porém por este tema estar relacionado diretamente ao processamento e demonstração dos dados que estão sendo enviados/recebidos em tempo real, a velocidade de tratamento e disponibilização das informações é de vital importância para que um projeto de *Big data* seja bem-sucedido. Desta form, quando falamos em Volume é necessário manter em mente que o tratamento dos dados (obtenção, gravação, atualização, enfim) deve ser feito em tempo hábil - muitas vezes em tempo real.

Se o tamanho do banco de dados for um fator limitante, o negócio pode ser prejudicado: imagine, por exemplo, o transtorno que uma operadora de cartão de crédito teria - e causaria - se demorasse horas para aprovar um transação de um cliente pelo fato de o seu sistema de segurança não conseguir analisar rapidamente todos os dados que podem indicar uma fraude, com este simples exemplo podemos verifica que o volume de informação gerado é crescente e a janela de tempo para a tomada de decisão é cada vez menor.

4.2.3 Variedade

Este tema já é conhecido quando falamos em *Big Data*, afinal um dos pontos mais importantes deste paradigma está relacionado a variedade de dados que é vista circulando por todos os pontos com conexões digitais na atualidade.

É possível encontrar nos dias de hoje dados de duas formas: em formato estruturados, isto é, armazenados em bancos como PostgreSQL e Oracle, e não estruturados, oriundos de inúmeras fontes, como exemplo de fontes podemos citar os e-mails, posts em blogs, arquivos de áudio, alguns tipos de vídeos, imagens, dados de GPS, páginas web, entre outros.

Como estes dados podem ser encontrados de ‘n’ maneiras um dos grandes desafios é como fazer para armazená-los e como uma tarefa ainda mais árdua como fazer para interpretá-los e analisá-los. Desta forma é necessário saber tratar a variedade como parte de um todo - um tipo de dado pode ser inútil se não for associado a outros.

4.2.4 Veracidade

O aspecto de veracidade deve ser sempre lembrado e considerado quando nos referimos a projetos de *Big Data*. O porquê? Não adiantaria termos a combinação de “volume + velocidade + variedade” se não pudéssemos garantir que os dados extraídos não pudessem ser cofiáveis.

Para que exista o máximo de certeza que as informações que serão utilizadas contenham consistência e verdade é necessário que se haja uma série de processos, que garantiram uma a certeza dos dados.

Como um exemplo para demonstrar a teoria da veracidade, imaginemos um exemplo com uma operadora de cartão de crédito, tente imaginar o problema que a empresa teria se o seu sistema bloqueasse uma transação genuína por analisar dados não condizentes com a realidade.

4.2.5 Caso a caso

É importante ter em mente que estes aspectos apresentados não precisam ser tomados como sendo obrigatório em todo projeto, esta pode não ser a definição perfeita.

Existem fontes que creditam que combinações diferentes, poderiam atender e transmitir uma noção aceitável do *Big Data*. Sob este conceito, é entendível que nem todos os aspectos podem estar presentes, podendo deixar valor ou veracidade por fora, por não serem essenciais aos resultados finais, por exemplo. Toda esta análise deverá ser feita de acordo com o que se

busca ao utilizar *Big Data* em um projeto. Como mencionado acima, muitos autores defendem a falta de veracidade e do valor como sendo desnecessários, pois tomam como princípio que já estejam implícitos no negócio - qualquer entidade séria sabe que precisa de dados consistentes; nenhuma entidade toma decisões e investe se não houver expectativa de retorno.

Seguindo este conceito, de não possuirmos todos os 4 Vs presentes no escopo de um projeto de *Big Data*, o autor Emerson Alecrim adiciona, os três primeiros 'Vs' - volume, velocidade e variedade - podem até não oferecer a melhor definição do conceito, mas não estão longe de fazê-lo. Entende-se que *Big Data* trata apenas de enormes quantidades de dados, todavia, você pode ter um volume não muito grande, mas que ainda se encaixa no contexto por causa dos fatores velocidade e variedade.

5 ANÁLISE E MINERAÇÃO DE DADOS: PRINCÍPIOS E MÉTODOS

Assim como observado nos capítulos anteriores, o cenário para utilização de *big data* pode ser encontrado em áreas diversas, da ciência a saúde, das finanças a internet, independente do quão diferentes possam ser, mas todos juntos encontram uma barreira comum: a quantidade de dados gerada diariamente em vários domínios de aplicação como, por exemplo, da Web, rede sociais, redes de sensores, dados de sensoriamento, entre diversos outros, estão na ordem de algumas dezenas, ou centenas, de Terabytes e os mesmos crescem rapidamente e superam não apenas nossas máquinas como nossa imaginação.

A partir deste princípio, se torna cada vez mais visível que a revolução não está apenas nas máquinas que calculam os dados, mas também nos dados em si e na maneira como os usamos.

Como inicialmente, não foi definido uma explicação rigorosa para o termo, porém, como este sempre esteve acompanhado da ideia de que o volume da informação crescera tanto que a quantidade examinada já não cabia mais na memória de processamento dos computadores, gerou diversos novos grandes desafios na forma de manipulação, armazenamento e processamento de consultas em várias áreas de computação, e em especial na área de bases de dados, mineração de dados e recuperação de informação. [L. C. da Silva, Ticiana - 2013]

Quando nos referimos a mineração de dados, temos como definição inicial que o mesmo consiste em um conjunto de técnicas que visa descobrir/encontrar conhecimento em grandes bases de dados, onde estas técnicas se baseiam em modelos que possuam a capacidade de reunir, de maneira resumida, os principais indicativos, assuntos e informações dos dados, extraindo conhecimento ou até mesmo realizando predições. [L. C. da Silva, Ticiana - 2013]

Por estes e outros motivos engenheiros tiveram de aprimorar os instrumentos que utilizavam para análise. Esta é a origem de novas tecnologias e algoritmos de processamento, como o processo de Data Mining. [L. C. da Silva, Ticiana - 2013]

Nos capítulos iniciais é apresentado também que um dos maiores desafios do *Big Data* é que existe um montante grande de volume de dados a ser analisado e que o processamento destes está além da capacidade computacional de uma máquina individual, o que faz com que se deva agregar aos cenários de *Big Data* novos modelos computacionais eficientes. Foi com a intenção de suprir um pouco este cenário que a computação baseada em *cluster* surgiu, pois a mesma permite trabalhar com o aproveitamento de um número considerável de processadores em paralelo. Com a computação em cluster se tornando mais interessante, se iniciou um maior

esforço em estudar ferramentas para uma melhor utilização do mesmo, foi neste aspecto que o framework *MapReduce* ganhou destaque, assim como sua implementação de código aberto, o *Hadoop*. Este se tornou uma bela opção pois o mesmo oferece um modelo simples por meio do qual os usuários podem expressar programas distribuídos relativamente sofisticados. [L. C. da Silva, Ticiana - 2013]

Existem algumas técnicas de mineração de dados, como a classificação, que fica categorizada como aprendizagem supervisionada. Os algoritmos que implementam esta técnica são do tipo preditivos, possuem este nome pois sua tarefa de mineração induz buscas com o intuito de retornar padrões, previsões ou tendências de dados, que por sua própria amostragem bruta não apresentam estes resultados. [L. C. da Silva, Ticiana - 2013]

Além da técnica mencionada acima, existem dois outros modelos muito conhecidos, sendo respectivamente a *Clusterização* e a Associação, porém estes ao contrário da Classificação, não possuem suas amostras conhecidas e suas definições como o número ou conjunto de classes de amostras conhecidos, fazendo com que os mesmos sejam classificados como um modelo de aprendizagem não-supervsionado. [L. C. da Silva, Ticiana - 2013]

Porém, é necessário mencionar que para que as técnicas mencionadas acima possam ser utilizadas, diversos autores, como [Ji et al. 2012] e [Cohen et al. 2009, Herodotou et al. 2011], citam um conjunto de boas práticas a serem seguidas nos algoritmos a serem utilizados, que serão tratados na próxima sessão.

5.1 Boas práticas para técnicas de Big Data

De forma geral, não existe fontes fixas que comprovem quais as definições e requisito estas técnicas e algoritmos devam possuir, porém, com o passar dos anos e da realização de projetos envolvendo o conceito de *Big Data*, foi-se verificando que algumas funcionalidades ou requisitos estão mais propensos a estarem presentes ou serem necessárias para o bom funcionamento de tais aplicações. [L. C. da Silva, Ticiana - 2013]

De acordo com [Cohen et al. 2009, Herodotou et al. 2011], é possível eleger seis princípios básicos que fazem com que um sistema possa ser utilizado em *Big Data*, onde estes são mais comumente conhecidos por ‘MADDER’, que vem a ser uma sigla consistida da letra inicial de cada princípio (sigla composta pelos nomes na língua inglesa), sendo eles: *Magnetic* (Magnetismo), *Agile* (Agilidade), *Deep* (Profundidade), *Data lifecycle awareness* (Consciência do Ciclo de Vida de Dados), *Elasticity* (Elasticidade) *e Robustness* (Robustez).

Podemos definir cada um deles como sendo:

- *Magnetic* (Magnetismo) – Um sistema magnético é capaz de trabalhar com qualquer fonte de dados, independente de possíveis discrepâncias para com os mesmos, além de haver a possibilidade da existência de esquemas desconhecidos ou da falta de estrutura, além da ausência de valores para alguns atributos.
- *Agile* (Agilidade) – Um sistema ágil tem como característica se adaptar a cotidiana e rápida evolução dos dados.
- *Deep* (Profundidade) – Um sistema de profundidade suporta necessidades analíticas mais profundas, que vão além, fazendo o uso de modelos estastísticos e técnicas de aprendizagem de máquina complexa.
- *Data lifecycle awareness* (Consciência do Ciclo de Vida de Dados) – Um sistema de com consciência de ciclo de vida dos dados, busca otimizar a movimentação de dados, processar e armazenar dados de *Big Data*, durante o ciclo de vida dos dados. Neste modelo, como a quantidade de dados circulantes é muito grande, o mesmo é necessário para:
    - Eliminar cópias de dados indiscriminadas que fazem com que as necessidades de armazenamento se inflem;
    - Reduzir gastos de recursos e realizar ganhos de performance devido à reutilização de dados intermediários ou metadados.
- *Elasticity* (Elasticidade) – Um sistema elástico é capaz de ajustar o uso de recursos e de custos operacionais aos requisitos demandados pelo usuário e do processamento da carga de trabalho.
- *Robustness* (Robustez) – Um sistema robusto é capaz de permanecer fornecendo serviço, mesmo que possíveis adversidades ocorram, como falhas de hardware, erros de software e corrupção de dados.

5.2 Fases de um processo de mineração

Tendo em mente as boas práticas apresentadas na sessão 5.1, é necessário mencionar que a Mineração de Dados não depende apenas das técnicas, métodos e algoritmos usados. Para o bom funcionamento, análise e aproveitamento dos dados, o processo de mineração deve ser dividido em fases que tem coo objetivo definir e padronizar as atividades da mineração. Atualmente existem diversos processos que executam esta tarefa e apesar das particularidades de cada, de uma forma geral, todos contêm a mesma estrutura. Nesta sessão, foi escolhido o

modelo CRISP-DM (*Cross-Industry Standard Processo of Data Mining*) para visualização, pois o mesmo é considerado na área, por diversos autores, como sendo o padrão de maior aceitação. [Amorim, Thiago - 2006]

O modelo CRISP-DM é composto por seis (6) fases, e apesar de sua composição, este não é um fluxo unidirecional, podendo ser alternado no decorrer de suas fases. A figura 3, pontua as seis fases existentes neste método. [Amorim, Thiago - 2006]

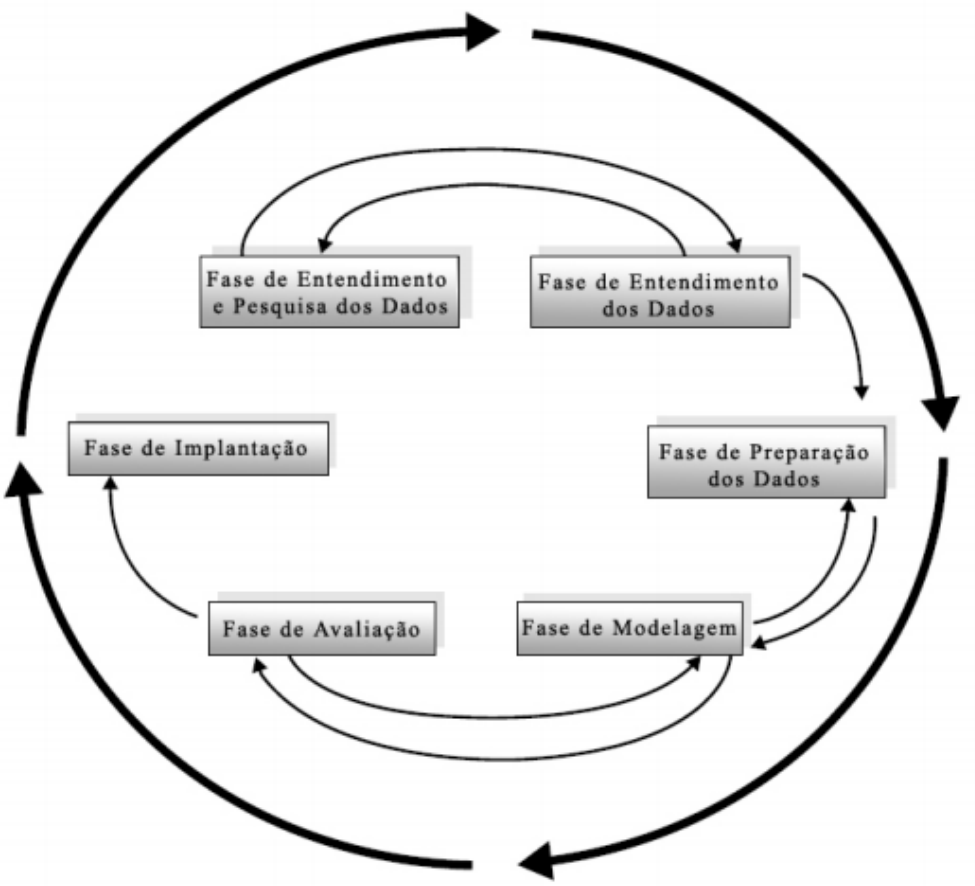

Figura 3: Fases do processo CRISP-DM
Fonte: [Oliveira Camilo, Cássio - 2009]

5.2.1 Entendimento e Pesquisa dos Dados

Nesta primeira fase, é obtido como ponto chave o entendimento do negócio, adquirindo um conhecimento sobre os objetivos do mesmo e suas definições, além de se buscar compreender qual destes objetivos devem ser atingindo com a mineração de dados. Após este levantamento de informações e conhecimento, o intuito é converte-los para uma definição de um problema de mineração de dados, traçando então um plano inicial para que no decorrer do processo os pontos focais traçados do projeto sejam alcançados. [Amorim, Thiago - 2006]

5.2.2 Entendimento dos Dados

Nesta fase o que se busca é compreender os dados, familiarizando-se com os meios de armazenamento e com os mesmos, pois existe a possibilidade de estes serem derivados de diversos locais e não possuírem uma formatação padrão, sendo fornecidos em diversos formatos. Nesta etapa acontece também uma coleta inicial dos dados mencionados, visando em todo momento o conhecer dos mesmos, para que assim seja possível identificar possíveis problemáticas e a qualidade das amostras. É comum também as técnicas de agrupamento e exploração visual sejam utilizadas nesta etapa. [Amorim, Thiago - 2006] [Oliveira Camilo, Cássio - 2009]

5.2.3 Preparação dos Dados

Esta fase é nomeada por muitos como sendo a mais exigente em termos de esforço, pois seu foco está em preparar os dados, visando a limpeza, transformação, integração e formatação dos mesmos. Este processo se faz necessário pelo fato de ser possível que os dados se originem de diversas formas diferentes, desta forma as ações citadas acima são necessárias. Nesta etapa geralmente se envolve os métodos de filtragem, combinação e preenchimento de valores não existentes. [Amorim, Thiago - 2006] [Oliveira Camilo, Cássio - 2009]

5.2.4 Modelagem

É neste ponto onde são aplicados os algoritmos de Mineração de dados, conjuntamente com a escolha das técnicas de modelagem a serem utilizadas. Comumente é possível se utilizar várias técnicas para mesmo problema de mineração, porém aluas das mesmas possuem requerimento específicos no formato dos dados, para estes casos, o retorno para a etapa de preparação dos dados se faz necessário. [Amorim, Thiago - 2006]

A grande maioria das técnicas de mineração existentes tem como base conceitos de aprendizagem de máquina, podendo ser: reconhecimento de padrões, estatística, classificação e *clusterização*. [Amorim, Thiago - 2006] [Oliveira Camilo, Cássio - 2009]

5.2.5 Avaliação

Por esta ser considerada como crítica, é necessária a presença de especialistas nos dados, conhecedores profundos das regras de negócio e de pessoal especialistas em tomada de decisão. Estes cuidados são necessários pois, a fase de avaliação tem como objetivo garantir que o modelo gerado ao fim desta etapa, atenda às expectativas e necessidades da organização. Porém deve-se manter e mente que se deve realizar uma análise criteriosa, pois a partir da mesma, se pode identificar a necessidade de se retornar a alguma das fases anteriormente mencionadas

De acordo com [HAN, J; KAMBER, M. - 2006], é nesta etapa que são realizados os testes e as validações necessárias, pois, o intuito final é a obtenção da confiabilidade nos modelos gerados, é requerido a execução de indicadores para que seja possível realizar a análise dos resultados, obtendo de retorno o esperado (matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística *kappa*, erro médio absoluto, erro relativo médio, precisão, *F-measure*, dentre outros)

5.2.6 Implementação

Nesta etapa, finalmente o projeto é implementado e este pode ser algo simples, como apenas o ganho de conhecimento ou uma ação como a geração de um macro simples. EM diversos casos, esta etapa estará ligada diretamente como cliente, que são os mesmos que realizaram a tarefa de execução para validação final, porém é importante mencionar que apesar de o analista não está ligado em primeira pessoa a esta etapa o mesmo possui como função de fazer com que o ciente compreenda. [Oliveira Camilo, Cássio - 2009]

5.3 Técnicas de Mineração de Dados

Atualmente, existem algumas técnicas para se realizar o processo de mineração de dados, onde estas suprem todas as outras formas de apresentação, permitindo assim que seja possível ter uma visão geral de um todo de forma apropriada ao tema. Sendo elas, as técnicas de classificação, a estimativa, a previsão, a análise de afinidades e Agrupamento (*Clustering*), onde estas serão, nas próximas sessões, definidas. [Carvalho, 2005].

5.3.1 Classificação

O método da Classificação é uma das técnicas mais utilizadas quando falamos em mineração de dados, isso se deve pelo fato de que o ato de se classificar, vem a ser uma das tarefas mais comuns para o ser humano no seu cotidiano, pois a todo instante o homem está lidamos com o classificar e compreender o meio em que está. [Oliveira Camilo, Cássio - 2009]

Quando nos referimos ao processo de classificar, no contexto de mineração de dados, a classificação se torna especificamente voltada a atribuição de uma das classes pré-definidas pelo analista, a novos fatos ou objetos que são apresentados à classificação. Esta técnica pode ser utilizada para entender dados existentes, assim como forma de predição de como novos dados irão se comportar no ambiente [Euriditionhome, 2004].

Como classificação, tem como objetivo identificar a qual classe um registro e especifico pertence, este método analisa o conjunto de entrada de dados, onde cada um destes possui uma indicação de a qual classe o mesmo pertence, isto se deve pelo fato de que este métoo tem como fim ‘aprender’ a classificar novo registros (aprendizado supervisionado). [Oliveira Camilo, Cássio - 2009]

Por mais semelhante que as classes sejam, na teoria, estas não são exatamente iguais, por esta razão para que as mesmas possam ser criadas de forma correta é necessário desprezar detalhes mantendo apenas as características principais. [Oliveira Camilo, Cássio - 2009]

A tarefa de se classificar, está em sua maior parte ligada a comparação de dois objetos/dados ou objetos/dados que supostamente pertençam a classes pré-definidas. Para que seja possível a comparação destes se faz a utilização de medidas de diferenças entre estes.

Quanto aos modelos para processamento, este método utiliza algoritmos que realizam as predições mencionadas. Os mais conhecidos são árvore de decisão [Quinlan 1986], redes bayesianas [Michie et al. 1994] e os vizinhos mais próximos [Cover and Hart 1967].

Como exemplo, podemos citar que este modelo pode ser utilizado para:

- Determinar se uma transação de cartão de crédito posse ser fraude;
- Identificar qual a turma mais indicada para um aluno em especifico;
- Diagnosticar onde uma determinada doença pode estar presente;
- Identificar quando uma pessoa pode ser uma ameaça para a segurança.

5.3.2 Estimativa

O método de Estimativa está associado a respostas continuas fazendo assim com que esta seja o contrário do método de Classificação mostrada acima. A Estimativa está associada ao ato de se estimar algum índice, que consiste em se determinar um valor mais provável diante de dados históricos ou de índices coincidentes, desta forma a arte de se estimar, simplificadamente falando, é determinar da melhor forma possível um valor, com base em outros que possuam uma situação similar. [Amorim, Thiago - 2006]

Neste método, são comumente utilizados os algoritmos de regressão e de redes neurais.

O método de Estimativa pode ser utilizado para:

- Estimar o valor gasto por uma família de três pessoas no período de volta às aulas;
- Estimar a pressão ideal de um paciente com base em idade, sexo e massa corporal.

5.3.3 Previsão

O método de Previsão é ligeiramente parecido com os modelos de Classificação e Estimação, é um modelo muito utilizado em Data mining e está associado a encontrar um valor futuro de um determinado atributo, baseando-se em dados de seu comportamento no passado.

Como um ponto não muito favorável temos o fato de que a única forma de se verificar se a previsão foi bem-feita é esperar o decorrer dos acontecimentos e verificar o qual foi a taxa de assertividade da previsão realizada. Sem dúvida, a previsão é uma das tarefas mais difíceis.

Neste método, são comumente utilizados os algoritmos de redes neurais, regressão e árvores de decisão, porém é possível a utilização de outros, como alguns métodos de classificação e regressão podem ser usados para predição, com as devidas considerações.

O método de Previsão pode ser utilizado para:

- Prever o valor de uma ação três meses no futuro;
- Prever o percentual que de aumento de tráfego na rede se a velocidade aumentar;
- Prever o vencedor do campeonato com base na comparação das estatísticas dos times.

5.3.4 Análise de Afinidades

O modelo de Análise de Afinidades tem como foco reconhecer padrões de ocorrências simultâneas em determinados eventos nos dados em que se está focando a análise. Ele busca determinar que fatos estão razoavelmente mais propensos a acontecerem ou quais itens de um montante de dados estão interligados, com certa chance de correlação.

Quanto aos algoritmos, de acordo com [Pelegrin et al., 2005], a utilização das regras de associação constitui-se no procedimento mais utilizado nestes casos.

Como exemplo de utilização para este método, é possível citar o exemplo mais clássico de análise de afinidades, o do carrinho de supermercado, onde desejamos ter o conhecimento de quais produtos são comprados em conjunto com mais frequência, pelos consumidores.

5.3.5 Análise de Agrupamento (*Clusterização*)

O modelo de Análise de Agrupamentos ou *Clusterização* tem como objetivo identificar e aproximar os objetos ou elementos similares, formando grupos homogêneos entre si podendo este número de grupos ser previamente estabelecido ou então se pode admitir ao algoritmo de agrupamento uma livre associação de unidades, onde desta forma, este número apenas será conhecido ao final de processo. [L. C. da Silva, Ticiana - 2013]

Esta tarefa difere do método da Classificação, pois uma clara diferença entre estes dois é que na Classificação as classes são pré-definidas pelo técnico, enquanto que na *Clusterização* não existe este pré-requisito, isto torna este método um modelo do tipo de aprendizado não-supervisionado. Em soma a esta diferença, podemos citar ainda, que este modelo não busca classificar, estimar ou predizer o valor de uma variável, seu objetivo é apenas identifica os grupos de dados similares. [L. C. da Silva, Ticiana - 2013]

Definindo um pouco mais, podemos citar o método *k-means* como sendo o método mais popular da *Clusterização*, além de ser um dos algoritmos mais importantes quando tratamos de mineração de dados [Wu et al. 2008]. Entretanto, quando nos referimos a qualidade e eficiência, a ordem de execução do *k-means* pode ser exponencial no pior caso. Porém, para sanar este quesito de qualidade e eficiência, foi proposto uma evolução do mesmo que dispões para os usuários uma melhor estratégia de inicialização dos centroides, sendo ele o algoritmo *k-means*++. [Arthur and Vassilvitskii 2007]

Permanecendo no quesito de algoritmos de *clusterização*, podemos citar como um dos mais importantes na atualidade sendo o DBScan(Density-based Spatial Clustering of

Application with Noise) [Ester et al. 1996]. Sua vantagem em relação aos outros algoritmos utilizados se encontra no fato de o mesmo agrupar dados em clusters de formato arbitrário, não sendo necessário assim o número de clusters anterior, além de lidar com valores exuberantes no conjunto de dados.

Em estudos como [He et al. 2011] e m [Dai and Lin 2012] e outros, é normalmente proposto uma implementação do *DBScan* com a utilização de *MapReduce*.

Podemos citar como exemplo de *clusterização*:

- Segmentação de mercado para um nicho de produtos específicos;
- Para a realização de auditorias, onde se é possível separar suspeitos de acordo com seus comportamentos;
- Reduzir para um conjunto de atributos similares registros com centenas de atributos.

6 INFRAESTRUTURA PARA ANÁLISE EM BIG DATA: TÉCNICAS PARA DESENVOLVIMENTO

Para tratar o problema de gerenciamento e análise dos dados em *Big Data*, diferentes soluções vem sendo propostas por especialistas e estudiosos, entre estas é possível se ouvir falar constantemente de computação em nuvem, paralelismo, bancos de dados *NoSQL*, técnicas como Data mining, *MapReduce* juntamente com *Hadoop*, além de sistemas baseados em estrutura de *arrays* multidimensionais ou *Distributed Hash Table (DHT)* [Dean and Ghemawat 2008], tais ferramentas foram desenvolvidas no intuito de tornar projetos de *Big Data* uma tarefa mais simples, para assim possibilitar que profissionais de diversas áreas possam fazer uso dela. Neste quesito, o mercado atual de ferramentas para análise em *Big Data* tem se tornado bastante atraente.

Nas sessões seguintes, é falado um pouco mais de algumas destas técnicas, onde estas, serão auxiliadoras no desenvolvimento do estudo de caso deste trabalho.

6.1 Data Mining

Data Mining é uma das novidades que vieram acompanhando o termo *Big Data* na área da Ciência da Computação, que assim como mencionado anteriormente, como o *Big Data*, o Data Mining também veio para ficar. Com a geração cada vez maior do volume de informação, se tornou essencial tentar aproveitar o máximo possível dos dados que estão circulando em todos os meios. Contudo, apenas recuperar informação não propicia todas as vantagens possíveis. [NAVEGA, SERGIO – 2002]

O processo de Data Mining permite que se investigue esses dados à procura de padrões que possuam valor. Desta forma, podemos dizer que a forma mais nobre de se utilizar esses vastos repositórios seja tentar descobrir se há algum conhecimento escondido neles e não apenas descobrir onde eles estão. [NAVEGA, SERGIO – 2002]

Porém como podemos definir Data Mining, de acordo com o *Gartner Group*, podemos definir este termo como sendo o processo de descoberta de significativas novas correlações, padrões e tendências de peneirar grande quantidade de dados armazenados em repositórios, utilizando tecnologias de reconhecimento de padrões, bem como técnicas estatísticos e matemáticos. [LAROSE, DANIEL T. – 2014]

Porém de uma forma mais simplificada podemos citar a definição de *Usama Fayyad* (Fayyad et al. 1996): "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis". [NAVEGA, SERGIO – 2002]

Nos dias atuais, existe uma variedade de ermos utilizados para se descrever este processo, incluindo *Analitics* analise preditiva, *Big Data*, aprendizado de máquina e descoberta de conhecimento em banco de dados; mas estes termos dispõem de um mesmo objetivo, o de realizar a mineração de partes de conhecimento de conjuntos de dados. [LAROSE, DANIEL T. – 2014]

Esse processo vale-se de diversos algoritmos (muitos deles desenvolvidos recentemente) que processam os dados e encontram esses "padrões válidos, novos e valiosos". É preciso ressaltar um detalhe que costuma passar despercebido na literatura: embora os algoritmos atuais sejam capazes de descobrir padrões "válidos e novos", ainda não temos uma solução eficaz para determinar padrões valiosos. Por essa razão, o processo de Data Mining ainda requer uma interação muito forte com analistas humanos, que são os principais responsáveis pela determinação do valor dos padrões a serem encontrados. Além disso, a o desenvolvimento da exploração de dados é também tarefa fundamentalmente confiada aos analistas, e este é um detalhe que não pode ser desprezado em nenhum projeto que queira ser bem-sucedido. [NAVEGA, SERGIO – 2002]

6.1.1 Os Passos do Data Mining

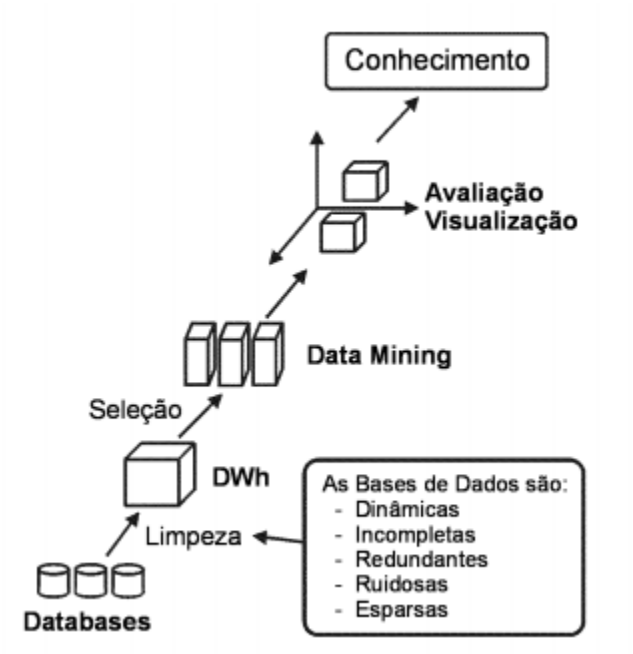

Figura 4: Passos do Data Mining
Fonte: Intelliwise - 2002

Neste capítulo é apresentado os passos fundamentais para uma mineração de dados bem-sucedida. De forma simplificada, o processo de mineração para a utilização de dados pode ser mostrado na imagem acima. Em seguida é descrito de forma simplificada o processo a ser executado.

Os dados são retirados de uma fonte de dados, podendo ser bancos de dados, relatórios, logs de acesso, transações, entre outros; a partir desta extração é realizada uma limpeza dos dados. Podem ser utilizados diversos tipos de limpeza, como: métodos de consistência, preenchimento de informações, remoção de ruído e redundâncias, etc. Disto se dão origem os repositórios organizados, que são mais conhecidos por *Data Marts* e *Data Warehouses*, onde estes podem ser úteis de diversas maneiras. Mas é a partir dos mesmos que se pode selecionar algumas colunas para atravessarem o processo de mineração. [LAROSE, DANIEL T. – 2014]

Tipicamente, este processo não é o final da história. Através da utilização de formas interativas e muito usualmente, visualização gráfica, os analistas refinam e conduzem cada vez mais estes processos, para que ao final os tão esperados e valiosos padrões apareçam.

É importante observar que todo esse processo parece indicar uma hierarquia, algo que começa em instâncias elementares e terminam em um ponto relativamente concentrado, mas muito valioso. [LAROSE, DANIEL T. – 2014]

Então surge a pergunta, o por que se fazer todo esse processo de limpeza? Para que obtenhamos uma resposta é importante que mantenhamos em mente o seguinte conceito, é necessário que se encontre padrões nestes dados, para isto é necessário que estes sejam sistematicamente “simplificados”, fazendo assim com que seja desconsiderado aquilo que é específico e então seja privilegiado aquilo que é genérico. Claro cada caso se definirá, mas porque este conceito é levado a diante, isto se deve porque na maioria dos cenários não parece haver muito conhecimento a extrair de eventos isolados.

Em outras palavras, não há como explorar profundamente eventos isolados, pois estas informações em particular, em sua maioria são apenas isso, particularidades.

Apenas com conhecimento genérico será possível fazer com que as empresas cresçam e lucrem, e este é o processo de como se obtiverá isto. Por essa razão devemos, em Data Mining, controlar nossa vontade de “não perder dados”. Para que o processo dê certo, é necessário sim desprezar os eventos particulares para só manter aquilo que é genérico.

6.1.2 Os dados e sua importância no processo de Data Mining

As empresas então a todo momento recebendo informações de todos os lugares, se portando assim como um organismo vivo, e estas necessitam atuar sobre tais dados. Neste processo, é necessário analisar as informações e verificar seus distintos níveis de importância e utilização, classificando as mesmas. [LAROSE, DANIEL T. – 2014]

Abaixo verificamos a figura 4, que contém a tradicional pirâmide da informação, onde é possível visualizar as etapas de abstração realizadas no decorrer de como subimos os níveis das informações após serem recebidas e analisadas:

Figura 5: Pirâmide da Informação tradicional
Fonte: Intelliwise - 2002

Quando trazemos as informações para uma empresa atual, esta pirâmide sofre alterações e começa a ser mostrado como a figura 5:

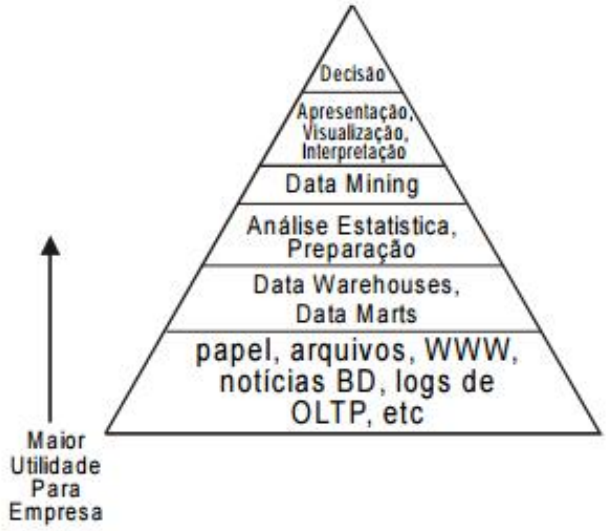

Figura 6: Pirâmide da Informação Aplicada a uma empresa
Fonte: Intelliwise - 2002

A característica principal que deve ser notada nesta nova pirâmide é a sensível redução que ocorre com o volume a cada etapa de subimos os níveis na pirâmide. Esta redução de volume é uma natural consequência do processo de abstração. [LAROSE, DANIEL T. – 2014]

Quando nos referimos a abstração neste processo, é referente ao sentido de representar uma informação através de correspondentes simbólicos e genéricos, este é um ponto importante, pois como já foi mencionado, quando buscamos o genérico, o "perder" uma parcela de dados se torna algo naturalmente necessário, para que busquemos assim conservar a essência da informação. É neste ponto que o processo de Data Mining se encaixa, pois, o mesmo tende com a utilização de seus algoritmos e técnicas a obtenção de localizar padrões através de processos de generalização, ou também conhecido como indução. [LAROSE, DANIEL T. – 2014]

6.1.3 Localizando padrões

Para que possamos entender o que buscar, vamos entender o que estamos buscando. Quando buscamos o genérico, estamos na maior parte do tempo buscando por padrões que nos levaram a estes dados, mas o que são padrões? Estes são unidades de informações que tendem a se repetir, ou então podem ser sequencias de dados que dispõem de uma estrutura que se repete.

A tarefa de localizar padrões não é privilégio do Data Mining. Nosso cérebro utiliza-se de processos similares, pois muito do conhecimento que temos em nossas mentes é um processo que depende da localização de padrões. Por este motivo, grande parte do que é estudado sobre o cérebro humano também pode nos se útil para entender o que deve ser feito para localizar esses dados repetidos.

Mas o que se consiste em ser localizar padrões? O que é indução? Para que possamos entender melhor estes conceitos, é apresentado a seguir um breve exemplo/exercício de uma indução de regras abstratas. Temos como objetivo obter algum tipo de expressão genérica para a sequência dada abaixo:

Sequência de dados simples: **ABCXYABCZKABDKCABCTUABEWLABCWO**

Neste exemplo, observamos a sequência de caracteres com o objetivo de tentar encontrar algo relevante. Ao uma análise da sequência, é possível visualizar algumas possibilidades, para isso, podemos considerar os seguintes passos:

- **Passo 1:** nesta primeira etapa podemos perceber que existe um conjunto de letras, na sequência de exemplo, que se repete bastante em sua extensão. É possível citar, por exemplo, as sequências “AB” e “ABC” e observar que as mesmas aparecem com frequência superior à das outras sequências o decorrer do dado.
- **Passo 2:** Após determinar as sequências “ABC” e “AB”, é possível dizer que elas segmentam o padrão original em diversas unidades independentes: “ABCXY” “ABCZK” “ABDKC” “ABCTU” “ABEWL” “ABCWO”.
- **Passo 3:** Nesta etapa, será dado início as Induções, que irão gerar algumas representações genéricas das unidades encontradas: “ABC??” “ABD??” “ABE??” e “AB???”.

No final desse processo, toda a sequência original será substituída por regras genéricas indutivas, que faz com que a informação original seja simplificada (reduzida), por algumas expressões simples. Este pequeno exemplo simples, tem como objetivo mostrar/apresentar um dos pontos essenciais do Data Mining, que vem a ser formas de como podemos fazer para extrair certos padrões de dados brutos. Contudo, mais importante do que simplesmente obter essa redução de informação, esse processo nos permite gerar formas de predizer futuras ocorrências de padrões. Desta forma chegamos exatamente o ponto onde todo este processo começa a mostrar o seu valor. [LAROSE, DANIEL T. – 2014]

Agora que iniciamos um exemplo de Data Mining, vamos nos aprofundarmos um pouco mais utilizando os caracteres apesentados no início desta sessão, partindo das expressões abstratas genéricas obtidas acima. Uma das mesmas mostra que toda vez que encontramos a sequência “AB”, podemos inferir que iremos encontrar mais três caracteres, mostrando que isso nos mostra um “padrão”. Na forma abstrata mostrada até o momento (sequência de caracteres) pode ficar difícil de perceber a relevância deste resultado, com intuito de melhor entendimento será utilizado uma representação mais próxima da realidade a partir deste momento. [LAROSE, DANIEL T. – 2014]

Adicionando valores reais ao nosso dado fictício, será imaginado que a letra 'A' esteja representando um item de um registro comercial, como exemplo vamos atribuir a esta o significado de “aquisição de pão”, considerando que nosso cenário seja realizado e uma transação de supermercado, adicionamos a letra 'B' o significado de “aquisição de leite” e pôr fim a letra 'C' é um indicador de que o leite que foi adquirido e o mesmo é do tipo desnatado. [LAROSE, DANIEL T. – 2014]

Com os significados mais reais para estes casos podemos analisar de uma forma mais significativa os dados, podemos agora notar que com estes dados uma regra com as letras “AB” possui o significado de que toda vez que alguém comprou pão, também comprou leite. Isto mostra que estes dois atributos estão associados e este fato nos foi revelado através do processo de descoberta de padrões. Com esta informação se pode pensar que a partir deste momento é interessante colocar o “leite” e o “pão” em locas próximos para venda, pois desta forma, a aquisição conjunta será facilitada. [LAROSE, DANIEL T. – 2014]

Vamos um pouco adiante neste exemplo, suponhamos que a letra ‘X’ tenha como significado “manteiga sem sal”, a letra ‘Z’ signifique “manteiga com sal” e a letra ‘T’ tendo como significado “margarina”. Parece que poderíamos tentar unificar todas essas letras através de um único conceito, uma ideia que resuma uma característica essencial de todos esses itens. Para isto, iremos introduzir a letra ‘V’, que significa “manteiga/margarina”, ou simplesmente, consideremos “coisas que passamos no pão”. [NAVEGA, SERGIO – 2002]

Com a ação acima, acaba de ser realizado uma indução orientada a atributos, que consistiu em substituímos uma série de valores distintos (porém com conceitos similares) por apenas um nome. Ao realizar esta ação, é possível notar que foi perdido u pouco das características dos dados originais, afinal, após essa transformação, já não é possível saber o que é manteiga e o que é margarina. Essa perda de informação é fundamental na indução e é um dos fatores que permite o aparecimento de padrões mais generalizados. E como proceder com essas alterações que a indução nos mostrou? Basta reajustar nossa sequência de cractéres original substituindo a letra V em todos os lugares necessários. [NAVEGA, SERGIO – 2002]

E então, após estas alterações, obteremos a seguinte sequência:

ABCVYABCVKABDKCABCVUABEWLABCVO

A partir deste momento, o sistema de Data Mining irá extrair, entre outras coisas, a expressão “ABCV”, que nos irá revelar mais um dado automaticamente: A maioria dos usuários que adquirirem pão e leite desnatado também adquirirem manteiga ou margarina. Co essa informação em mãos, fica mais fácil imaginar uma distribuição de produtos nas prateleiras de um supermercado, com o intuito de incentivar ainda mais o hábito comprovado pela análise realizada. [NAVEGA, SERGIO – 2002]

Em uma linguagem mais lógica, pode-se dizer que pão e leite estão associados (na aquisição de manteiga, podendo então, chegar a seguinte regra

(Pão, Leite) ⇒ Manteiga

O lado da esquerda desta expressão (Pão, Leite) é chamado de Antecedente, e o lado da direita de Consequente.

6.1.4 Técnicas de mineração utilizadas em Data Mining

No exemplo mostrado anteriormente, pode-se verificar que a partir da mineração e análise de dados é possível se fazer certas induções e se descobrir padrões. Nesta sessão é possível visualizar algumas técnicas que se utilizam de princípios similares.

É possível visualizar algumas destas técnicas e mais na sessão 5.3 deste trabalho.

6.1.4.1 Regras Caracterizadoras

Se realiza através da obtenção de regras que caracterizam um conceito satisfeito por todos (ou pela maioria) dos exemplos disponíveis. Assim, é possível descobrir formas de sumarizar certas características que podem revelar padrões nos dados.

Exemplos:

- Sintomas de uma doença específica podem ser sumarizados por uma regra caracterizadora;
- Geração de regras que caracterizem quais os estudantes de graduação que se decidiram por prosseguir com uma carreira acadêmica (MBA, doutorado).

6.1.4.2 Regras Discriminantes

Neste tipo de regra, o que se busca é obter regras que discriminem (separem) um conceito alvo em relação a outros conceitos (classes contrastantes).

Exemplo:

- Para distinguir uma doença, procura-se por regras que sumarizem as características que separam esta doença das outras.
- Tenta-se achar as regras que discriminem uma loja bem-sucedida de várias outras não tão bem-sucedidas.

6.1.4.3 Regras Associativas

Este é o tipo de regra que utilizamos anteriormente. Nesta se procura estabelecer regras que interliguem/unam um conceito a outro. A utilidade deste procedimento é muito grande, conforme pode ser visto nos exemplos abaixo[NAVEGA, SERGIO – 2002]:

- Achar todas as regras que tenham "Coca-Cola dietética" como consequentes. Isto irá auxiliar no planejamento de lojas para vender melhor este produto (privilegiam-se os antecedentes dessas regras).
- Achar todas as regras que tenham "iogurte" no antecedente. Isto irá auxiliar na determinação do impacto nas receitas, caso este produto seja retirado das prateleiras.
- Achar todas as regras com "salsicha" no antecedente e "mostarda" no consequente. Isto irá auxiliar na obtenção de melhores regras para determinar quais os itens que devem ser vendidos em conjunto com salsichas para aumentar as vendas de mostarda.

6.1.4.4 Regras de Evolução Temporal

Neste tipo de regra a preocupação é detectar associações entre itens ao longo do tempo. Descobre-se padrões de compras após um evento inicial de aquisição.

Exemplos:

- Consumidor comprou um desktop hoje, irá comprar um HD externo em 6 meses. Isto permite que se faça uma oferta desse produto a todos os que estão nesta situação.
- Um consumidor adquiriu uma máquina fotográfica, em 4 meses terá muita probabilidade de comprar um cartão de memória. Desta forma é possível se fazer uma promoção especial para estes clientes.

6.2 Computação em Nuvem

A computação em nuvem tem como enfoco proporcionar soluções com custo baixo de forma eficiente para o armazenamento de montantes de dados. Atualmente, existem diversas definições e conceitos para a computação em nuvem. Neste estudo, iremos utilizar a definição de [Mell and Grance 2009], onde de acordo com o mesmo, se pode definir computação em nuvem como sendo um modelo que provê acesso sob demanda a um conjunto de recursos

computacionais, onde estes podem ser configurado de acordo com as necessidades, como CPU, armazenamento, memória e outros.

Estes recursos podem ser fornecidos e liberados de forma rápida, utilizado o mínimo de esforço de gerenciamento ou assistência do provedor da nuvem.

É possível citar algumas propriedades que são fundamentais quando para se diferenciar a computação em nuvem de sistemas distribuídos convencionais (como sistemas em grade, *clusters* e P2P), sendo estas:

6.2.1 Serviço sob demanda

Este está relacionado a possibilidade do usuário utilizar os serviços computacionais dispostos, quando for necessário, sem nenhuma interferência humana para a distribuição do provedor de serviço.

6.2.2 Elasticidade rápida

Este, tem como foco permitir que mais recursos possam ser escalados de acordo com o crescimento da demanda. É a partir desta característica que se deriva a abstração para o consumidor, isto é, para o mesmo é mostrado como se o prover destes serviços pareçam muitas vezes ser ilimitados.

Os recursos fornecidos em sistemas na nuvem são controlados de forma automática e sua utilização é medida no nível de abstração que for apropriado para cada tipo de serviço

6.2.3 Pagamento de acordo com a utilização do serviço

Este modelo define o valor a ser pago pelo consumidor, que será estimado de acordo com a utilização dos serviços e por seu tempo. Onde está utilização pode ser monitorada, verificada, controlada e reportada, gerando assim total transparência para todos os envolvidos na utilização do serviço

6.2.4 Nível de qualidade de serviço (SLA)

Este é o modelo onde se garante que pelo menos será disposto o nível de qualidade de serviço mínimo ao usuário. Onde os recursos físicos e virtuais são dinamicamente atribuídos e reatribuídos conforme a demanda dos consumidores. [Coelho da Silva 2013]

6.2.5 Agrupamento ou Pooling de Recursos

Os recursos de computação de cada fornecedor são concebidos para servir vários clientes, num modelo *multi-tenant*, com diferentes recursos físicos e virtuais, distribuídos e alocados dinamicamente. Como exemplo podemos citar os recursos de armazenamento, processamento, memória, largura de banda de rede, entre outros.

6.3 MapReduce e Hadoop

MapReduce é um modelo de programação que possibilita que um montante de dados seja processado a partir de um algoritmo paralelo e distribuído, fazendo com que seja possível organizar o processamento de maneira a aproveitar as múltiplas máquinas de um *cluster*. Enquanto se mantém pelo tempo possível, do processamento e dos dados que ele precisada mesma máquina. [L. C. da Silva, Ticiana - 2013]

Quando nos referimos a manipulação dos dados de forma paralela, como no caso de leitura ou escrita de dados, dois problemas são frequentemente enfrentados:

I. Se o número de discos rígidos é *n* vezes maior, as chances de ocorrer uma falha é de *n* vezes maior, podendo ocasionar a perda de dados. Para que este problema seja sanado, é se utilizada a replicação, que consiste em se manter copias de segurança dos dados em diferentes discos;
II. O ouro problema enfrentado está relacionado ao fato de muitas das tarefas de análise de dados, tem a necessidade de se combinar dados ‘espalhados’ em diversos discos.

Porém com a utilização de *MapReduce*, os problemas citados não geram dor de cabeça, pois o mesmo oferece um modelo de programação que visa abstrair tais inquietudes. [L. C. da Silva, Ticiana - 2013]

O modelo de programação desta técnica é composto de um programa formado por duas operações básicas: *Map* e *Reduce*, onde a operação *Map* recebe um par chave/valor e gera um conjunto intermediário de dados, também no formato chave/valor, e a operação *Reduce* é executada para cada chave intermediária, com todos os conjuntos de valores intermediários associados àquela chave. De uma forma geral a operação *Map* é utilizada para encontrar algo dependendo apenas de um único registro, e a operação *Reduce* é utilizada para fazer a sumarização do resultado, onde recebe múltiplos resultados do mapeamento com a mesma chave e combina seus valores. [L. C. da Silva, Ticiana - 2013]

O esquema de funcionamento das operações de *Map* e *Reduce*, podem ser visualizados na Figura 7. [L. C. da Silva, Ticiana - 2013]

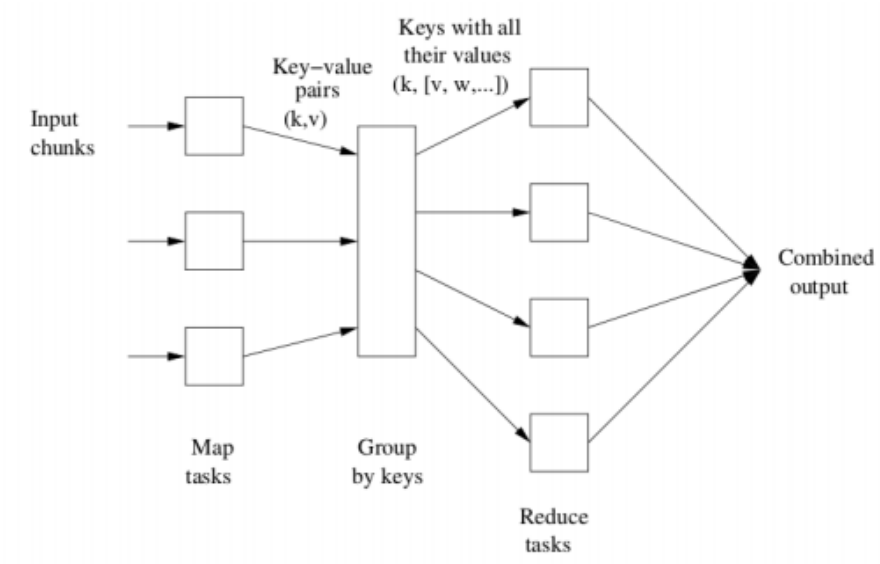

Figura 7: Funcionamento operações Map e Reduce
Fonte: [Rajaraman and Ullman - 2012]

6.3.1 Passo-a-passo da execução

1. A biblioteca *MapReduce*, no programa do usuário ira inicialmente dividir os dados de entrada em M pedações, após isto se iniciará o processo de enviar várias cópias do programa em um *cluster* de computadores.
2. Uma das cópias do passo anterior será considerada ‘especial’ – denominada cópia *master*, enquanto as outras serão chamadas de *workers* e serão comandadas pela *master*, que é a cópia que distribui os trabalhos. Existem M tarefas de *Map* e R tarefas de *Reduce* para serem assinaladas.
3. O *master* selecionará *workers* que estejam ociosos e os assigna uma tarefa de *map* ou de *reduce*.
4. Um *worker* que possui uma tarefa de *map* le o conteúdo que corresponde ao pedaço da entrada. Ele interpreta os pares chave/valor a partir dos dados de entrada e passa como parâmetro para a função de *map* do usuário. Estes pares de chave/valor intermediários que foram produzidos pela função de *map* são então armazenados em memória.
5. De tempos em tempos os pares de dados dos buffers são escritos em disco, particionados em R regiões pela função de particionamento. A informação do local de armazenamento

desses pares no disco é passada ao *master*, que repassar a mesma para os *workers* através das tarefas de *reduce*.

6. Quando um *worker* de *reduce* é informado da localização dos dados pelo *master*, o mesmo usa uma chamada de procedimento remota para buscar os dados do disco local dos *workers* de *map*. Quando os dados já foram lidos, ele ordena os dados pelas chaves intermediárias, para que todas as ocorrências de uma mesma chave sejam agrupadas juntas. Esta operação de ordenação se faz necessária, pelo fato de muitas chaves diferentes serem mapeadas para a mesma tarefa de *reduce*. É utilizada uma ordenação externa caso a quantidade de dados intermediários utilizados seja muito grande para caber em memória.
7. O *worker* de *reduce* passa pelos dados intermediários, já ordenados, e para cada chave que ele encontra, o mesmo passa a chave dos valores intermediários para a função de *reduce*. A saída de cada função de *reduce* é adicionada ao final de um arquivo de saída para aquela partição.
8. Quando todas as tarefas de *map* e *reduce* forem finalizadas, o *master* retorna o programa do usuário.

6.3.2 Tolerância a falhas

Pelo trabalho de processamento que o *MapReduce* executa, lidar com falhas é algo vital para o bom funcionamento de todo o projeto que utilizem o mesmo. O autor [Paiva, Ricardo - 2011] explica de forma simples o processo de administração de falhas, sendo este: inicialmente o processo *master* envia um *ping* periodicamente para cada *worker*. Caso o *master* não receba uma resposta durante um período de tempo, o mesmo assume que aquela máquina falhou. Todas as tarefas de *map* que foram completadas pelo *worker* são *resetadas* para seu *status* inicial e são escalonadas novamente para um outro *worker* responsivo. Aso haja alguma falha todas as tarefas necessitam ser refeitas, pois os arquivos de saída são armazenados no disco local da máquina que falhou, tornando os mesmos impossíveis de serem acessados. O mesmo acontece com as tarefas de *map* ou *reduce* que estão sendo executadas. Porem as tarefas *reduce* completadas não precisam ser executadas novamente, afinal os arquivos de saída destas são salvas no sistema global. [Paiva, Ricardo - 2011]

Já quando no referimos a falha do processo *master*, o conrole necessário é mais complexo, afinal o processo *master* é o elo entre a execução das tarefas de *map* e *reduce*. Para que seja feito este controle, o processo *master* deve executar checkpoints periódicos de suas estruturas de dados, caso uma destas falhe, uma nova instância pode ser levantada, recuperando o a saída anterior. Como o *MapReduce* pressupões que existe apenas um processo *master*, falhas no mesmo não são desejáveis. [Paiva, Ricardo - 2011]

Ao final da execução do programa, algumas máquinas, apesar de ainda responderem, podem apresentar um tempo de resposta muito inferior à média das outras máquinas. Por exemplo, falhas nos discos podem reduzir a taxa de leitura de 30MB/s para 1MB/s. Para evitar que isto atrase a execução do programa, algumas cópias das tarefas restantes são iniciadas (tarefas *backup*), quando o programa está perto de terminar. A tarefa será marcada como completada quando a tarefa primária ou uma tarefa de backup der retorno. [Paiva, Ricardo - 2011]

Quando o *MapReduce* é mencionado, é quase impossível não se falar também do *Hadoop*, pois de uma forma geral, foi o mesmo que trouxe o *MapReduce* como sendo uma solução para processamento paralelo de dados, por este motivo, o *Hadoop* será apresentado a seguir.

O *Hadoop* é um projeto que oferece solução para problemas relacionados a projetos de *Big Data*, onde ele é ser uma coleção de subprojetos relacionados a computação distribuída, sendo hospedado pela *Apache Software Foundation.* Seu núcleo é composto basicamente por duas (2) partes, o *Hadoop Distributed Filesystem* (*HDFS*), que vem a ser um sistema de arquivos distribuídos, que tem como tarefa o armazenamento dos dados, e a outra parte é claro é o *Hadoop MapReduce*, que se incube da análise e processamento dos dados. Ambos *HDFS* e *Hadoop MapReduce* possuem a característica de confiabilidade como marca, o que torna o *Hadoop* um sistema consistente e robusto para projetos que englobam dados vitais para as empresas que o utilizam. [Paiva, Ricardo - 2011]

Pelo fato de este modelo possuir código aberto, ele permite que suas funcionalidades sejam moldadas de acordo com a demanda de cada projeto.

Atualmente o Hadoop é vastamente utilizado em aplicações que envolvam o termo *Big Data* por grandes companhias, podendo citar como exemplo o *Facebook* e o *Twitter*,

Ainda que o *MapReduce* e o *HDFS* sejam os subprojetos mais conhecido do *Hadoop*, o mesmo oferece uma gama de serviços complementares ou que adicionem abstrações em um nível mais elevado. Estes subprojetos provem um leque de opções quando se necessita lidar com vários formatos de dados, criando além um grande nível de abstração para com o

desenvolvedor, estes atributos e características aumentam ainda mais sua capacidade e robustez. [Paiva, Ricardo - 2011]

É possível citar os seguintes subconjuntos, que possuem um maior destaque:

- **Avro -** é um sistema que tem por objetivo difundir os dados em série, são responsáveis por fornecer *RPCs (Remote Procedure Calls*) eficientes e independentes da linguagem utilizada, armazenamento persistente de dados, estruturas de dados poderosa e mais. Este subconjunto é muito utilizado com linguagens de programação dinâmica;
- **Pig** – vem a ser uma plataforma para grandes conjuntos de dados que possui uma linguagem de programação de alto nível para realizar a análise dos dados. Além disso, possui a infraestrutura necessária para avaliar os programas criados, como um compilador especial que transforma as aplicações desenvolvidas nessa linguagem em uma sequência de programas *MapReduce*;
- **HBase -** é uma base de dados distribuída, muito utilizada nos dias atuais, criada para armazenar enormes tabelas (milhões de colunas x bilhões de linhas). Quando o assunto é armazenamento o *HBase* é baseado em um modelo orientado a colunas;
- **ZooKeeper** - fornece um serviço centralizado para coordenação de aplicações distribuídas. Mantém as informações de configuração das aplicações distribuídas, além de fornecer a sincronização destas. Este modelo é muito utilizado em aplicações distribuídas, onde este subconjunto tem a possibilidade de fornecer uma interface simples para auxiliar o desenvolvedor;
- **Hive** - este pode ser classificado como sendo uma espécie de *Data Warehouse* distribuído, facilita a utilização de grandes conjuntos de dados em ambientes de armazenamento paralelo.

7 ESTUDO DE CASO

Neste capítulo, é apresentado um caso real da utilização dos conceitos envolvendo *Big Data*, reunindo as técnicas de *MapReduce,* para lidar com o paralelismo deste cenário, *DBScan* para cuidar da grade quantidade de informações que estarão em tramite, juntamente com o *Hadoop,* além de a infraestrutura ser baseada em computação em nuvem. Este caso de uso visa apresentar como as técnicas de mineração são utilizadas em projetos de *Big Data*. [L. C. da Silva, Ticiana - 2013]

A Universidade Federal do Ceará (UFC), possui o seguinte cenário para desenvolvimento: a cidade de Fortaleza quer desenvolver um modelo que consiga prever como está o engarrafamento no trafego na cidade em tempo real, seria um serviço similar ao que o aplicativo *waze* faz, ode este determinaria, quais são os pontos críticos e quais os favoráveis a se passar, mapeando desta forma todas as ruas existentes na cidade e os acontecimentos do momento. [L. C. da Silva, Ticiana - 2013]

Para que as informações necessárias para o desenvolvimento deste projeto sejam obtidas, foi coletado material de diversas fontes, utilizando *tweets*, foto-sensores, sites de trânsito e de GPS. Em cidades grandes é possível obter informação sobre o trânsito de diversas fontes, como os dados extraído de redes sociais, essas informações podem ser utilizadas como complemento (pois em sua maioria as mesmas estão sendo reportadas em tempo real ou quase real, o que possibilita orientar o setor público do que está acontecendo, fazendo com que seja possível se tomar decisões e medidas para os acontecimentos), para as informações provindas de câmeras e sensores físicos (que garantem uma veracidade maior). [Ribeiro Jr et al. 2012]

A parir desses dados coletados é possível identificar o comportamento do transito nos locais, possibilitado assim que sejam traçados planos para sanar os problemas que estão acontecendo [Ribeiro Jr et al. 2012]. Estes fatos comprovam que a obtenção de todas essas informações pode ser muito poderosa para que se possa adquirir o conhecimento necessário para a execução de diversas tarefas e este é o motivo do desenvolvimento do projeto da UFC, estes fatos também mostram o motivo de tantas fontes para ganho de informação. [L. C. da Silva, Ticiana - 2013]

Podemos identificar este coo sendo um projeto considerado como *Big Data* por alguns motivos:

I. Pelo grande volume de dados a ser considerado para análise e processamento;
II. Pela variedade de dados utilizados, por se trabalhar com diversas fontes, os dados serão provindos de forma desestruturada;

III. Pela velocidade que estas informações chegam, por deverem ser tratadas e utilizadas da forma mais veloz possível.

7.1 Infraestrutura

Neste estudo de caso, buscamos, a partir de dados de transito coletados, definir em que pontos estão ocorrendo congestionamento na cidade de Fortaleza. Para que este possa ser desenvolvido, será utilizado o algoritmo de *DBScan*, para o trabalho com volume dos dados, este será utilizado em sua versão paralelizada. Já para lidar com o paralelismo será utilizado o *MapReduce* com plataforma para a execução do *DBScan*. [L. C. da Silva, Ticiana - 2013]

Neste caso iremos utilizar o ambiente em nuvem, que trabalha com a plataforma *OpenNebula*. É utilizado um total de 11 máquinas virtuais, tendo o *Ubunto* como sistema operacional, 8GB de memória *RAM* e 4 unidades de CPU. [L. C. da Silva, Ticiana - 2013]

Foi necessária a instalação do *PostgreSQL* nas *VMs* para que fosse possível manipular os objetos espaciais.

7.2 Tratamento dos dados

Neste cenário proposto, estamos lidando com a velocidade do trânsito, desta forma que definimos que um engarrafamento se caracteriza pelo fato de os veículos se encontrarem em baixa ou nula velocidade, nesta primeira etapa, aplicaremos um filtro nos dados coletados para que apenas informações que possuam registro de velocidade abaixo de 20km/h, onde estes serão os dados utilizados como válidos. [L. C. da Silva, Ticiana - 2013]

Em uma segunda etapa do processo, será considerado as dimensões espacial: latitude e longitude, é nesta etapa que irá ocorrer a *clusterização* por proximidade geográfica dos pontos ode foram detectados os dados com baixa velocidade, fazendo com que seja possível visualizar onde se encontram os pontos com velocidade baixa e a partir do momento em que consideramos a dimensão espacial, será possível representar os pontos que tem potencial de engarrafamento. [L. C. da Silva, Ticiana - 2013]

Para que possamos definir nossos pontos de engarrafamento, será necessário identificar os conjuntos que possuem alta densidade nesse grupo de dados, mesmos que estes apresentem baixas velocidades. A seguir, será definido os passos necessários para a paralelização do *DBScan*, com a ajuda do modelo de programação *MapReduce*. [L. C. da Silva, Ticiana - 2013]

1. Nesta fase, será executada a primeira etapa de Map, onde cada registro do conjunto de dados será descrito como um par (de chave/valor), onde a chave será atribuída a rua referente do dado e o valor será referente a posição geográfica.
2. Na segunda fase, será executado o *Reduce*, onde o mesmo receberá uma lista de valores de chaves iguais, isso quer dizer as posições geográficas da mesma rua. E é neste momento onde o algoritmo do *DBScan* entra em ação.

   O resultado deste processo é armazenado em um banco de dados, onde as informações do *ID* de cada cluster e as informações de latitude, longitude e outras informações necessárias.
3. Como em uma cidade as ruas se cruzam e a divisão dos dados é realizada por rua, é necessário descobrir quais *clusters* (‘ruas’) diferentes se cruzam ou quais podem ser unidos. Isto quer dizer que dois *clusters* podem ser unidos, caso sua distância seja menor que o *eps* estipulado, isto quer dizer que caso os dados fossem processados em um mesmo *Reduce* ou na mesma partição, estes estariam em apenas um *clusters.* Isto faz com que o *clusters* seja armazenado como um objeto geométrico no banco de dados, este fator é importante pelo fato de que, na etapa de *merge*, apenas objetos que estão a uma distância de no máximo tantos *eps*, poderão prosseguir.
4. Esta fase pode ser denominada de merge de *clusters*, onde sua descrição também é feita através de um processo de *MapReduce*. Possuimos a função *Map* como sendo a identidade e a função *Reduce* recebendo como chave o menor ID de *clusters*, onde este deve ser reunido em um só. Caso essa união aconteça, é necessário atualizar as informações dos pontos pertencentes aos *clusters*.

Na sessão seguinte é mostrado os algoritmos, pertencentes a solução proposta, sendo implementados.

7.3 Implementando os Algoritmos

Quando trabalhamos com *MapReduce* é necessário definir duas classes que estendam de *MapReduceBase* e implementem uma das seguintes interfaces *Mapper* e *Reducer*. Os tipos de saída dessas funções são parametrizadas, onde estas, devem implementar a interface *Writable* (o *Hadoop* fornece alguns modelos básicos que seguem esta interface). [L. C. da Silva, Ticiana - 2013]

```
1  public static class  ClusterMapper extends MapReduceBase implements Mapper<LongWritable,
       Text, Text, Text>
2      {
3            //registro é formado por uma tripla <Rua, lat, long>
4            public void map(LongWritable key, Text value, OutputCollector<Text, Text>
                 output, Reporter reporter) throws IOException
5            {
6
7                Text street = new Text();
8                String line = value.toString();
9                String[] tokenizer = line.split(",");
10               street.set(tokenizer[0]);
11               String values=tokenizer[1]+" "+tokenizer[2];
12               value.set(values);
13           output.collect(street,value);
14           }
15     }
```

Figura 8: Primeiro MapReduce - Particionamento
Fonte: [L. C. da Silva, Ticiana - 2013]

O procedimento, *ClusterMapper*, executado na Figura 8 tm coo objetivo converter os dados que foram recebidos em arquivos, onde estes são formados por um conjunto triplo de valores <*nomedarua,latitude,longitude*> e em pares de chave-valor <K,V>, de tal forma que K é o nome da rua e V uma tupla com os valores de <latitude,longitude>. Estas informações são então passadas para a fase de *Reduce*. [L. C. da Silva, Ticiana - 2013]

Este processo é possível graças ao *Hadoop*, pois este por padrão particiona os arquivos que serão processados por linha, desta forma cada linha é passada uma a uma ao procedimento *Map*. [L. C. da Silva, Ticiana - 2013]

Permanecendo no primeiro *MapReduce* na fase de *Reduce*, como mostra a Figura 9, uma mesma chave recebe um conjunto de valores. O algoritmo de *DBScan* juntamente com a utilização do índice *KD-tree* [Bentley 1975] é aplicado aos pontos de uma mesma rua, sendo essa fase paralelizada para ruas distintas que pertencem ao conjunto de dados de entrada. Os resultados de cada cluster criado nessa fase são armazenados em um banco de dados, onde neste cenário se faz a utilização do *PostgreSQL*. [L. C. da Silva, Ticiana - 2013]

Na Figura 10, é apresentado a classe eu ontem o método principal, onde o objeto *JobConf* é criado. É a partir deste que se torna possível criar todas as configurações iniciais, assim como os diretórios de entrada e saída de dados e dos parâmetros de *eps* e *minPoints* necessários. É válido identificar que no algoritmo da Figura 9, existe um método para se retomar o parâmetro passado para o *JobConf* no método principal. [L. C. da Silva, Ticiana - 2013]

```
1  public static class ClusterReducer extends MapReduceBase implements Reducer<Text, Text,
       Text, Text> {
2          private static double eps;
3          private static int minPoints;
4          public void configure(JobConf job) {
5              eps = Double.parseDouble(job.get("eps"));
6              minPoints = Integer.parseInt(job.get("minPoints"));
7              super.configure(job);
8          }
9          public void reduce(Text key, Iterator<Text> values, OutputCollector<Text, Text>
               output, Reporter reporter) throws IOException {
10
11         String path = key.toString().replaceAll(" ", "");
12         BufferedWriter out = new BufferedWriter(new FileWriter(file));
13         String[] line;
14         while (values.hasNext()) {
15             line = values.next().toString().trim();
16             out.write(line[0]+","+line[1]+"\n");
17         }
18         out.close();
19         dbscan.loadPoints(file);
20         dbscan.createIndex();
21         dbscan.dbscan(key.toString(), eps, minPoints);
22         output.collect(key, new Text("points clustered"));
23        }
24   }
```

Figura 9: Primeiro MapReduce - Aplicação do DBScan nos dados de uma mesma partição
Fonte: [L. C. da Silva, Ticiana - 2013]

```
1  public class FirstMapReduceKdTree {
2        public static void main(String[] args) throws IOException {
3            JobConf conf = new JobConf(FirstMapReduceKdTree.class);
4            conf.set("eps", args[2]);
5            conf.set("minPoints", args[3]);
6            conf.setJobName("ClusteringKdTree");
7            conf.setOutputKeyClass(Text.class);
8            conf.setOutputValueClass(Text.class);
9            conf.setMapperClass(ClusterMapper.class);
10           conf.setReducerClass(ClusterReducer.class);
11           conf.setInputFormat(TextInputFormat.class);
12           conf.setOutputFormat(TextOutputFormat.class);
13           FileInputFormat.setInputPaths(conf, new Path(args[0]));
14           FileOutputFormat.setOutputPath(conf, new Path(args[1]));
15           conf.setJarByClass(FirstMapReduceKdTree.class);
16           JobClient.runJob(conf);
17     }
18 }
```

Figura 10: Classe de preparação dos dados que chama as funções *Map* e *Reduce*
Fonte: [L. C. da Silva, Ticiana - 2013]

A fase que vem a seguir é realizada através de dois (2) procedimentos que estão armazenados no banco utilizado (*PostgreSQL*). O primeiro procedimento quando chamado tem a função de criar um objeto geométrico dos clusters gerados na fase anterior, já o segundo procedimento tem como objetivo encontrar quais destes clusters deveriam ser apenas um, onde deveriam estar contidos em apenas uma partição durante a aplicação do *DBScan*. Para descobrir quais destes devem ser fundidos, é então analisado quais dos objetos estão a uma distância menor ou igual que os *eps*. [L. C. da Silva, Ticiana - 2013]

Obs.: A função utilizada para a criação dos objetos geométricos utilizada para este caso está sendo a *st-convexhull* e para a comparação da distância entre dois objetos se está utilizando o *ST-Distance*. As das pertencem a biblioteca do *PostGIS*. [L. C. da Silva, Ticiana - 2013]

Para descobrir se dois *clusters* devem se fundir, o algoritmo da Figura 11 a seguir irá receber dois dados de entrada inicial. E então é feito os cálculos para se verificar se os clusters são vizinhos e se o *merge* destes deve ser feito, sempre levando em consideração o valor estipulado dos *eps*. Caso o retorno seja positivo para a realização da fusão os pontos pertencentes a cada *cluster* serão renomeados para assim, possuírem o mesmo identificador. [L. C. da Silva, Ticiana - 2013]

```
1   public static class MergeReducer extends MapReduceBase implements Reducer<Text, Text
        , Text, Text>{
2
3       private static double eps;
4       private static int minPoints;
5       public void configure(JobConf job) {
6       eps = Double.parseDouble(job.get("eps"));
7       minPoints = Integer.parseInt(job.get("minPoints"));
8       super.configure(job);
9       }
10      public void reduce(Text key, Iterator<Text> value, OutputCollector<Text, Text>
            arg2, Reporter arg3) throws IOException {
11      String[] line;
12      String cluster;
13      ArrayList<String> clustersToMerge = new ArrayList<String>();
14      while(value.hasNext()){
15          clustersToMerge.add(value.next().toString());
16      }
17      for(int i = 0; i < clustersToMerge.size(); i++){
18          line = clustersToMerge.get(i).split(" ");
19          if(dbscan.mergeDBScan(line[0], line[1], eps, minPoints)){
20              for (int j = i+1; j < clustersToMerge.size(); j++) {
21                  if(clustersToMerge.get(j).startsWith(line[1]+" ")){
22                      cluster = clustersToMerge.get(j);
23                      cluster = cluster.replaceAll(line[1]+" ",line[0]+" ");
24                      clustersToMerge.set(j, cluster);
25                  }
26                  else if(clustersToMerge.get(j).endsWith(" "+line[1])){
27                      cluster = clustersToMerge.get(j);
28                      cluster = cluster.replaceAll(" "+line[1]," "+line[0]);
29                      clustersToMerge.set(j, cluster);
30                  }
31              }
32          }
33      }
34  }
35  }
```

Figura 11: MapReduce de Verificação de fusão de clusters
Fonte: [L. C. da Silva, Ticiana - 2013]

Nesta fase do processo de *MapReduce* possuímos a função *Map* como sendo a identidade, porém temos a função *Reduce*, que pode ser visualizada na Figura 11, é paralelizada com base no número de representantes a sofrerem merge recebidos como chave. O ato de atualizar os identificadores é realizado no algoritmo da Figura 11.

Figura 12: Clusters encontrados na fase de fusão, saída inicial
Fonte: [L. C. da Silva, Ticiana - 2013]

7.3.1 Resultados do Processo de Mineração

Após todo o processo de implementação realizado cima se pode iniciar o recebimento de dados de retorno, após a execução destes foi possível estimar que, quando a quantidade de pontos a serem processados pelo *DBScan* aumenta, o tempo de processamento também aumenta, mostrando que a solução proposta neste caso é de natureza escalável. Foi possível analisar que com este processamento realizado, as utilizações das informações de entrada foram usadas de forma concreta retornando valores de saída esperados, assim como mostrado na Figura 11.

Este retorno de saída só foi possível de ser criado, pelo uso dos métodos mostrado no decorrer deste trabalho, fazendo com que o montante de dados trabalhados e a velocidade para ao qual ele devesse ser mostrado, não se tornassem um empecilho para a realização desta aplicação, pois mesmo que seja possível a utilização de outros métodos para a realização do mesmo, estes tornam o desenvolvimento necessário muito mais enxuto e simples.

8 CONCLUSÃO

Com o avanço tecnológico e a importância dos dados se tornando cada vez mais valioso, em diversos setores da sociedade, o armazenamento e processados para sua futura utilização destes vem sendo cada vez mais requisitados pelas companhias atuantes no mercado. Estes dados ganharam proporções enormes e estes foram então denominados *Big Data*, porém o trabalho com este não é simples. Para o auxílio ao desenvolvimento de trabalhos de *Big Data*, estudiosos foram com o passar do tempo, desenvolvendo técnicas, modelos e algoritmos para que o processamento deste montante fosse simplificado. Foi apresentado, no decorrer deste trabalho, alguns modelos e técnicas utilizados em projetos, no quesito de mapeamento e análise dos dados.

Podemos, ao final deste trabalho, concluir que o conceito de *Big Data* vem a cada dia, ganhando mais espaço no meio de diversas áreas da sociedade, não sendo exclusiva apenas para a área de *TI*.

Mesmo que o *Big Data* seja um termo relativamente novo, este se torna/fixa como sendo uma boa solução para diversas dificuldades encontradas/existentes no quesito de manuseio de grande volume de dados, tendo com o foco a análise e o manuseio deste montante de dados.

No decorrer das sessões apresentadas, foi mostrado um pouco mais do termo *Big Data*, onde se pode visualizar um pouco de seus princípio e modelo para funcionamento, destacando pôr fim a parte de análise e mapeamento dos dados que circulam este. Foi destacada diversas técnicas e modelos para que este trabalho possa ser realizado, mostrando alguns dos métodos utilizados e seu poderoso uso para o desenvolvimento de aplicações.

Neste quesito de mapeamento, foi ganho certo foco na utilização do *Haddop* e das técnicas que o precedem, apresentando alguns algoritmos para a realização da mineração e as boas práticas necessárias para sua utilização. Ao final é apresentado um cenário onde seu desenvolvimento é realizado através dos conceitos de *Big Data*, mostrando ser possível a execução do trabalho onde uma grande quantidade de dados é utilizada.

É importante mencionar que, mesmo que o *Big Data* esteja ganhando espaço, este por ser considerado ‘novo’, possui ainda algumas limitações que devem ser exploradas e desenvolvidas, onde desafios como armazenamento, processamento, análise, entre outros deve ser superado com o intuito de tornar este um modelo cada vez maior e mais robusto, porém devemos manter em mente que estes desafios geram oportunidades de pesquisa neste setor, com objetivo de fazer com que a análise para *Big Data* possa ser utilizada pelas companhias, melhorando seus produtos, serviços e a qualidade de vida de todos.

REFERÊNCIAS

[Adami, Anna] Anna Adami - **Big Data** - http://www.infoescola.com/informatica/big-data/

[DEVMEDIA] DEVMEDIA - **Significado de Data Mining**
http://www.significados.com.br/data-mining/

[CSC Tecnologia] CSC Tecnologia - **O QUE É O CLOUD COMPUTING?**
http://www.csc.com/pt/offerings/63346-o_que_%C3%A9_o_cloud_computing

[Elton Meira - 2014] Elton Meira - **Big data e mineração de dados** – 2014
http://pt.slideshare.net/eltonmeira/big-data-e-minerao-de-dados-41373638

[Emerson Alecrim – 2013/2015] Emerson Alecrim - **O que é Big Data?** – 2013/2015
http://www.infowester.com/big-data.php

[Euriditionhome, 2004] Euriditionhome - **Data Mining Tutorials, Resources**.- 2004
http://datamining.eruditionhome.com

[Jornal Conexão Sebrae/MS] Jornal Conexão Sebrae/MS - **A importância do armazenamento de dados para as micro e pequenas empresas -**
http://www.perallis.com.br/news/a-importancia-do-armazenamento-de-dados-para-as-micro-e-pequenas-empresas

[MARTINS, CLÁUDIO - 2014] CLÁUDIO MARTINS – **Fundamentos sobre MapReduce**, 2014
http://www.devmedia.com.br/fundamentos-sobre-mapreduce/28644

[Marcos Vieira – 2014] Marcos Vieira - **Entendendo Big Data** – 2014
https://www.ecommercebrasil.com.br/artigos/entendendo-big-data/

[Tourion, Cesar - 2012] Cesar Tourion - **Você realmente sabe o que é Big Data?** -2012
https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/voce_realmente_sabe_o_que_e_big_data?lang=en

[Ohl, Rodolfo] Rodolfo Ohl - **Big Data: como analisar informações com qualidade** -
http://corporate.canaltech.com.br/coluna/big-data/Big-Data-como-analisar-informacoes-com-qualidade/

[Paiva, Ricardo - 2011] Ricardo Paiva - **O que é e como funciona o Map Reduce usado pelo Google** – 2011
http://blog.werneckpaiva.com.br/2011/08/como-funciona-o-map-reduce-usado-pelo-google/

[WIKIPEDIA] WIKIPEDIA – **MapReduce**
https://pt.wikipedia.org/wiki/MapReduce

[Arthur and Vassilvitskii 2007] Arthur, D. and Vassilvitskii, S. k-means++: - **The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics** – 2007

[Amorim, Thiago - 2006] Thiago Amorim - **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados** – 2006

[Carvalho, 2005] Luís Alfredo Vidal de Carvalho - **Data Mining A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração** – 2005

[Claudio José Silva Ribeiro - 2014] Claudio José Silva Ribeiro - **Big Data: os novos desafios para o profissional da informação** – 2014

[Cohen et al. 2009] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., and Welton, C. **Mad skills: new analysis practices for big data. Proceedings of the VLDB Endowment** - 2009

[Cover and Hart 1967] Cover, T. and Hart, P. **Nearest neighbor pattern classification. Information Theory, IEEE Transactions on** – 1967

[Dai and Lin 2012] Dai, B.-R. and Lin, I.-C. - **Efficient map/reduce-based dbscan algorithm with optimized data partition. In Cloud Computing (CLOUD)** – 2012

[Ester et al. 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. - **A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD** - 1996

[Fonte, Flávio - 2013] Flávio Fonte - **Big Data - O que é o hadoop, map reduce, hdfs e hive** – 2013

[HAN, J; KAMBER, M. - 2006] HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques. Elsevier**, 2006.

[Herodotou et al. 2011] Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F. B., and Babu, S. **Starfish: A self-tuning system for big data analytics. In CIDR, volume 11** – 2011

[J. Sadalage, Pramod - 2013] Pramod J. Sadalage, Martin Fowler - **NoSQL, um guia conciso para um mudo emergente da persistência poliglota** – 2013

[Jonathan Stuart Ward and Adam Barker - 2013] Jonathan Stuart Ward and Adam Barker - **Undefined By Data: A Survey of Big Data Definitions** – 2013

[L. C. da Silva, Ticiana - 2013] Ticiana L. C. da Silva, Antonio Cavalcante - **Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem** – 2013

[LAROSE, DANIEL T - 2014] LAROSE, DANIEL- **Discovering Knowledge in Data: An Introduction to Data Mining** – 2014

[MG (DINO) - 2015] Belo Horizonte, MG (DINO) - **A Evolução no Armazenamento dos Dados ao longo dos anos** – 2015

[Michie et al. 1994] Michie, D., Spiegelhalter, D. J., and Taylor, C. C**. Machine learning, neural and statistical classification** – 1994

[NAVEGA, SERGIO – 2002] NAVEGA, SERGIO - **Princípios Essenciais do Data Mining** – 2002

[Oliveira Camilo, Cássio - 2009] Cássio Oliveira Camilo - **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas** – 2009

[Oliveira Camilo, Cássio - 2009] Oliveira Camilo, Cássio | da Silva, Carlos - **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas** - 2009

[Paiva, Ricardo - 2011] Ricardo Paiva - **O que é e como funciona o Map Reduce usado pelo Google** – 2011

[Quinlan 1986] Quinlan, J. R. - **Induction of decision trees. Machine learning** - 1986

[Rafael Russo - 2013] Rafael Russo - **A História e evolução do Armazenamento Digital** - 2013

[Ribeiro Jr et al. 2012] Ribeiro Jr, S. S., Rennó, D., Gonçalves, T. S., Davis Jr, C. A., Meira Jr, W., and Pappa, G. L. **Observatório do trânsito: sistema para detecção e localização de eventos de trânsito no twitter**. SBBD - 2012

[RONALDO GOLDSCHMIDT - 2005] RONALDO GOLDSCHMIDT, EMMANUEL PASSOS - **Data mining: um guia Prático** – 2005

[Schonberger, Viktor Mayer – Big Data] Viktor Mayer Schonberger; Kenneth Cukier – **Big Data, como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana** – 2013

[Thomas H. Davenport, Dados Demais!] Thomas H. Davenport – **Dados Demais! Como desenvolver habilidades analíticas para resolver problemas complexos, reduzir riscos e decidir melhor** - 2013

[Wu et al. 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. - **Top 10 algorithms in data mining. Knowledge and Information Systems** - 2008