UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC

CURSO DE CIENCIA DA COMPUTAÇÃO

ÁLVARO AMBRIZ DOMINGOS

METODOLOGIA PARA MINERAÇÃO DE DADOS NÃO ESTRUTURADOS NO SOFTWARE TABLEAU (ESTUDO DE CASO) COM BIG DATA

CRICÍUMA

2016

ÁLVARO AMBRIZ DOMINGOS

METODOLOGIA PARA MINERAÇÃO DE DADOS NÃO ESTRUTURADOS NO SOTWARE TABLEAU (ESTUDO DE CASO) COM BIG DATA

Trabalho de Conclusão de Curso, apresentado para obtenção do Grau de Bacharel no curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientador: Prof. MSc. Paracelso de Oliveira Caldas

CRICÍUMA

2016

ÁLVARO AMBRIZ DOMINGOS

METODOLOGIA PARA MINERAÇÃO DE DADOS NÃO ESTRUTURADOS NO SOFTWARE TABLEAU (ESTUDO DE CASO) COM BIG DATA

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Banco de Dados Relacional e Distribuído

Criciúma, 01 de dezembro de 2016.

BANCA EXAMINADORA

Prof. Paracelso de Oliveira Caldas – MSc. - (UNESC) - Orientador

Prof. Gustavo Bisognin - MSc. - (UNESC)

Prof. Gilberto Vieira da Silva - Esp. - (UNESC)

Dedico este projeto primeiramente ao grandioso Deus por me ter dado força e coragem, e aos meus queridos pais que desde sempre investem na minha formação acadêmica, pelo infinito amor, e valores ensinados. Sou grato a vocês pela educação que um dia passaram-me e pelo homem que me estou a tornar.

AGRADECIMENTOS

A Deus acima de tudo quero agradecer pois, somente através da sua graça e paz foi possível chegar até aqui. Agradeço aos meus queridos pais, Francisco Gabriel e Maria de Fátima pelo apoio, amor, carinho, e por incentivarem me sempre em todos os momentos da minha vida, pelo suporte necessário para a minha formação pessoal e acadêmica. Sem dúvida nenhuma, se vocês não estivessem ao lado de mim durante todas as minhas escolhas, teria sido muito mais difícil. Muito obrigado por todo carinho, conselhos e força.

Aos meus irmãos principalmente Gabriel, Dulce e o Mauro pela força que deram-me ao longo desses anos todo longe de casa. Ao meu orientador professor Mestre Paracelso de Oliveira Caldas por todo apoio, paciência, incentivo e empenho dedicado a mim desde o início deste trabalho. Por todos os ensinamentos compartilhados durante as orientações, que com certeza agregaram muito ao desenvolvimento deste trabalho.

A minha querida namorada Agnes Julia por esta sempre ao pé de mim e por ter compreendido a minha ausência (e o meu estresse), principalmente nos momentos finais do TCC.

Aos meus amigos César, Humberto, Josemar, que estiveram muito próximos a mim nesses anos compartilhando todas as angustias e incertezas, mais também muitas risadas e momentos de companheirismo e parceira.

Sem esquecer a professora Doutora Merissandra por me ter aturado sempre que eu fosse tirar qualquer dúvida a respeito do tcc, agradeço-te muito pois foste para mim mais que uma professora.

RESUMO

O armazenamento das informações em formato digital cresceu muito nos últimos anos. Não apenas pessoas que são responsáveis por produzir dados, os equipamentos industrias, operacionais, electrónicos também se tornaram grandes geradores de registros como exemplos as redes sócias, servidores, aparelhos de smartphones, microcomputadores espalhados nos mais variados contextos entre uma afinidade de aplicações. Percebeu-se que maior parte desses dados são dados não estruturado, ou seja, dados que não podem ser analisados e armazenado pelo método de banco de dados tradicionais. O termo Big Data se refere a toda esta quantidade de dados que se encontram na ordem dos dados estruturados, semiestruturados e não estruturados. Este trabalho teve como o objetivo realizar um estudo de caso, e um processo usando o software Tableau para mineração de dados não estruturados com Big Data, extraindo as informações por intermédios de visualizações.

Palavras-chave: Banco de dados. Gestão. Big Data. Tomada de Decisões.

ABSTRACT

The storage of information in digital format has grown a lot in recent years. Not only people who are responsible for producing data, industrial, operational, and electronic equipment have also become great generators of records such as social networks, servers, smartphone devices, microcomputers scattered in the most varied contexts among an affinity of applications. It was noticed that most of this data is data not structured, and data that cannot be parsed and stored by the traditional database method. The term Big Data refers to all this amount of data that is in the order of structured, semi-structured and unstructured data. The objective of this work was to carry out a case study and a process using Tableau software for unstructured data mining with Big Data, extracting the information through visualization intermediates

Keywords: Database. Management. Big data. Decision-making.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

BD	Banco de Dados
DDL	Data Definition Language
DM	Data Mining
DML	Data Manipulation Language
HTML	Hiper Text Markup Language
IMS	Information Management System
SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language
XML	eXtended Markup Language

SUMÁRIO

1 INTRODUÇÃO

Atualmente, vivemos uma era que tem sido marcada por enormes quantidades de dados. Dados esses que têm sido gerados por instituições como empresas, universidades, hospitais, a internet entre outros.

Muitos desses dados são não estruturados. Entende-se por dados não estruturados aqueles que são provenientes de documento do formato doc, docx, pdf, XML, HTML, redes sociais e outras infinidades de formatos.

Desta forma, são extremamente relevantes as atenções e os estudos desses tipos de dados visto que eles têm uma representatividade enorme além de terem significados importantes nas decisões que envolvem a sociedade em sua totalidade.

As empresas passaram a ter dificuldades de acessar e armazenar enormes quantidades de dados não estruturados pelos métodos tradicionais, visto que dados não estruturados são dados que não podem ser analisados ou armazenados em banco de dados tradicionais.

Desta necessidade surgiu o conceito de Big Data que tem sido uma peça fundamental para tomada de decisões. Big Data refere-se a conjuntos de grandes quantidades de dados cuja a dimensão é além da capacidade de ferramentas de software de banco de dados tradicionais para capturar, armazenar, gerenciar e analisar (MANYKA, 2011).

Big Data surgiu nestes últimos anos devido à enorme quantidade de dados que está sendo gerado hoje. Mais também pode ser uma tecnologia que ajuda as organizações no seu processo decisório.

Desta forma este estudo visa criar uma metodologia para mineração de dados não estruturados usando a tecnologia do Big Data.

1.1 OBJETIVOS

O objetivo deste projeto é desenvolver uma metodologia de mineração de dados em bases de dados não estruturados, com a utilização do software Tableau em um estudo de caso de empresas que usam o Big Data e

quais informações elas pesquisam.

Os objetivos específicos deste estudo consistem:

a) Compreender e aplicar o conceito de Big Data;
b) Compreender e aplicar o conceito de dados não estruturados;
c) Aplicar as técnicas de mineração de dados e visualização de dados no software Tableau;
d) Desenvolver um estudo de caso de empresas que usam big data e informações pesquisadas para mineração de dados em base de dados não estruturados.

1.2 JUSTIFICATIVA

As tomadas de decisões antes do advento Big Data eram feitas somente usando bases de dados estruturados, desta maneira a gestão ficava muito distante da realidade, tendo em vista que a maior parte das decisões precisavam usar outras fontes, e as ferramentas convencionais não dispunham de tal suporte.

Apesar de existirem ferramentas para se obter a extração dos dados, ainda faltam ferramentas que utilizem do mesmo princípio para a extração, armazenamento e consequentemente a manipulação de dados não estruturados.

E quando se lida com enormes volumes de dados, que vem de diferentes fontes e formatos, é um grande desafio para as organizações minerar, limpar, organizar e transformar esses dados em informações relevantes.

O Big Data resolve essa questão, pois, sua arquitetura foi projetada para o uso de grandes quantidades de dados estruturados, semiestruturados, não estruturados e complexos (por exemplo, redes sociais, vídeos, imagens, arquivos de e-mails, transações comercias).

Big Data representa uma nova tecnologia que está sendo um novo método de compreensão do mundo, a partir da manipulação dos dados, e do processo decisório de negócios. E também ajuda as empresas a serem mais

eficientes.

Atualmente, as empresas recorrem a tecnologia do Big Data para tomarem decisões de informações e de analítica que permitem operarem melhor e mais rapidamente.

A tecnologia Big Data vem solucionando problemas como, por exemplo, uma análise do genoma do (DNA) de células cancerosas que é utilizada para determinar o tratamento indicado para um paciente, esta análise demora de duas a três semanas. Atualmente na era do Big Data, o processo pode ser feito em 20 minutos (RODRIGO, 2014).

De acordo com o Isaca (2013) a análise de Big Data pode impactar positivamente nos seguintes processos: desenvolvimento de produto, desenvolvimento do mercado, eficiência operacional, experiência e lealdade do cliente, previsões de demanda do mercado.

Sendo assim a metodologia irá proporcionar o uso do software Tableau sobre estes dados obtendo os resultados desejados e tirando o melhor proveito deles.

2 GESTÃO

A evolução das empresas em termos de modelos estruturais e tecnológicos tendo a gestão e a administração como novos paradigmas, têm vindo exigir uma nova postura nos estilos organizacionais e gerenciais voltados para realidade diferenciada. O grande desafio desta última década têm sido a competência e a capacidade que várias organizações enfrentam para gerenciar, inovar, administrar e se adaptarem ao mais alto nível.

A gestão surgiu após a revolução industrial, os profissionais da área decidiram buscar soluções para os problemas que não existiam antes, usavam métodos de ciências, para administrar os projetos da época o que deu início a administração. Ela também pode ser definida como um dos ramos da ciência humana por tratar com grupos de pessoas procurando assim manter suas relações, a estrutura das organizações e os recursos existentes (DRUCKER,2002).

A palavra gestão remete a ideia de gerenciar, administrar, planejar, e está ligada a administração, ou colocar ideias e gerenciar ações focando os seus objetivos. Entretanto sua atenção surgiu quando alguns economistas se dedicaram a estudar o aumento da produtividade através do trabalho, nos dias de hoje é uma ferramenta muito importante na tomada de decisões empresarias. Fayol foi um dos primeiros a definir funções básicas da gestão: planejar, organizar, coordenar, controlar. Em seus estudos e pesquisas sempre destacou que ela é fundamental numa organização. Sempre inicia com uma estratégia de coordenação para um plano, contudo o gestor deve definir o seu objetivo para o colocar os planos da empresa em prática controlando, mantendo e definindo (FAYOL, 2002).

Segundo Drucker (2002) gestão é o conjunto de processos para obter resultados ou valores da empresa que direcionam as decisões dos gestores, cujo impacto se verificará no patrimônio da organização.

Gestão é a maneira mais eficaz de assumir o controle de uma determinada situação dentro de uma organização, podendo assim utilizar os

recursos existentes com eficiência possibilidade de tomada de decisões condicionadas, assim, gestão é planejamento, é organização é controle. Planeamento é o processo que vai determinar o que deve ser feito e como deve ser feito, é mais do que uma previsão. Organização é o processo que estabelece relações formais entre os clientes e a empresa para atingir os objetivos propostos. Controle é o processo de comparação do desempenho atual da organização, verificação e análise das perdas e razões que acontecem na sua origem, pretendendo definir ações necessárias para corrigir e evitar perdas futuras (TAYLOR, 2011).

Paladini (2009) também contribui quando diz que ela existe para auxiliar as organizações a atingir os seus objetivos, que seriam difíceis ou mesmo impossíveis de se atingir se as organizações agissem de formas isoladas. As empresas para atingirem os seus objetivos formulam e implementam as suas estratégias e planos para alcançarem as suas metas. Basicamente a missão da gestão é fazer com que as organizações consigam atingir os seus objetivos, pois sem ela a empresa não consegue definir estratégia de crescimento.

Além da gestão ser uma técnica de administrar, gerenciar e controlar usa-se ela em vários ramos como a contabilidade, psicologia, matemática, direito, economia estatística e na informática.

A gestão divide-se em: gestão institucional, e gestão operacional.

2.1 GESTÃO INSTITUCIONAL

Dentro de uma organização cada elemento tem o seu cargo, e um conjunto de tarefas que por ele serão executadas. Os gerentes são membros da organização responsáveis pelo bom desempenho da organização (DRUCKER, 2002).

Gerentes têm a responsabilidade e a autoridade formal para usar os recursos da empresa e tomar decisões. Abaixo serão descritos os tipos de níveis de gestão (HENRY, 2011).

O nível de gestão institucional caracteriza-se por ser um dos meios mais fortes para a estratégia, ou seja, o envolvimento da totalidade dos recursos

disponíveis na determinação e formulação de políticas gerais. O modelo da gestão institucional compreende o sistema que prevaleceu até a década de 1930, que é ligado à hierarquia, sem autonomia e o envolvimento criativo do trabalho (MATTOS, 2005).

Ela também pode ser chamada de gestão financeira estratégica. As etapas da gestão institucional geralmente envolvem a identificação do momento atual, identificando a situação desejada e determinando os passos precisos para alcançar o que se deseja (VAZ, 2013).

Na gestão institucional o gerente lida com os regulamentos e leis existentes. As leis e regulamentos são usados para desenhar um projeto bem-sucedido. Gestão institucional é o processo de colocar fora os planos da empresa e organizar recursos de produção disponíveis para ajudar uma empresa ou organização a ter sucesso (HENRY, 2011).

Durante o processo realizado ao longo da execução da gestão institucional, o gerente institucional pode avaliar vários aspectos da empresa, como recursos financeiros e humanos da empresa, e bem como as condições de mercado que eventualmente podem vir a afetar o bom desempenho da mesma (TAYLOR, 2011).

Drucker afirma que é preciso conhecer as condições do mercado que uma organização opera. Isto é fundamental na formação de um plano estratégico bem analisado e pesquisado (DRUCKER, 2002).

As análises feitas das condições do mercado determinam posição da empresa em comparação com outras.

Esta análise em gestão institucional também envolve as oportunidades disponíveis para a lucratividade no mercado e as ameaças que serão apresentadas pelos seus concorrentes. E também ajuda a formular estratégias que podem incluir normas que restringem e beneficiem os lucros das empresas no mercado inserido (VAZ, 2013).

Dentro de uma sociedade a gestão institucional existe para poder ajudar a analisar operações e também determinar o que foi feito com sucesso ou não. As informações sobre os sucessos e fracassos empresarias passados podem

ajudar a empresa a elaborar planos de gestão alcançando assim sucessos (DRUCKER, 2009).

2.2 GESTÃO OPERACIONAL

O nível de gestão operacional consiste em atividades e decisões de responsabilidade da gestão dos recursos que são dedicados à produção e o fornecimento dos produtos e serviços (SCHOOL, 2009).

A gestão operacional refere-se principalmente ao planejamento decisório da organização, e a supervisão nos contextos de produção, fabricação ou a prestação de serviços (DRUCKER, 2002).

Ela usa técnicas para criar rotinas e procedimentos dentro da empresa, e decidir quais objetivos e quais caminhos seguir (HENRY, 2011).

Essas técnicas e estratégias aplicadas ajudam a empresa a organizar os recursos humanos, financeiros e matérias e acima de tudo a tomar decisões, visto que a tomada de decisões é a verdadeira essência da gestão (MCAFEE, 2012).

Focando nas entregas, garante assim que uma organização transforma com sucesso as entradas e saídas de uma maneira mais eficiente. Essas mesmas entradas podem ser representadas como qualquer coisa: de matérias, equipamentos, e tecnologias para os recursos humanos (PEREIRA, 2014).

Gestão operacional tem fundamentos em várias áreas, como por exemplo entender as tendências mundiais na gestão do fornecimento, afim de suprir a demanda do cliente que muitas vezes são críticas e exigentes (TAURION, 2011).

O nível de gestão operacional, bem como as relações custo-eficácia, tornou-se cada vez mais importante numa época em que os recursos podem ser expectativas de oferta e de clientes (TAURION, 2011).

E com a recente explosão dos dados o nível de gestão operacional começou a ser uma peça indispensável nas empresas, por causa do grande volume de dados que o mundo vem gerando (EVANS, 2013).

O crescimento explosivo dos dados faz com que a maioria das empresas gastem com infraestrutura para mineração de dados, visto que os recursos

existentes não são suficientes e apresentam altos custos sendo acessíveis apenas as médias e grandes empresas (SIDLER, 2014).

A capacidade das análises dos gestores é fundamental na hora de se obter um bom desempenho nas resoluções dos problemas, porque de nada adianta aplicar os dados caso não haja processos bem definidos, e se a empresa não sabe o que se pretender fazer com os dados (SIDLER, 2014).

A área operacional da empresa precisa saber lidar com os dados para melhor tomada de decisões. Neste contexto o Big Data contribui efetivamente pois não tem problemas com a quantidade de dados, mas proporciona condições para saber o que fazer com eles (MCAFEE,2012).

Visto que não se tem como gerenciar o que não pode se medir, os gestores recorrem ao uso da gestão operacional com o auxílio do Big Data e outras ferramentas podendo medir, e saber radicalmente mais sobre os negócios decisórios da empresa.

O objetivo do gerenciamento de grandes quantidades de dados nas empresas, é de assegurar um nível elevado de qualidade e acessibilidade dos dados para a inteligência de negócios e grandes aplicações de dados.Com o uso do Big Data nas empresas o gestor tem facilidade de medir, gerir de forma inteligente previsões e decisões (MCAFEE,2012).

3 BANCO DE DADOS

Antes do surgimento de banco de dados as organizações armazenavam os seus dados em fichas de papel, que posteriormente eram guardados, organizados em arquivos físicos através de pastas. Extrair essas informações como manter esses arquivos organizados era uma tarefa difícil. Além disso, o acesso a informações dependia da localização geográfica dos arquivos. Estes arquivos físicos evoluíram para arquivos digitais. Foi assim que na década de 1960 e 1970, surgiram os primeiros conceitos fundamentais de bancos de dados relacionas na empresa IBM, através de pesquisas feitas descobriram que estava muito caro empregar número de pessoas para fazer trabalhos de como armazenar e indexar (MACHADO, 2008).

Os sistemas de banco de dados surgiram como opção aos métodos antigos no gerenciamento de dados em computadores. Visto que nos sistemas antigos os dados eram armazenados em arquivos do sistema operacional, que era responsável em criar ambiente para o acesso, manipulação e controle dos arquivos. Para cada interação do sistema operacional entre o sistema e o arquivo era necessário que um programa específico do sistema operacional fosse executado que fornecia aquela função (SILBERSCHATZ, 2006).

A expressão banco de dados é um termo que originou do inglês Data Banks, que eventualmente foi trocado pela palavra Data bases- Bases de dados por possuir significado, mas apropriado (SILVA, 2005).

Um banco de dados é um sistema de armazenamento de dados, que faz interação com os usuários podendo permitir diversas execuções como: inserir registros, consultar, excluir, editar, atualizar e várias outras. Ou seja, é um sistema de coleção de dados organizados, integrados, que formam uma representação de dados que pode ser utilizada por todas as aplicações relevantes sem duplicação de dados. Banco de dados pode ser utilizado tanto para o uso pessoal, como para o uso organizacional (empresas, escolas, hospitais), um exemplo de uso do banco de dados numa empresa é onde os dados armazenados são dados de trabalhadores e funções que eles exercem,

numa universidade dados armazenados de alunos e cursos, e muitas outras áreas de atuação (DATE, 2004).

Banco de dados é nada mais do que uma coleção de dados padronizados, organizados guardados em um sistema computadorizado, de modo a permitir as buscas e consultas destes mesmos dados. Portanto eles estão cada vez mais presentes em nosso dia a dia, visto que a maioria das atividades que realizamos envolvem, direta ou indiretamente, o uso de uma base de dados (SILBERSCHATZ,2006).

Banco de dados são utilizados em diversas áreas da sociedade, os bancos de dados são parte essencial do controle do negócio de diversas organizações, motivo pelo qual as empresas fornecedoras de sistemas de banco de dados estão entre as maiores empresas do mundo, exemplo da Oracle, que têm os bancos de dados Mysql e Oracle, e fazem parte da lista de produtos das empresas, mas diversificadas, como o caso da Microsoft e a IBM (SILBERSCHATZ, 2006).

3.1 SISTEMA DE GERENCIAMENTO DE BANCO DE DADOS

Para se realizar operações em um banco de dados é necessário utilizar um sistema para manipulação e gerenciamento por ele, e este sistema é conhecido como Sistema de gerenciamento de Bancos de Dados (SGBD). SGBD tem como finalidade proporcionar um ambiente tanto conveniente quanto eficiente para a recuperação e o armazenamento das informações do banco de dados (DATE,2004).

O primeiro Sistema de Gerenciador de Banco de Dados comercial surgiu no final de 1960 com base nos primeiros sistemas de arquivos que existiam na época. Os SGBDs evoluíram desses sistemas de arquivos onde o armazenamento era em disco. Alguns bancos de dados possuem alternativas de armazena-los em arquivos e que seja escrito um programa para sua manipulação, os SGBD´s oferecem uma gama de vantagens quando comparando com as coleções de arquivos (SILBERSCHATZ, 2006).

Estas são algumas das vantagens do Sistema de gerenciamento de Bancos de Dados:

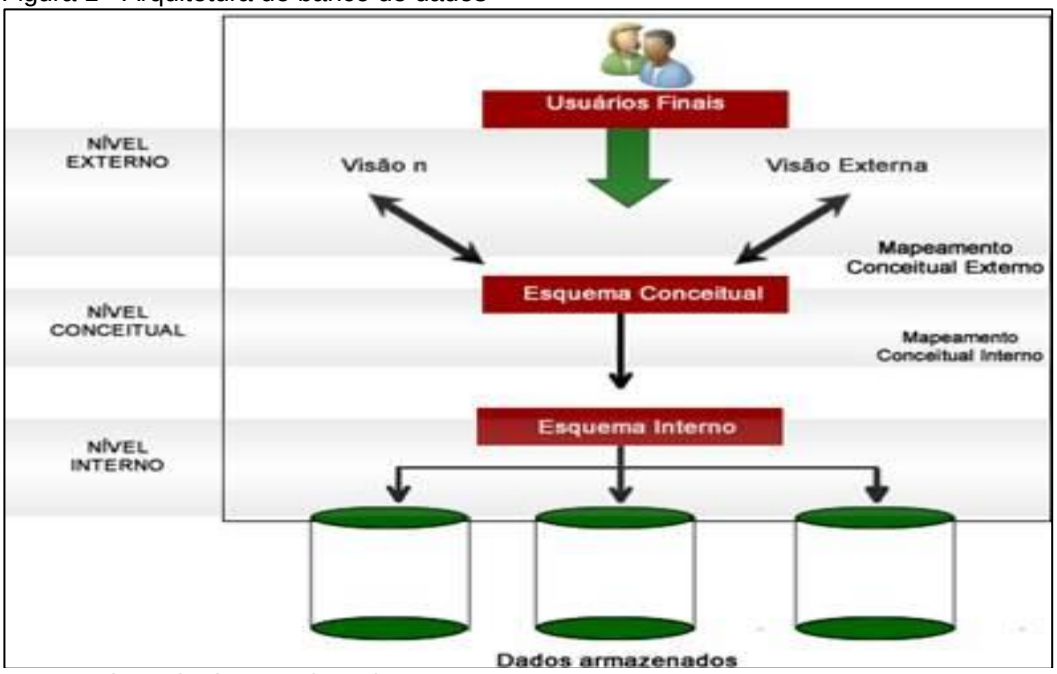
a) Independência de dados: em uma aplicação os dados são apresentados de forma tratada, ocultando detalhes que só são visíveis através de SGBD;
b) Acesso eficiente aos dados: um SGBD é desenvolvido especificamente para acessar e manipular o banco de dados, desta forma ele possui técnicas específicas e eficazes para visualização e manipulação de dados;
c) Integridade e segurança dos dados: através do SGBD é possível configurar níveis de acesso, determinado quais dados são visíveis e quais podem ser manipulados para cada nível de usuário, além de respeitar as restrições do banco ao executar uma operação;
d) Administração de dados: o SGBD oferece uma administração centralizada dos dados, desta forma os administradores, que conhecem a estrutura de banco e os dados armazenados de forma geral, diferente de usuários do sistema de aplicação, que tem geralmente conhecimento apenas dos dados que está habituado a utilizar, possuem mais facilidade para organizar a representação dos dados, realizar melhorias na estrutura, garantindo a eficiência do banco;
e) Acesso concorrente e recuperação de falhas: o SGBD garantirá que a transação executada por um usuário não afete as transações de outro usuário que esteja acessando o banco ao mesmo tempo e garante a recuperação dos dados em caso de falha;
f) Tempo reduzido de desenvolvimento no aplicativo: o SGBD, por ser um sistema específico para a manipulação do banco de dados, possui uma serie de rotinas e funções para a manipulação de dados que são comuns a diversos aplicativos que acessam o banco (RAMAKRISHNAN,2008).

3.2 ARQUITETURA E MODELAGEM DE BANCO DE DADOS

A arquitetura de sistemas de banco de dados conhecido como arquitetura ANSI/SPARC, é uma arquitetura para sistemas de banco de dados que têm como o objetivo viabilizar as características da abordagem de banco de dados, sobretudo na independência de dados (DATE, 2004).

Ela se divide em três níveis, conhecidos como: nível externo, nível conceitual, e o nível interno.

Figura 1 - Arquitetura de banco de dados



Fonte: Adaptado de Date (2004).

A figura 1 mostra como é constituída a arquitetura de banco de dados, cada nível tem uma função importante na estrutura de um banco de dados, ou seja, existe vantagens e desvantagens em usar cada um desses níveis na arquitetura do banco de dados.

3.2.1 Nível externo

O nível externo é o nível do usuário individual, ou seja, um determinado usuário que pode ser um programador de aplicações como um usuário final de qualquer grau. Em geral o usuário está interessado em apenas uma parte do banco de dados, é fornecido ao usuário o acesso apenas a essa parte dos dados armazenados. E essa parte de dados visualizada por determinado usuário chama-se visão externa. Visões externas são definidas através de esquemas externos. O nível externo trata desse esquema (DATE, 2004).

Segundo Silberschatz (2006) O nível externo descreve um modelo de dados que demostra a estrutura de armazenamento físico do banco de dados, e o caminho do acesso aos dados guardados, ou seja, consiste em uma série de diferentes pontos de vista externo do banco de dados, descreve parte do banco de dados para determinados grupos de usuários. Fornece um forte e flexível mecanismo de segurança ocultando partes do banco de dados a partir de determinados usuário, ele permite aos usuários acessar dados de uma forma personalizada para as suas necessidades, de modo que os mesmos dados podem ser vistos por diferentes utilizadores de diferentes maneiras ao mesmo tempo.

3.2.2 Nível conceitual

Nível conceitual trata da representação de toda informação armazenada do banco de dados, ou seja, uma visão externa de forma abstrata em comparação da maneira como os dados são armazenados fisicamente. Ela também é bastante diferente do modo como os dados são visualizados por qualquer usuário, a visão conceitual pretende ser uma visão dos dados mostrando realmente como esses dados são (DATE, 2004).

O nível conceitual possui um esquema que descreve a estrutura de todo banco de dados para os usuários, esse esquema conceitual oculta os detalhes da estrutura do armazenamento físico e concentra-se na descrição de entidades, tipos de dados, operações e restrições (ELMASRI; NAVATHE, 2011).

A visão conceitual mostra a estrutura lógica do banco de dados inteiro como é visto pelo administrador, mostra a visão dos dados armazenados no banco de dados, a relação entre os dados, visão completa dos requisitos de dados organizados, independentemente de qualquer forma de armazenamento, e as restrições sobre os dados (SILBERSCHATZ, 2006).

3.2.3 Nível interno

O nível interno é uma representação de baixo nível do banco de dados, consiste em ocorrências múltiplas de tipos múltiplos de registros internos. O registro interno é um termo ANSI/SPARC para a estrutura de registros armazenados, nível interno é descrito por meio de um esquema interno, que não define apenas os vários tipos de registros que neles estão armazenados como também especifica os índices que existem, e como os campos armazenados são representados. Nível interno é o mais próximo do armazenamento físico, por estar relacionado com a forma como os dados são armazenados, emprega-se o modelo de dados físico para descrever os detalhes de armazenamento (DATE, 2004).

Nível interno consiste também na relação da parte física do banco de dados no computador, como os dados são armazenados na base de dados, sua implementação física do banco de dados para poder alcançar um bom desempenho de tempo e utilização de espaço de armazenamento (SILBERSCHATZ, 2006).

Nível interno têm um esquema interno, que descreve a estrutura de armazenamento físico do banco de dados. E esse esquema utiliza um modelo de dados físico e descreve detalhes completos do armazenamento de dados e o caminho de acesso ao banco de dados (NAVATHE, 2005).

3.2.4 Modelos de dados

Modelo de dados é um conjunto de conceitos que usados para descrever a estrutura de um banco de dados, ou seja, refere-se à descrição

dos tipos de informações, dados que estão armazenados em um banco de dados (SILBERSCHATZ, 2006).

Segundo Date (2004) modelos de dados podem ser definido como uma coleção de construtores de alto nível para a descrição dos dados, onde este modelo oculto diversos detalhes de baixo nível.

A modelagem de dados tem como objetivo possibilitar a apresentação de uma visão única, não redundante e resumida dos dados de uma aplicação (ABREU,2002).

3.2.5 Modelos de dados hierárquico

Banco de dados hierárquicos foram os primeiros modelos de dados da IBM, conhecido como Information Management System (IMS), que foi lançado em 1960. O primeiro SGBD´s disponível comercialmente e que se tornou popular foi o hierárquico, modelos de dados hierárquicos foram os modelos que existiam antes do modelo relacional. Ele foi desenvolvido para superar as limitações do sistema de ficheiros Modelos de dados hierárquicos foram utilizados para o processamento de transações em que os volumes de transações eram elevados e tinham uma estrutura lógica básica e era representado por uma árvore de cabeça para baixo (GOMES, 2007).

Um sistema de banco de dados hierárquico é um conjunto de registros que possuem ligações entre si.

Onde o registro é uma coleção de atributos que possui apenas o valor dos dados e vínculos que representa a ligação entre dois registros (SILBERSCHATZ, 2006).

Segundo Date (2004) modelos de dados hierárquicos consistem em umas coleções de registros conectados uns aos outros por meio de ligações.

O modelo de dados hierárquicos é tanto similar ao modelo de redes no sentido de que os dados e relacionamentos são representados por ligações de registros.

Neste modelo os dados são estruturados em hierarquias ou árvores, onde os nós da hierarquia contêm ocorrências dos registros, e cada registro é

uma coleção de atributos, o registro da hierarquia precede a outros é o registro pai, e os outros registros são conhecidos como registros filhos.

3.2.6 Modelo Relacional

O modelo relacional é um dos modelos de dados atual, foi criado por Codd (1970) tornou-se um padrão de fato para aplicações comercias, devido a sua forma simples e desempenho. Este modelo consiste basicamente na relação de colunas e linhas, de uma tabela, sendo as colunas conhecidas como atributos e cada linha como uma entidade, uma relação de um conjunto de valores. A tabela é a coleção dos conjuntos de atributos e entidades (SILBERSCHATZ, 2006).

Modelo relacional de dados descreve da maneira informal o banco através de tabelas que respeitam restrições de integridade e possuem recursos para a manipulação das tabelas.

A finalidade do modelo relacional é representar os dados de uma forma simples, através de um conjunto de tabelas inter-relacionadas.

Este modelo faz com que o banco de dados se torna mais flexível, tanto na forma de representar as relações entre os dados como na tarefa da modificação (DATE, 2004).

Modelo relacional (MR) caracteriza-se em conceitos; de matemática de teorias dos conjuntos; e lógica de predicado.

Os primeiros sistemas comercias baseados em MR foram disponibilizados em 198, e vem sendo implementados em vários sistemas como Oracle, Acess, MySql e muitos outros sistemas.

Atualmente é um dos modelos de dados, mas utilizados em sistemas de gerenciadores de banco de dados comercias, visto que o modelo oferece uma gama de vantagens:

a) Suas estruturas de dados são simples, constituídas de relações e tabelas bidimensionais cujos elementos são itens dos dados.

b) Oferece uma sólida base para a consistência dos dados, permitindo a manipulação de relações orientadas a conjuntos através de linguagens não procedurais, baseadas na teoria dos conjuntos.

c) Possui mecanismos que permitem a definição de acessos aos dados, segurança, controle de redundâncias, controle de integridade, recuperação a falhas e controle de transações (ELMASRI, 2011);

3.2.7 Modelo de dados orientados a objetos

Modelo de dados orientados a objetos surgiu em meados de 1980 para armazenamento de dados complexos e não adequados aos sistemas relacionais. O modelo de banco de dados orientados a objetos é baseado nos conceitos de orientação a objetos difundidos em linguagem de programação, o seu objetivo principal é tratar os tipos de dados complexos como um tipo abstrato que pode ser um objeto (DATE, 2004).

O modelo de dados orientados a objetos é um modelo que estende as funcionalidades de um banco no modelo relacional, adicionando encapsulamento, métodos e identidade de objeto. São sistemas de banco de dados que trabalham de forma direta com linguagens orientadas a objetos usando o sistema nativo da linguagem (SILBERSCHATZ, 2006).

Um banco de dados orientados a objetos representa a forma de integrar diretamente as linguagens de programação complexas com persistência, e características dos dados. Umas das propriedades, mais importantes do modelo orientados a objeto é a sua identidade, visto que sem ela não é possível saber com qual objeto se comunicar, seja para obter ou atualizar dados. Neste modelo, tantos as informações quanto as ações que atuam sobre essas operações estão encapsuladas em uma entidade única, ou seja, os dados e comportamentos dos objetos são modelados e armazenados no mesmo sistema gerenciador de banco de dados (HOROWITZ, 2002).

O comportamento de um objeto em um banco de dados orientados a objetos é representado por meio de métodos, que consistem em uma assinatura única dentro do objeto. Os sistemas gerenciadores de bancos de

dados orientados a objetos não possuem grandes desigualdade em termos de funcionalidades exigidas quando relacionados aos SGBD´s relacionais. A principal exigência do sistema gerenciador de banco de dados orientados a objetos é o armazenamento de objetos, suas operações de forma a prover uma integração transparente com a aplicação, sem a necessidade de uma camada de tradução dos dados (GUPTA, 2005).

Entretanto, Weildler (2004) diz que apesar de todas as ferramentas que existem disponíveis no mercado, a utilização do modelo de banco de dados orientados a objetos permanece bastante pequena, quando comparado ao modelo relacional. Para Weidler algumas das razões que as organizações adotam o uso do banco de dados relacionais está ligada à cultura empresarial, visto que o tempo necessário e custos altos hesitam emigrar para o modelo de banco de dados orientados a objetos.

4 DADOS

Dados são uma sequência de símbolos quantitativos ou qualitativos. Quantitativos são dados que se refere a um número, e dados qualitativos refere-se à qualidade de alguma coisa, uma descrição de cores, textura etc. Dados podem ser umas entidades matemáticas ou uns símbolos sintáticos que podem ser descritos através de representações formais e estruturais, eles podem ser textos, letras símbolos quantificados, assim como imagens e sons, é uma coleção de fatos, como: números, palavras, medições, ou mesmo apenas descrições de coisas (SEQUEIRA,2008).

Dados são uns conjuntos de informações, relativamente aos computadores dados são informações convertidas em formato digital, binário.

Os dados são os itens mais fundamentais dentro da informação. É um termo que serve para identificar meros pedaços distintos da informação digital.

Dados podem ser a representação de alguma coisa de maneira especifica e que podem existir uma variada de formas como; números, textos.
E quando são utilizados no contexto de meros de transmissão, os dados referem-se à informação em formatos acima referido. Como exemplos nas telecomunicações, os dados muitas vezes referem-se ao formato digital em vez de analógico (ABREU,2005).

No contexto da computação, dados são símbolos ou sinais de entrada, armazenados, e processados por computador, e posteriormente geram a informação como saída. Em banco de dados Elmasri (2011) afirma que dados são fatos conhecidos que podem ser registrados e possuem significado implícito, exemplo o número de matricula de um aluno é um dado, pois para a instituição ele é um fato conhecido, contém uma semântica, é representado por símbolos sendo assim passível de registro.

Esses mesmos dados são encontrados em diversos formatos variados, dependendo do seu estado, suas características, descrição e o seu formato na web. E eles classificam-se por dados estruturados, dados semiestruturados e dados não estruturados.

4.1 DADOS ESTRUTURADOS

Os dados armazenados em banco de dados tradicionais, como exemplo do banco de dados relacionais, banco de dados orientados a objetos são dados estruturados, ou seja, são sempre especificados método que nenhum dado pode ser armazenado no banco de dados se não tiver de acordo com a descrição definida pelo esquema do banco de dados. E este armazenamento é adequado para as aplicações tradicionais onde um esquema do banco de dados são previamente definidas, os dados são regulares, organizados e adaptam-se ao método criado (DATE, 2006).

Dados estruturados são conjuntos de dados ou informações com um alto grau de organização, são dados que geralmente são organizados em uma estrutura de modo que seja identificável, de tal forma que a sua inclusão em banco de dados relacional seja transparente e que pode ser pesquisado ou mesmo consultado facilmente (SILBERCHATZ, 2006).

Os dados estruturados também podem ser definidos como dados que são mantidos em campo fixo dentro do registro ou arquivos. São dados contidos em banco de dados relacionais, que estão organizados em blocos semânticos entidades similares, e são agrupados com a entidade do mesmo grupo.

Possuem as mesmas descrições para todas as entidades de um grupo (RAMAKRISHNAN, 2008).

Normalmente os dados estruturados são armazenados em esquemas definidos como banco de dados. E geralmente é tabular, com linhas e colunas que definem claramente os seus atributos (DATE, 2004).

Dados estruturados são fácil de organizar, analisar com os seus parâmetros definidos. Sua organização possibilita o acesso rápido as informações em relação aos dados não estruturados.

4.2 DADOS SEMIESTRUTURADOS

Os dados semiestruturados são conjuntos de dados com representações sem esquemas, são dados que não possuem descrição, dados que não estão

organizados de forma complexa em um armazenamento especializado (ABITEBOUL, 2008).

Dados semiestruturados são tipo de dados que não são armazenados em banco de dados convencionais. São dados que geralmente não estão organizados em modelos racionais, como exemplo de tabela ou um gráfico baseado em um objeto (DEUTSCH, 2009)

Peter definiu dados semiestruturados como informações que não são armazenados nos bancos de dados relacionais, mas que possuem propriedades organizacionais que tornam fáceis de serem analisadas (PETER,2011).

Dados semiestruturados são cruzamentos entre os dados estruturados e os dados não estruturados. São tipos de dados que não tem a estrutura do modelo de dados rigidamente organizados (MELLO, 2007).

A estrutura dos dados semiestruturados muda-se frequentemente quanto aos seus valores. Existe uma estrutura básica para os dados, porém a mesma estrutura está implícita na forma em que os dados são apresentados.

É preciso realizar métodos para se conseguir a estrutura. Existem algumas aplicações, onde os dados possuem esquemas predefinidos e rígidos um exemplo são dados científicos e Extensible Markup Language (XML) (BUNEMAN, 2004).

Com os dados semiestruturados, marcas e vários outros tipos de marcadores são usados para identificar determinados elementos dentro dos dados.

As consultas sobre os dados semiestruturados não são tanto eficientes como em dados estruturados. Mais pode se executar consultas em dados semiestruturados criando métodos específicos para esses tipos de dados.

Os dados semiestruturados apresentam algumas características principais:

a) Os dados semiestruturados não tem um esquema definido;
b) A estrutura é irregular;
c) A estrutura está encaixada nos dados;
d) Possui uma estrutura extensa;

e) A estrutura é fortemente evolutiva e descritiva.

A gestão dos dados semiestruturados exige reconsiderar concepção dos componentes do sistema gerenciador de banco de dados. Para tal surge dadosXML uma nova forma de tratar dados semiestruturados.

Sendo que XML é muitas vezes utilizada para gerenciar dados semiestruturados (DEUTSCH, 2001).

XML surgiu como um novo padrão para as trocas dos dados existentes na Web. Do ponto de vista de bancos de dados, não apenas o uso do XML como o formato para a troca dos dados mais também como a representação interna dos dados garantindo assim um desempenho no suporte a execução eficiente em consultas sobre grandes quantidades de dados na Web (IBM,2011).

4.3 DADOS NÃO ESTRUTURADOS

Nos últimos anos vem crescendo o aumento no volume dos dados digitais, criados e armazenados quer seja por pessoas particulares, por empresas que são a maior fonte de dados. A história mostra-nos que as empresas sempre utilizaram os sistemas de computação e os bancos de dados para armazenar maior partes dos dados corporativos em formatos estruturado, como tabelas relacionais, arquivos de formato fixo.

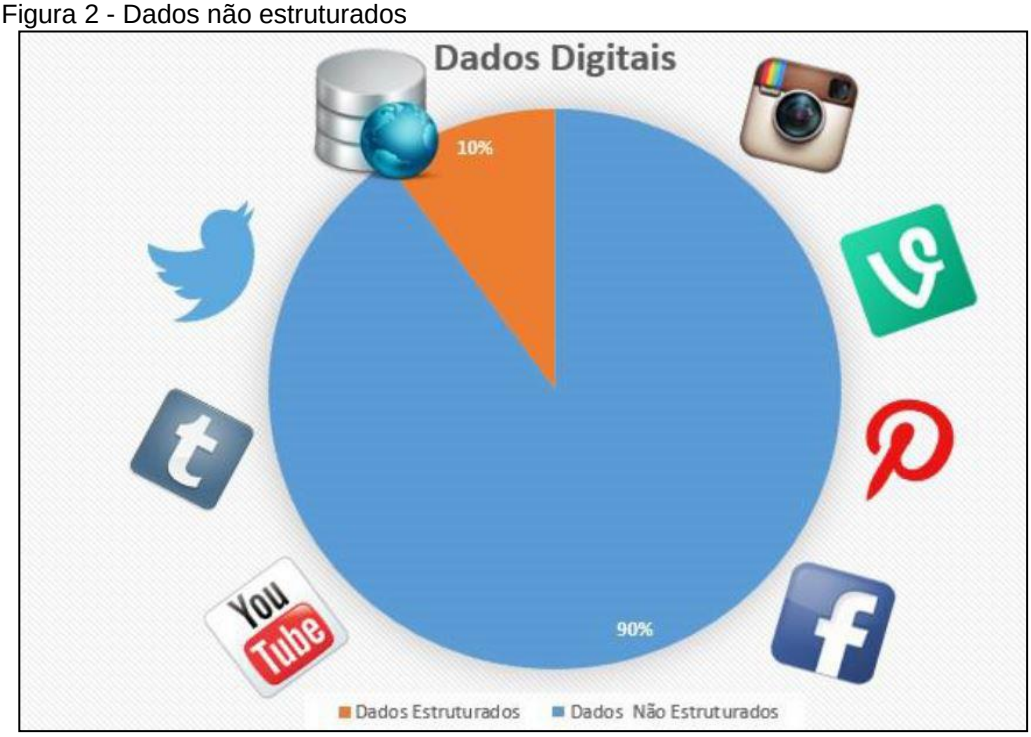
Porém, atualmente grandes partes dos dados são armazenados em documentos criados por ferramentas como: Office Excel, Microsoft Word. E com o avanço da digitalização de documentos, produção de vídeos, formatos de áudio, o conjunto de dados não estruturados cresceu (MICROSOFT, 2011).

De acordo com a pesquisa feita a estimativas, de 90% dos dados armazenados são mantidos fora do banco de dados relacionais (GUIA DAMA, 2012).

Da maioria dos dados gerados nos últimos anos, apenas 10% destes mesmo dados são dados estruturados. Os outros 90% são dados não estruturados e geralmente se encontram na sua grande parte nas redes sócias como exemplo do Twitter, Instagram, Facebook e várias outras (IBM, 2011).

A figura abaixo mostra detalhadamente o percentual dos dados não estruturados em relação a outros formatos de dados.

Figura 2 - Dados não estruturados



Fonte: Taurion (2015) ADPTADO.

Dados não estruturados são bancos formados por conjuntos de dados textuais como, por exemplo, e-mails, documentos, relatórios, registros médicos, planilhas, são coleções de dados que não tem um formato ou uma estrutura organizada. Refere-se a toda informação que não são tratados em banco de dados tradicional linha-coluna (NONOHAY, 2012).

Segundo Navathe (2005), dados não estruturados são formas para descrever os dados que não estão contidos numa base de dados ou qualquer um outro tipo de estrutura de dados. E geralmente os dados não estruturados podem ser textuais ou não textuais, textuais como e-mails, apresentações como o Power Point, documentos de Word, e dados não estruturados não textuais como imagens JPEG, arquivos de áudio MP3 e arquivos de vídeos em flash, são dados que podem ter qualquer tipo de formato ou sequência, que não seguem nenhuma regra e não são previsíveis.

Os dados não estruturados são conjuntos de dados que seguem uma forma não ordenada de itens como páginas, planilhas, tabela de banco de dados ou outros conjuntos de dados lineares.

O termo conjunto de dados não estruturados está associado aos dados que geralmente não inclui um modelo de dados pré-definido, e que não combina com tabelas relacionais, e geralmente são pesados textos que podem incluir números, datas, fotos, e que se tornam difíceis de serem identificados utilizando bancos de dados convencionais (BETTENBURG, 2014).

Um dos maiores desafios com os dados não estruturados é que eles sejam contextualizados para serem entendidos e utilizados. Alguns pesquisadores afirmam que pesquisar e analisar dados não estruturados são como procurar uma agulha num palheiro (BACCHELLI, 2013).

Com o volume dos dados não estruturado aumentando tão rapidamente, muitas empresas recorrem as soluções tecnológicas para ajudá-los a gerenciar, melhorar e armazená-los.

As empresas acreditam que seus armazenamentos de dados não estruturados envolvem informações que podem ajudá-los a tomar melhores decisões dos seus negócios e projetos (GUERROUJ, 2012).

Para o auxílio do problema, as empresas voltam-se para diferentes soluções de software projetados para procurar dados não estruturados e também extrair informações importantes. E a principal vantagem do uso dessas ferramentas é a capacidade de recolher informações importantes que podem ajudar a empresa a ter sucesso (IBM, 2011).

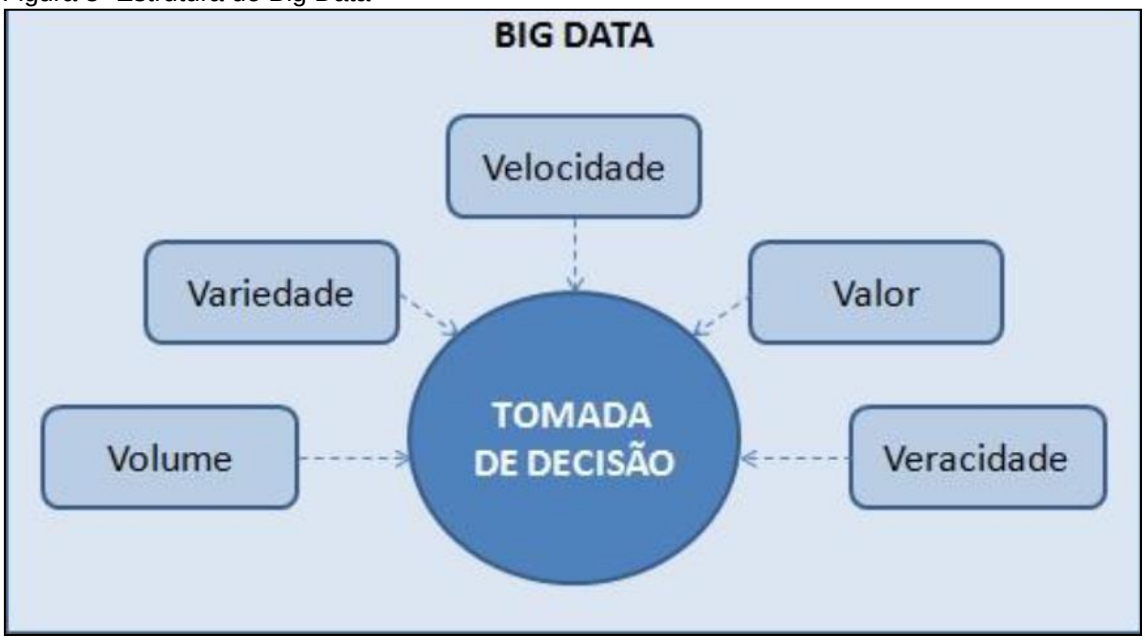
5 BIG DATA

A quantidade de informações que são geradas por diversas fontes como, por exemplo, Web, as redes de sensores e vários outros meios de comunicação, estão na ordem de centenas dados não estruturados. E com esta quantidade de dados não estruturados, novos grandes desafios surgem na forma de armazenamento, manipulação e no processamento de consultas em muitas áreas da computação, especialmente na área de bases de dados, mineração dos dados e recuperação das informações (VIEIRA, 2012).

O conceito do Big Data consiste em dados que excedem a capacidade do armazenamento de banco de dados de sistemas convencionais (DUMBILL, 2012).

Ela também pode ser definida como uma tecnologia que descreve quantidade de dados estruturados, dados semiestruturados e não estruturados, que tem a capacidade de explorar esses dados para obter melhores informações para a empresa, ressaltando que há um desafio no que se refere a gestão e a análises desses dados (LOHR, 2012).

Figura 3- Estrutura do Big Data



Fonte: Taurion (2011). ADPTADO

O Big Data pode ser caracterizado por 5Vs. Primeiro volume, o volume é a dimensão mais comum no que concernem os conceitos do Big Data, visto que o termo volume vem tendo maior atenção pela forma com que os dados vêm aumentando. Cada dia são gerados volumes de dados pela sociedade (TAURION,2012)

Primeiro: Big Data surgiu para tratar volumes extremos de dados estruturados, semiestruturados e dados não estruturados. É difícil analisar esse volume de dados por sistemas que não tem o auxílio do Big Data, visto que é um desafio armazenar esse grande volume de dados com as ferramentas tradicionais (TANKARD, 2012).

Segundo: velocidade, Big Data, aproveita o conhecimento desenvolvido associado a evolução tecnológica, devolvendo o resultado das consultas extremamente mais rápido em comparação os tradicionais, visto que é necessário que as transmissões das informações sejam feitas a uma velocidade bastante dinâmica, porque muitas das vezes precisamos agir em tempo real, exigindo um processamento que acompanhe esta velocidade (TAURION, 2012).

Terceiro: variedade, Callebaut (2012) afirma que quando se fala em Big Data, fala-se em diferentes tipos de dados que podem ser logs de servidores, de imagens, textos, documentos digitalizados, e outros conteúdos gerados pelas redes sócias. A existência desses dados estruturados, semiestruturados e não estruturados são variedade de dados. Big Data opera esses variados formatos de dados com a precisão e a variedade de forma análoga ao tratamento estruturado.

Quarto: veracidade é preciso destacar o conteúdo mais necessário em meio a tanta informação, Big Data proporciona confiança no tratamento das informações garantindo o máximo de consistência possível, além de garantir a separação das informações mais importantes para compreender melhor os clientes e o seus comportamentos (IMB, 2011).

Quinto: valor, o uso do Big Data torna-se importante quando todos os esforços são direcionados para gerar valor, que é a informação da análise e

sua conversão em material que serão aplicados nas decisões de uma certa empresa.

Big Data de certa forma é responsável por ganhos nas tomadas de decisões caracterizando assim o valor da informação obtida (TAURION, 2012).

A compreensão desses conceitos e das suas aplicações irá resultar em uma transformação no mundo dos negócios, conhecido como uma economia mais inteligente e analítica.

Segundo Boyd e Crawford (2011) saber lidar com o surgimento de uma era do Big Data é fundamental para questionar as incertezas e as mudanças rápidas no nosso processo decisório, pois, as decisões atuais terão considerável impacto no futuro, razões que levam as organizações aplicarem está tecnologia e porque vem apresentando resultados satisfatórios.

Aplicando os 5vs do Big Data em determinadas áreas permite maior eficiência na previsão de tendências do mercado, o conhecimento de atitudes para fidelizar com os seus clientes e melhorar os processos internos da empresa para atingir o sucesso. Visto que a ideia central do conceito Big Data é a tomada de decisão em tempo real sobre uma gama de dados (TANKARD,2012).

5.1 APLICAÇÕES DO BIG DATA

Como estamos vivendo em uma era que os dados são gerados a cada milésimo de segundos, obter informações através desses dados é um fator diferenciador no cenário de negócios, na saúde, nas universidades e muitas outras áreas.

A tecnologia Big Data faz parte de praticamente todas as áreas que envolvem o tratamento de dados da sociedade. E sua aplicação é particular dependendo da área para atingir a meta que ser quer chegar, ou ajudar na resolução de um determinado problema (IBM, 2011).

Aplicando Big Data é possível detectar problemas de desempenho numa determinada máquina, ajudar a transformar os dados exploráveis, a analisar os

fluxos de dados, e numa rede permite analisar e prever os comportamentos dos consumidores (CRAWFORD, 2011).

Existem vários casos que a tecnologia do Big Data foi utilizada para solucionar problemas que antes eram impossíveis e difíceis de serem resolvidos. No Brasil existem alguns estudos, um deles, o da pesquisadora Breitman (2012), afirma que “A descoberta do pré-sal foi possível por causa do Big Data”, neste caso trata-se dos dados sísmicos advindos de sondas que estão no fundo do mar, que geram um alto volume de dados de entradas de diferentes variáveis que são filtradas e processadas para os conhecimentos confiáveis da superfície avaliada.

São inúmeros casos que o Big Data vem solucionando, como é o caso de uma empresa norte-americana de transportes e logística, UPS tinha o desafio de aperfeiçoar o seu cronograma de rotas, tornando mais eficientes e menos dispendiosas possíveis. Era necessário fazer um estudo matemático e milimétrico, cruzando todas as rotas de todos os seus veículos, avaliando também velocidade, pressão, e o tempo gasto em cada entrega. A imensidão de variáveis tornava impossível à mente humana chegar a resultados sólidos. Entretanto com a aplicação do Big Data a empresa conseguiu chegar a uma solução com suas rotas. Depois de um período de testes, resultados práticos foram alcançados, quase cinco milhões de litros de gasolina economizados.

Outro caso é o do banco norte-americano J.P Morgan que trabalha com estudos de algoritmos complexos, conceitos de redes neurais e cruzamento de dados por meio por meio de sistemas de Big Data para prever tendências e entender quando comprar ou vender ações (CHARLLES, 2014).

Aplicação do Big Data na medicina permitiu os médicos a decifrar cadeia de DNA completa por apenas alguns minutos ele traz resultados que ajudam a descobrir curas de doenças que até então eram difíceis de serem diagnosticados (RODRIGO, 2014).

Uma pesquisa da Tata Consultancy Services fez uma entrevista com mais de 1.217 empresas americanas no ano de 2012 e 2013. A pesquisa contatou que 53% das empresas já utilizam Big Data, e muitas delas relatam que tiveram resultados satisfatórios.

6 PROCESSO DE MINERAÇÃO DE DADOS

Antes do conceito geral sobre mineração de dados, deve-se primeiro entender o que é o processo de mineração de dados como um todo, incluindo todas as suas etapas, fases e processos.

KDD como é denominado significa Descoberta de Conhecimento em bases de dados que em inglês seria Knowledge Discovery in Databases, sendo que a mineração de dados é apenas umas das etapas do KDD que seria no caso a etapa de processamento dos dados.

Mineração de dados é uma pratica para pesquisar, ou analisar grandes quantidades de dados automaticamente, afim de descobrir padrões e tendências que vão além de uma analises simples (BEAL, 2011).

Mirkin (2012) definiu mineração de dados como um processo analítico para a exploração de grandes quantidades de dados. Mineração de dados pode ser também entendido como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomadas de decisões.

A mineração de dados define o processo automatizado de capturar e analisar os dados para extrair um significado, sendo usado tanto para descrever características do passado ou para predizer tendências para o futuro.

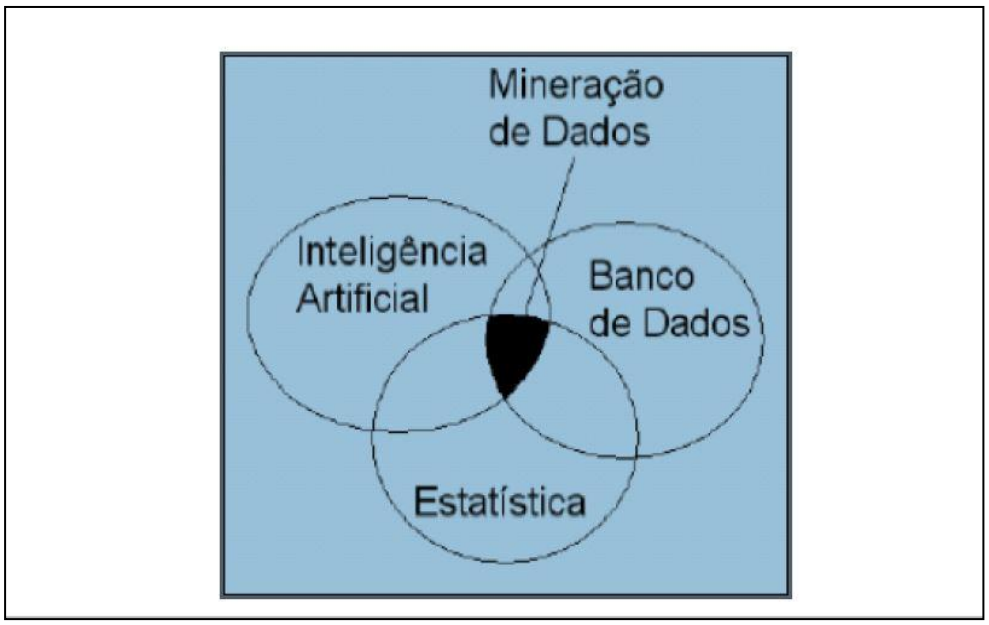
É uma prática aplicada em diversas áreas que usam o conhecimento para extração de informações, como empresas, industrias e instituições de pesquisas.

Mineração de dados é uma tecnologia que surgiu da intersecção das áreas como a estatística, inteligência artificial (KAMBER, 2009).

Esta tecnologia é um conjunto de várias outras ciências e técnicas usadas para como extração de conhecimento como é o caso dos os algoritmos matemáticos, estatística como por exemplo, para o segmento dos dados e a estatística para avaliar a probabilidade de eventos futuros.

A figura 12 mostra o processo e outras técnicas utilizada para a mineração de dados.

Figura 12: Mineração de dados



Fonte: Silva (2010)

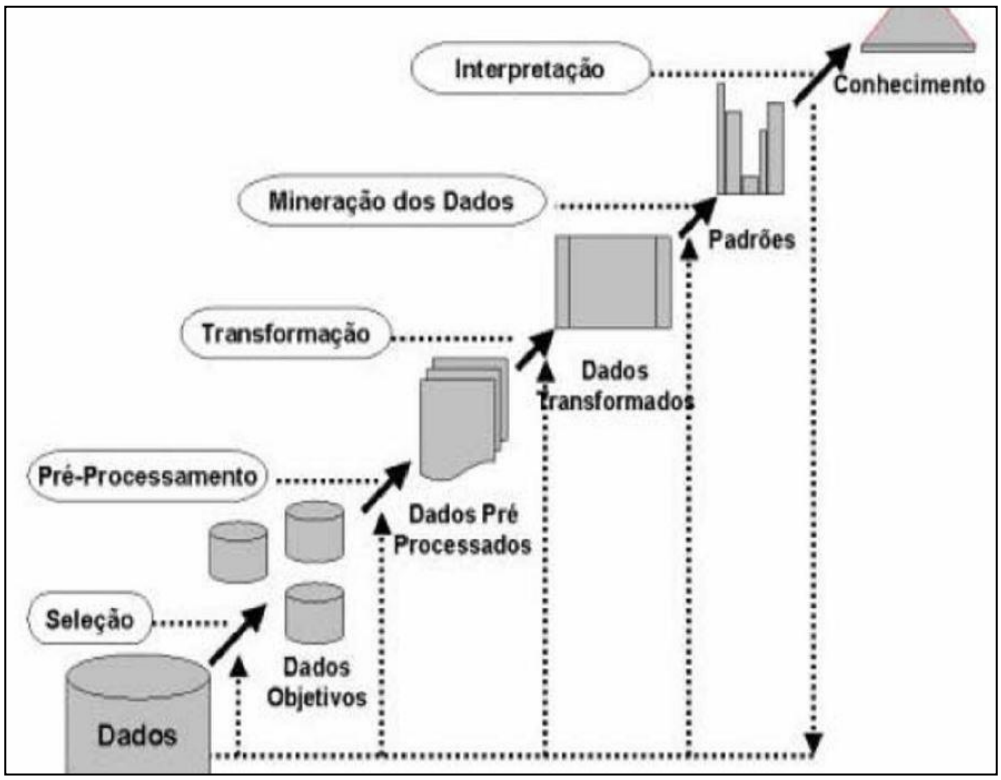
Para encontrar soluções ou, extrair conhecimento existem diversos métodos para mineração de dados. É preciso estabelecer metas para alcançar o que está sendo proposto (MARTINS, 2006).

Hoje em dia o processo de KDD que é conhecido popularmente como mineração de dados ou Data Mining é uma tecnologia que tem sido muito usado em casos onde é necessário a extração do conhecimento.

Podemos citar alguns casos aonde tem sido usado por exemplo: na tendência de consumo de vários clientes, na detecção de fraudes em arrecadações, na segmentação de mercados, previsão do volume de vendas, no planejamento de produção e principalmente na previsão de mercados financeiros, e várias outras áreas.

O processo de KDD é constituída por várias etapas operacionais como: interpretação mineração de dados, transformação, pré-processamento, e a seleção dos dados.

Figura -11 processos de KKD



Fonte: Donny (2013)

6.1 PRÉ – PROCESSAMENTO

O pré-processamento é a etapa que se responsabiliza por consolidar as informações relevantes para o algoritmo minerador, ou seja tem como objetivo a redução da complexidade de um determinado problema.

Gold Schmidt (2005) afirma que a principal função da fase do pré-processamento é na preparação dos dados que a serem aplicados na mineração de dados.

6.2 Seleção de dados

Seleção de dados também conhecido como redução de dados, ela realiza a identificação e a seleção das informações mais relevantes, entre as informações existentes na base de dados original que serão utilizados durante o processo. (AMORIM, 2006)

6.2.1 Limpeza de dados

A limpeza de dados consiste em tratar os dados que foram selecionados na etapa anterior, ou seja para não ter que comprometer na qualidade dos resultados a serem obtidos no final do processo é necessário que algumas correções sejam efetuadas na base de dados em caso de houver informações nulas, redundantes ou inconsistentes. Tendo, assim, a veracidade e a integridade dos fatos que foram representados por eles (GOLDSCHMIDT, 2010).

6.3 MINERAÇÃO DE DADOS

A mineração de dados é a etapa mais importante da metodologia, visto que é a etapa aonde são definidas as técnicas e tarefas, incluindo os algoritmos, que serão utilizados no processo da mineração, sendo assim, realizada a busca pelo conhecimento pretendido a ser extraído em uma base de dados

6.3. 1 Tarefas

Existem muitas tarefas em Data Mining, onde cada uma dessas tarefas extrai um tipo de conhecimento diferente em uma base de dados e também, estão diretamente associados no domínio das aplicações e ao interesse do usuário

As tarefas dependem muito dos objetivos que se pretende atingir com a aplicação, e não podemos esquecer também que pode se usar as tarefas isoladamente ou fazendo combinações entre elas dependendo da finalidade que queremos atingir.

6.3 .2 Previsão

A previsão como o nome já diz, é a avaliação do valor futuro, ou seja, de algum índice, baseando-se em dados do comportamento passado. Muitos especialistas afirmam que a previsão é uma das tarefas mais difíceis, e que o único modo de se ter a certeza se porventura a tarefa foi bem realizada é esperar o acontecimento, e posteriormente verificar os resultados obtidos.

Previsão é um dos métodos que muito tem sido utilizado para resolver certos problemas no mundo real, onde a eficácia de uma certa decisão depende muito dos eventos anteriores e ela mesma, é uma das tarefas que ajuda na redução de riscos gerados pelas incertezas e acaba ajudando no planejamento e nas tomadas de decisões.

Muitos algoritmos podem ser usados na previsão, com características e funcionalidades diferentes, os mais usados são: Árvore de decisão, Estatística, Regressão e RNA.

8 TABLEAU

Segundo a IBM (2011), nos próximos anos o mundo irá gerar 50 vezes a quantidade de dados do que atualmente, visto que com a expansão da internet e vários outros meios de informação a probabilidade é sempre de gerar mais dados.

Tableau surgiu para transformar as oportunidades em realidades, usando técnicas de visualização para explorar e analisar as bases de dados (WHITE, 2011).

Tableau é uma ferramenta de inteligência de negócios que permite os usuários a ver e entender os seus dados. É uma aplicação que as pessoas usam para qualquer nível de habilidade fácil de usar e que permite análises rápidas

Tableau usa a capacidade em memória ou pode ser usada de forma hibrida consultando diretamente, permitindo que exista um desempenho rápido e em grandes conjuntos de dados com múltiplas fontes.

O software também permite os usuários a se conectar a volumes de dados em vários formatos e visualiza-los em tempo real. É uma das ferramentas mais eficazes no que concerne a visualização de dados, possibilita as organizações representam os dados em formatos mais compreensíveis (DONNEY, 2013).

Para os administradores, há muitas vantagens em usar as fontes de dados do Tableau. Como fonte de dados pode ser usada muitas pastas de trabalho. Uma fonte de dados que inclui uma extração significa que economiza espaço no servidor e tempo de processamento.

As atualizações da extração podem ser agendadas por extração, e não por pasta de trabalho, e quando uma pasta de trabalho que usa uma fonte de dados do Tableau for baixada, a extração de dados permanecerá no servidor, resultando em menos trafego de rede (DONNEY, 2013).

9 TRABALHOS CORRELATOS

Big Data tem sido utilizado especificamente quando se trata de lidar com qualquer tipo de dados, e transformar negócio mudando assim a maneira de como analisar as informações usadas e tomadas de decisão.

Existem grandes quantidades trabalhos desenvolvidos na área de Big Data em monografias, teses, artigos entre outras, relatando a importância do Big Data na solução de problemas.

9.1 NoSQL DATABASE: NEW ERA OF DATABASES FOR BIG DATA ANALYTICS-CLASSIFICATION, CHARACTERISTICS AND COMPARISON

Este trabalho cientifico foi escrito por A B M Moniruzzaman e Prof. Dr. SyedAktherHossain em 2013 para o InternationalJournalofDatabaseTheoryandApplication. O objetivo do artigo é prover a classificação, características e avaliações de banco de dados NoSQL em aplicações Big Data. Oferecendo uma visão independente das vantagens e desvantagens de várias soluções NoSQL no processamento de grandes volumes de dados.

A pesquisa análise os conceitos do banco de dados NoSQL e Big Data, e realiza uma análise comparativa entre alguns bancos NoSQL através de uma tabela, cruzando modelos de banco orientado a documentos.

Os pesquisadores concluíram que os bancos NoSQL surgiram para liderar o desenvolvimento de aplicações de Big Data e analise de negócios, aplicações que já haviam chegado ao limite dos bancos de dados relacionais tradicionais.

9.2 A TOMADA DE DECISAO NO CONTEXTO DO BIG DATA (estudo de caso)

Este trabalho foi desenvolvido como projeto de graduação da Vivian Passos Canary, para a Universidade Federal do Rio Grande do Sul, Brasil.

Este trabalho foi desenvolvido com o objetivo verificar o efeito dos fatores 5vs (volume, variedade, velocidade, valor e veracidade) do Big Data no processo de tomada de decisão de executivos de diferentes níveis hierárquicos em um Sistema de Credito Corporativo.

A pesquisadora ao longo do trabalho fez um estudo de caso para saber dos gestores o que eles têm feito para tomada de decisão nas suas empresas, e que tipo de ferramentas usam para o auxílio na tomada de decisão, visto que os gestores são as pessoas que tem como responsabilidade tomar decisão e criar regras para o bom funcionamento da empresa.

A pesquisadora concluiu que a contribuição dos gestores envolvidos no trabalho foi possível aproveitar os resultados obtidos para auxiliar a sua reflexão a respeito do uso do Big Data no tratamento da informação no seu processo decisório, minimizando assim, os riscos incorridos em uma tomada de decisão equivocada.

9.3 A INFLUENCIA DO BIG DATA NO BUSINESS INTELLIGENCE

Este artigo cientifico foi desenvolvido por Diego Delsoto, como projeto de graduação na Faculdade IBTA, Campinas, Brasil.

O artigo de pesquisa trata sobre o impacto que o Big Data tem no Business Intelligence. O pesquisador concluiu ao longo da pesquisa que o Big Data permitirá que sejam feitas análises completamente add-hoc (aleatórias) e com bastante velocidade, acelerando os processos de negócio e tornando-os mais assertivo. Toda essa perspectiva do Big Data mostrará para as organizações a capacidade de armazenamento e a forma inteligente de se trabalhar em tempo real com seus dados, oferecendo uma análise mais profunda e exploratória, enquanto que os sistemas de Business Intelligence fornecerão uma análise mais estruturada. Por fim, ambos serão importantes para obter vantagem competitiva e um plano estratégico nos seus processos de negócio, obtendo controle e gerenciamento de desempenho, relatórios analíticos, painéis de visualização.

9.4A ESTUDY OF NOSQL AND NEWSQL DATABASES FOR DATA AGGREGATION ON BIG DATA

Este trabalho foi desenvolvido como projeto de mestrado de Ananda SentrayaPerumalMurugan para o Royal Instituteof Technology de Estocolmo, Suiça. Este trabalho foi desenvolvido com o objetivo de comparar uma solução NoSQL com uma solução NewSQL a fim de identificar qual se encaixaria melhor em uma aplicação de análise de dados utilizada pela Scania.

O departamento de TI da Scania utiliza um banco de dados relacional para armazenar os dados de todos os veículos e prover analise em cima destes dados. Por ser utilizado um banco relacional tradicional, esta aplicação consome grande volume de memória e tempo, por isso foi proposto uma solução distribuída para solucionar este problema. Foi realizada uma análise pratica comparando uma solução NoSQL com uma solução NewSQL, a fim de identificar qual seria mais adequada para a solução do problema. A comparação foi feita analisando o desempenho dos bancos de em um caso de uso comum e com uma tabela comparando características dos sistemas

9.5 NoSQL DATABASE: NEW ERA OF DATABASES FOR BIG DATA ANALYTICS-CLASSIFICATION, CHARACTERISTICS AND COMPARISON

Este trabalho cientifico foi escrito por A B M Moniruzzaman e Prof. Dr. SyedAktherHossainem 2013 para o InternationalJournalofDatabaseTheoryandApplication. O objetivo do artigo é prover a classificação, características e avaliações de banco de dados NoSQL em aplicações Big Data. Oferecendo uma visão independente das vantagens e desvantagens de várias soluções NoSQL no processamento de grandes volumes de dados.

A pesquisa analisa os conceitos do banco de dados NoSQL e Big Data, e realiza uma análise comparativa entre alguns bancos NoSQL através de uma tabela, cruzando modelos de banco orientado a documentos.

Os pesquisadores concluíram que os bancos NoSQL surgiram para liderar o desenvolvimento de aplicações de Big Data e analise de negócios.

11 DESENVOLVIMENTO DA METODOLOGIA

Os dados utilizados neste trabalho foram criados com base nas informações que nos foi fornecida pela empresa Duda Imóveis. O arquivo da nossa base de dados tem por volta de 2GB, e estão no formato Excel.

Os nossos arquivos, possuem várias patentes juntas e cada patente possui informações importantes como, números e data de venda, classificação por estados, cidades e informações adicionais dos clientes. Esses dados são suficientes para serem geradas informações que auxiliem no processo de decisão muito importante para empresas.

11. 1 Ambiente de Processamento

Para o trabalho, foram utilizados a base de dados e o software Tableau. Para auxiliar na manipulação dos arquivos Excel foi utilizado um conector que fornece uma maneira simples de fazer um mapeamento. Outra questão importante é que os dados originais precisam ser ajustados, pois o Tableau por padrão particiona os arquivos a serem visualizados por linha de arquivos.

Esse comportamento obviamente não funciona com os arquivos Excel pois precisamos que o conector recebe uma entidade Excel para que ela consiga extrair informações patentes, ou seja o conector determina qual é o formato da entrada dos dados e como ele deve ser visualizado, transformando do formato Excel para o formato visual.

Posteriormente o Tableau, manipula as informações no formato visual, pois é preciso fazer a transformação dos dados do formato original Excel para serem enviados para o armazenamento.

Apesar dessa transformação dos dados ser relativamente simples o maior problema é trabalhar com arquivos Excel, ou XML que acabam por consumir muita memória durante seu processamento. Para este caso existe uma biblioteca no software tableau que lê os documentos em Excel, converter cada entidade em patente para a planilha e após isso enviar para o dashboard.

Aqui também tratamos o seguinte problema: Como executar as visualizações nos dados de patentes e como armazenar estes dados dão suporte consultas eficientes. Para a resolução deste problema o software usado consegue armazenar os dados e abstrai em forma de consultas SQL. No momento do processamento a consulta SQL é transformada em visualizações.

Finalmente, precisamos ainda adequar os nossos dados a estrutura que ele fornece para visualizar, então nosso trabalho aqui é transformar os dados Excel em um arquivo com formatos visual e importar os dados para dentro do Tableau.

10 TRABALHO PROPOSTO

O presente trabalho consistiu em realizar uma metodologia para mineração de dados em bases de dados não estruturados no software Tableau. Neste projeto foram analisadas técnicas de mineração de dados, e também foi feito um estudo da ferramenta Tableau usadas para visualização e recuperação dos dados. Após estudos e analisadas as técnicas de mineração de dados foi selecionada uma técnica melhor que se adequa com a implementação deste estudo. Também foi desenvolvido um estudo de caso, na empresa Duda Imóveis para a coletas de dados e informações afim de se ter uma ideia de como é estruturado um banco de dados de uma empresa do setor imobiliário e quais informações ela contém. Após o desenvolvimento do estudo criou-se uma base de dados para a demonstração da mineração em uma base de dados não estruturados com o auxílio do software Tableau para a visualização dos dados.

10.1 METODOLOGIA

Para atingir os objetivos deste trabalho de conclusão de curso foi necessário seguir algumas etapas e estabelecer o que era necessário em cada uma delas até a conclusão.

Foi realizado também um levantamento bibliográfico a partir de artigos, livros, e-books, pesquisas na internet e em publicações, a fim de coletar dados para a elaboração da fundamentação teórica deste projeto que teve início com as pesquisas relacionadas a gestão. Na gestão, foram abordados o que é a gestão, tipos de gestão e os níveis de gestão. Em banco de dados abordou-se o que é um banco de dados: sistemas gerenciadores de bancos de dados: arquitetura e modelagem de banco de dados: modelos de bancos de dados.

Em seguida foi feita a fundamentação teórica sobre os dados: tipos de dados: o que são dados: dados estruturados: dados semiestruturados: e dados não estruturados. Fundamentou-se também sobre a tecnologia Big Data, sua

definição: arquitetura e suas aplicações em várias áreas como em medicina, empresas e na agricultura.

10.2 ESTUDO DA FERRAMENTA TABLEAU

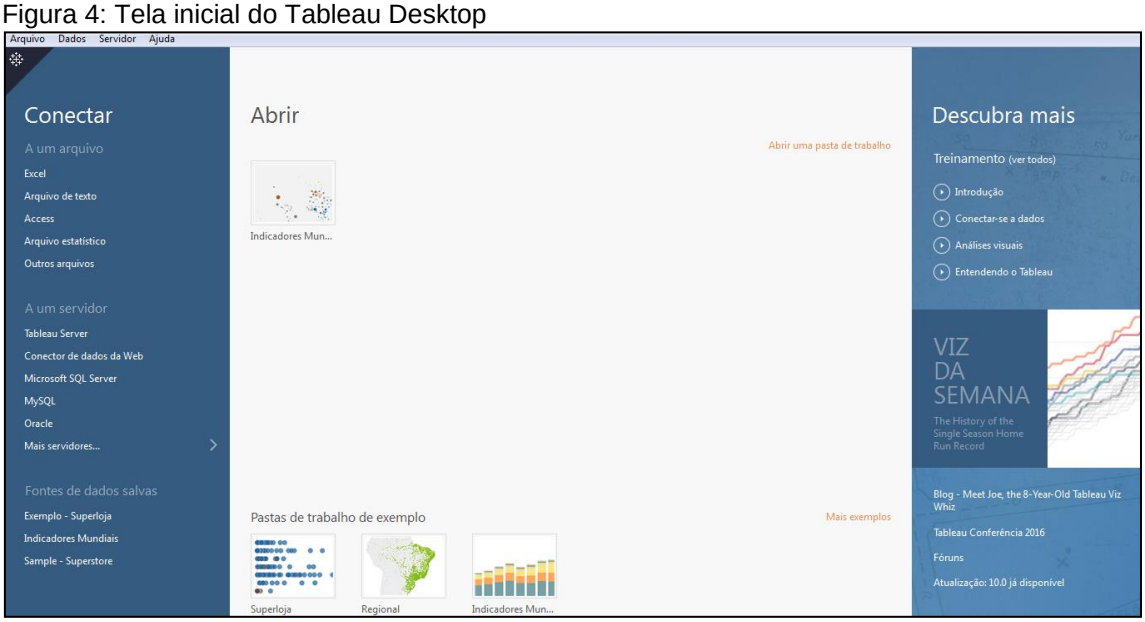
A ferramenta escolhida para a parte prática deste trabalho afim de visualizar os dados foi Tableau software versão 9.3.0. Tableau software é uma ferramenta gratuita que oferece quatro edições Tableau Desktop, Tableau Server, Tableau Online e o Tableau Public.

a) Tableau Desktop: oferece o acesso local, com licença de avaliação gratuita por 14 dias;
b) Tableau Server: permite o acesso web e móvel com uma infraestrutura gerenciada pela organização contratante;
c) Tableau Online: permite acesso web e móvel, porém é oferecida como um serviço que é gerenciada pela própria Tableau;
d) Tableau public: oferecida gratuitamente, especificamente a jornalistas e alguns usuários que desejam compartilhar seus dados na web**.**

Com o Tableau é possível criar pastas de trabalho, pode conter planilhas, painéis e histórias. A planilha é formada por uma visualização que pode ser gráfico, tabelas. E a ferramenta também trata como dimensão qualquer campo como datas ou cadeias de caracteres.

Os painéis que também são conhecidos como dashboards são formados por conjuntos de visualizações que estão presentes nas planilhas, enquanto que uma história contém uma sequência de planilhas ou painéis que trabalham juntos para transmitir informações.

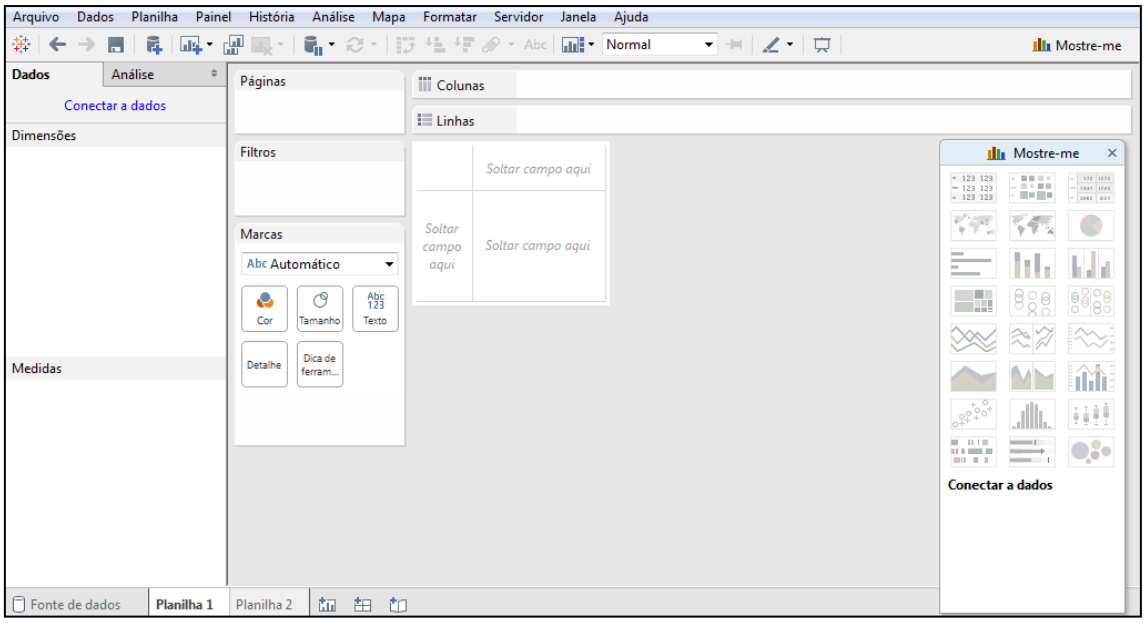
Figura 4: Tela inicial do Tableau Desktop



Fonte: Tableau (2016).

A figura 4 mostra a página inicial do Tableau Desktop, a partir da tela inicial é possível criar uma nova conexão a uma fonte de dados. O software oferece mais de trinta conectores nativos que permitem a conexão com diversas fontes como: banco de dados arquivos, Excel CVS, data warehouse corporativo ou dados baseados na web. Sua visualização de dados é muito interativa e simples, dando opções aos usuários em destacar seções de pesquisas em gráficos sem ter habilidades ou assistência em TI. Além disso, o usuário não tem que tomar todas as medidas para fazer painéis mobile, o próprio Tableau reconhece automaticamente se o usuário está usando o aplicativo móvel e faz os ajustes.

Figura 5: planilha do Tableau Desktop



Fonte: Tableau (2016)

A figura 5 é a demonstração de uma nova planilha, o Tableau organiza os campos em dimensões e medidas que podem ser arrastados com o objetivo de criar visualizações. Por meio do software é possível personalizar qual nível de dimensão que se deseja utilizar. Um outro recurso interessante é a barra de ferramentas “Mostre me”, por meio da qual o Tableau indica quais melhores tipos de visualizações disponíveis que podem selecionados pelo usuário.

Umas das principais vantagens do Tableau software é permite que um utilizador, sem necessidade de qualquer formação pode acessar, interagir, agregar, filtrar e produzir componentes graficamente ricos a partir dos seus dados.

10.2.1 ESTUDO DE CASO

O estudo de caso para a conclusão deste projeto consistiu em realizar uma metodologia para mineração de dados não estruturados, utilizando o software Tableau. O grande problema é estabelecer quais são as informações

necessárias e como traze-las do Big Data. O enfoque aqui é mostrar como a técnica da mineração de dados foi modificada para ser aplicada em neste trabalho.

Visto que com o crescimento dos dados gerados nos últimos anos, e que a maior parte desses dados são dados não estruturados composto pelos formatos doc, docx, pdf, XML, HTML, redes sócias e outras infinidades de formatos. As empresas passaram a ter imensas dificuldades em lidar com esses dados como por exemplo: armazenar, consultar, gerenciar, e tomar decisões afim de subtrair informações para melhor tomada de decisões.

10.2.2 DESCRIÇÃO DO ESTUDO DE CASO

Por meio desse estudo de caso foi possível analisar o comportamento de uma empresa do setor imobiliário e descobrir, por exemplo, padrões como: Em que trecho das cidades ocorrem mais procura de imóveis?
Quais são os tipos de imóveis que os clientes mais procuram?

Este estudo pode auxiliar na busca por soluções eficazes. Dessa forma, a crescente popularidade desse tipo de busca, nos leva a crer que em pouco tempo, podem se tornar a principal fonte avaliação em tempo real.

Por meio desses dados foi possível analisar o comportamento da procura nas cidades e descobrir, por exemplo, padrões como: detectar em que trechos da cidade ocorrem mais procura com maior frequência, e que tipos de informações sobre os imóveis os clientes mais procuram. Tal descoberta pode auxiliar na busca de soluções eficazes.

Para isso, foi feito um estudo de caso na empresa Duda Imóveis para coletas dos dados.

10.2 .3 Apresentação da empresa

A empresa Duda imóveis foi fundada em 14 (quatorze) de agosto de 1984 por José Mondardo, é uma empresa familiar onde todos participam na

colaboração das filais e matriz. Duda imóveis é uma empresa do ramo de imóveis voltado para o setor imobiliário, que presta seus serviços na cidade de Criciúma rua São José número 44, centro de Criciúma.

A Duda imóveis atua no mercado imobiliário desde sua fundação oferecendo confiança, segurança, credibilidade e atendimento diferenciado a quem deseja alugar, vender, comprar um imóvel ou mesmo administrar o condomínio.

Sendo uma empresa familiar a Duda imóveis conta com o auxílio de Eduardo Gudi Mondardo que está sempre junto ao pai tomando a frente dos setores de locação e vendas, assim como a administradora de condomínio, na região sul do estado. A empresa conta com setores próprios como o jurídico, financeiro e todos outros para melhor atendimento como comodidade e rapidez.

Na cidade de Criciúma a empresa possui três agencias, a primeira iniciada em 1984 com a sua primeira loja na Av. Getúlio Vargas com os setores de vendas, locação e administração de condomínio. Esta agencia trabalha com imóveis totalmente locados pela a própria empresa, configurando uma loja exclusiva para venda de imóveis de terceiros que sempre foi o seu ponto forte.

A empresa conta atualmente com mais de 100 colaboradores, uma equipe com mais de 40 de funcionários e mais de 50 corretores e tendo parceiras com grandes construtoras do momento.

10.2.3 Coleta de dados

A coleta de informações foi realizada através de entrevista e perguntas abertas, com o gerente de TI da empresa o senhor **Sander Fortuna** e com gestora de recursos humanos **Dhieine Bez Batti**, pois são estes que cuidam da parte administrativa operacional da empresa. As entrevistas e perguntas para a elaboração do estudo de caso foram feitas no mês de maio de 2016, do dia 3 a 13 de mesmo mês.

No decorrer das entrevistas e perguntas procurou-se saber respostas de como é estruturado o banco de dados da empresa e quais são os tipos de informações que eles procuram, e como tem feito para lidar com grandes volumes de dados.

Foram coletadas informações sobreo que o cliente quer saber a respeito dos imóveis, como é o prognostico de vendas, além de avaliar situações de determinadas coisas que são atualmente uteis e não tem no mercado. Coisas que os clientes procuram saber mais não estão guardadas de formas estruturadas no banco de dados da empresa, ou seja, informações que não existem no banco de dados.

Para se ter uma ideia de como é montado o sistema de banco de dados de uma empresa do setor imobiliário, foi mostrado a estrutura do banco de dados da empresa, de como é o seu funcionamento para atender as necessidades dos clientes e as ações que eles executam para as tomadas de decisões.

Também foi possível saber do gerente da empresa Duda imóveis quais são as dificuldades que a empresa vem enfrentando no que concerne a extração, e no armazenamento de dados afim de transformar esses dados em informações uteis. Foram feitas perguntas também sobre a tecnologia do Big Data visto que é uma tecnologia que muitas empresas usam para lidar com grandes quantidades de dados, e se a mesma empresa conhece essa tecnologia e se faz o uso dela.

E finalmente, qual é a visão da empresa sobre a tecnologia Big Data, quais são as vantagens e as desvantagens para o seu tipo de negócio. Quais problemas o foram solucionados com o auxílio do Big Data.

10. 2 .4 Análise e interpretação dos dados

Por esta pesquisa ter sido exclusivamente elaborada por uma única empresa, a abordagem utilizada para análise considera-se que, a interpretação dos dados foi qualitativa, visto que ela fornece a possibilidade de os dados serem apresentados de forma estruturada, e posteriormente analisados.

Segundo Silverman (2009), uma vantagem da pesquisa qualitativa é utilizar dados que ocorrem naturalmente para encontrar as sequencias de como os significados dos participantes foram exibidos e, assim, estabelecer e fazer o uso melhor da informação.

A partir das informações coletadas ao longo do estudo de caso baseado em entrevistas e perguntas necessárias para a elaboração do estudo de caso, foi feito um modelo de banco de dados para a manipulação dos dados, capaz de fornecer elementos para a metodologia da demonstração do conceito de mineração de dados com o Big Data.

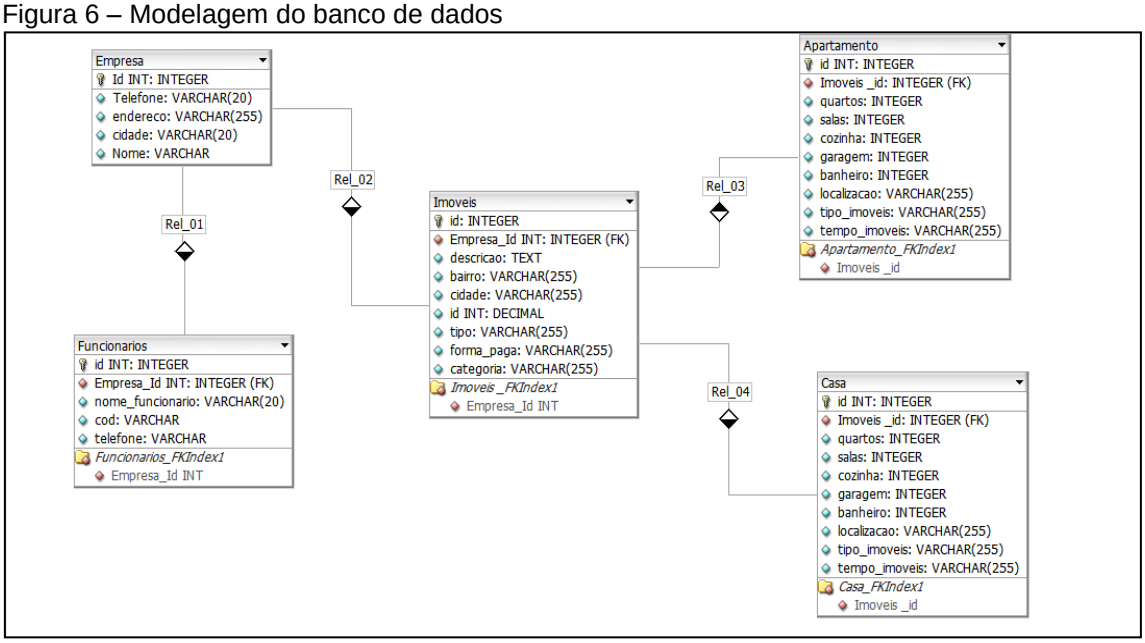
10.3 MODELAGEM DE DADOS

Na criação da base de dados, foi necessário realizar um questionário na empresa Duda imóveis sobre os dados mais importantes existentes em um banco de dados, para o setor imobiliário.

A modelagem do banco de dados, é necessária ao desenvolvimento do projeto pois, através desse modelo implementado será possível ter uma percepção maior de como irá ficar a ligação entre as tabelas, atributos, diminuindo assim possíveis erros que poderiam ocorrer no momento da execução.

Para a modelagem do banco de dados, foi utilizado o DBDesigner 4, um software livre usado para a modelagem de dados visual, sendo assim a figura 6 apresenta o modelo físico do banco de dados.

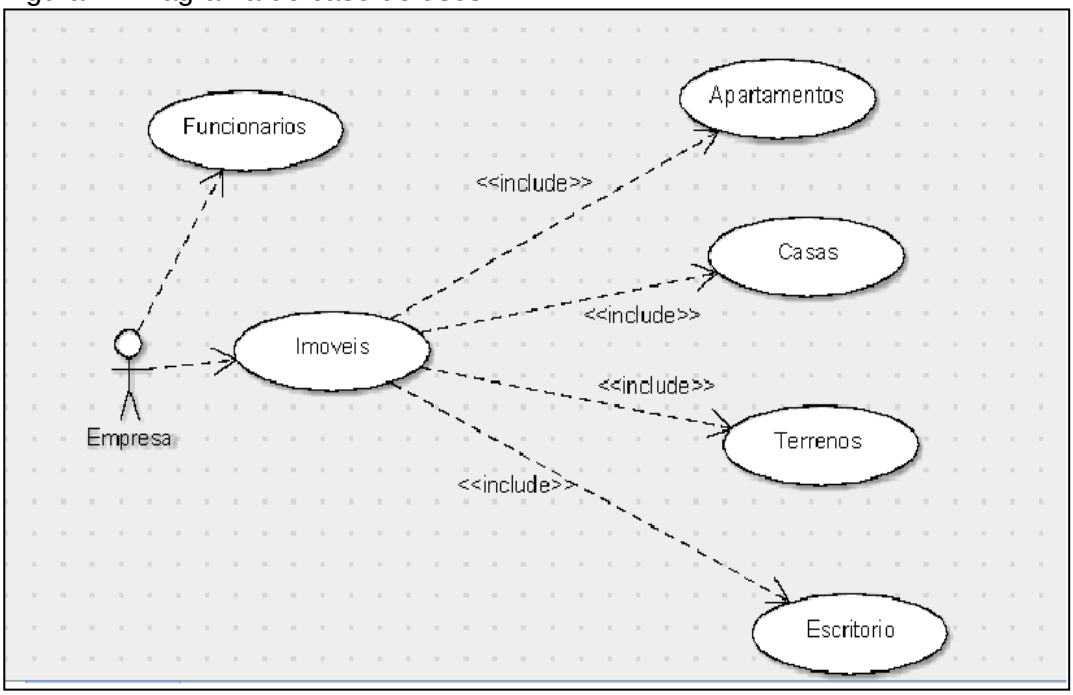
Figura 6 – Modelagem do banco de dados



Fonte: Do autor

O modelo possui todas informações necessárias para a identificação da empresa, desde o cliente, imóveis, e entre outras funcionalidades que se comunicam no banco de dados para a buscar e consultas dos dados. O diagrama de caso de uso, demostra algumas funcionalidades da base de dados, veja a figura 7.

Figura 7– Diagrama de caso de usos



Fonte: Do autor

10.3.1 DESENSEVOLVIMENTO DA METODOLOGIA

Após a configuração das ferramentas e instaladas corretamente e feito alguns testes, deu se o inicio ao desenvolvimento da metodologia. Inicialmente foi feito um estudo no Tableau e alguns testes, onde foram aplicadas algumas funcionalidades da tecnologia, depois se deu início do projeto principal.

10.3.3 Extração dos dados

Os websites na sua totalidade são criados para o auxílio na visualização de informações e não para a exposição dos dados de forma estruturada. Mais a extração dessas informações de forma manual pode levar muito tempo, e ser tão suscetível a erros de extração.

Neste contexto, a extração das informações de forma automatizada possui um papel fundamental para a análise dos dados expostos. Visto que, a

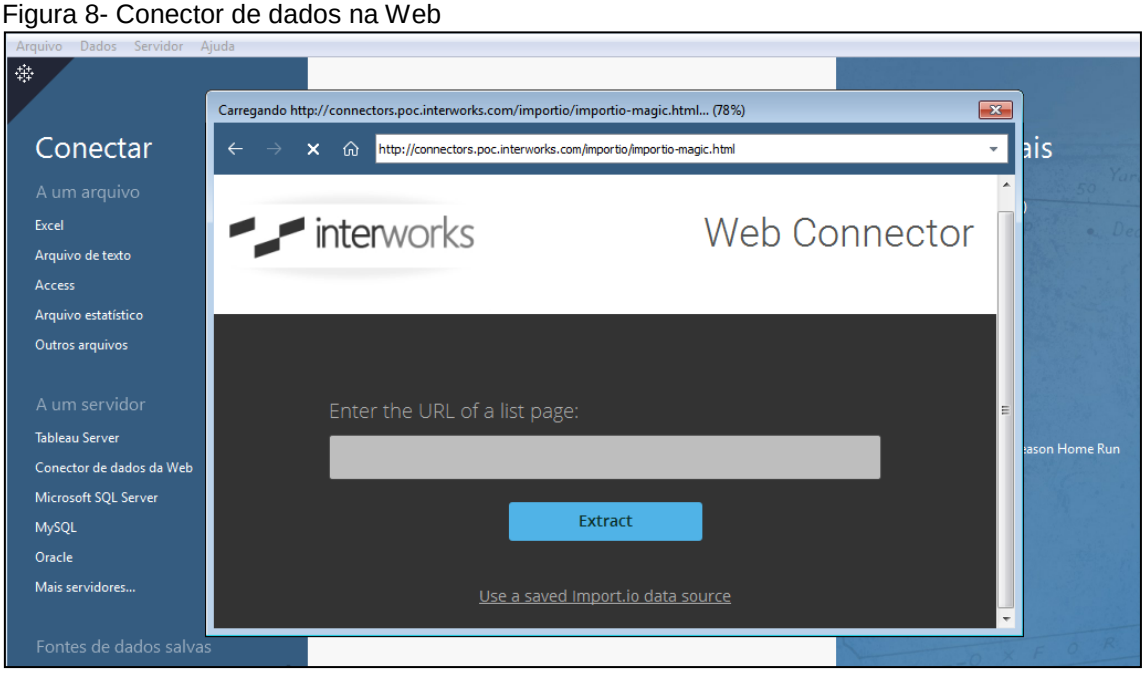
extração dos dados na Web possibilita analisar as informações em tempo real produzidas, alimentadas em plataformas digitais pelos próprios usuários,
Além de auxilia-los diretamente e responder diversos problemas.

A manipulação dos dados neste trabalho foi considerada Big Data, pela presença do grande volume de dados a ser gerenciado, pela variedade dos dados que foram obtidos de diferentes fontes, os quais não possuem um formato de informação definida, outra razão é a velocidade com que essas informações chegam aos conjuntos de dados. Finalmente, são dados que a todo instante são usados por conectores web coletam e bem como a sua velocidade.

Para se conectar aos dados na web a partir do Tableau Desktop é necessário criar um conector de dados na web, este conector pode ser um arquivo HTML que inclui o código Java Script ou um framework. Este conector de dados da Web deve ser hospedado em um servidor da Web executado localmente no seu computador, ou um servidor da Web de terceiros ou no Tableau Server.

Para este projeto foi desenvolvido um conector Web. Este conector fará a interação dos dados existentes na internet com o Tableau, ou seja, partir deste conector é possível pesquisar, ou buscar uma determinada informação na web e fazer a mineração no software além visualizar os dados que se precisa.

Figura 8- Conector de dados na Web



Fonte: Do autor

A figura 9 mostras detalhadamente o software Tableau configurado com o conector Web, como é feita a conexão na Web com o Tableau, usando o Web conector para poder acessar as páginas que se pretender extrair diversas informações.

Um Web conector é um código implementado que permite o Tableau Desktop acessar qualquer fonte de dados na Web, visto que, é necessário conector Web acessar os dados na internet a partir do Tableau Desktop.

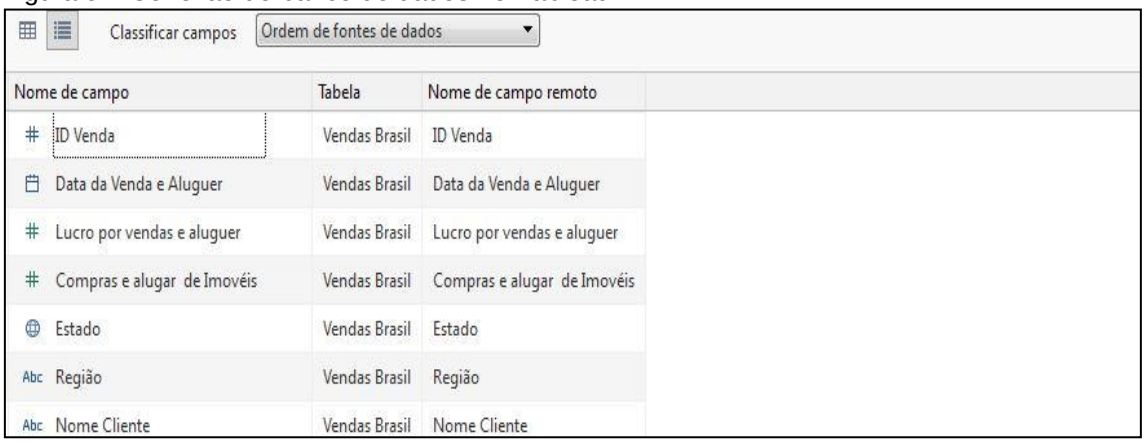
Após a configuração de toda estrutura para se conectar aos dados pelo Tableau Desktop, foi acessado a base de dados criada para fazer uma simulação de cadastro de alguns imóveis, e várias outras ações afins de se fazer uma manipulação de dados para dentro do Tableau. A seguir foi possível extrair e visualizar estes mesmos dados, ou seja, podendo ainda analisa-lo pelo método tradicional.

10.3 .2 Estrutura da conexão com o banco de dados.

Para a realização da conexão do banco de dados no Tableau, foi criada uma classe Connection Tableau com as configurações necessárias para a interação entre aplicação e o Tableau. Neste caso para fazer a interação com o

software o banco de dados utilizado foi o Excel Acess, na figura 8 pode se observar como o nosso banco de dados foi conectado ao sofwareTableau.

Figura 9 – Conexão do banco de dados no Tableau

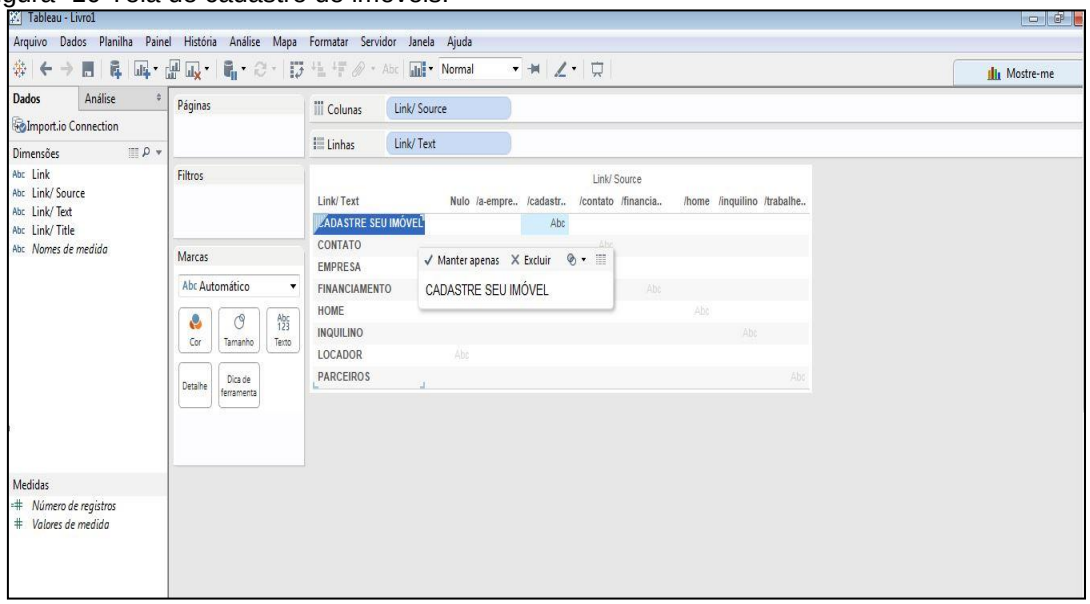


Fonte: Do autor

Depois de feita a conexão com o banco no software Tableau, os dados serão extraídos para a fonte de dados, e o software automaticamente vai executando as consultas, carregando meta dados, gravando os dados, e criar o banco de dados de extrações para o campo de visualizações.

Com esta conexão foram feitos cadastros de vários imóveis como: casas, apartamentos, salas, terrenos, escritórios, garage, e vários outros tipos de imóveis. A construção destes cadastros é como se fosse uma extração de dados para o dashboards que será gerenciado pelo próprio Tableau na figura 11 observamos como foram feitos o cadastro dos imóveis.

Figura- 10 Tela de cadastro de imóveis.



Fonte: Do autor

Depois do cadastro dos imóveis, foi feito a extração dos dados e começou-se a visualizar os imóveis cadastrados. Nos imóveis cadastrados, procuramos extrair e visualizar as informações como:

a) Em quais regiões tiveram maior procura de imóveis.
b) Quais foram os lucros e prejuízos que tivemos em cada cidade.
c) Subcategorias de imóveis que maior número de vendas e locações.
d) Tipos de imóveis tiveram maior procura em todas as regiões

10. 4 EXTRAÇÃO DO CONHECIMENTO ATRAVÉS DA VISUALIZAÇÃO

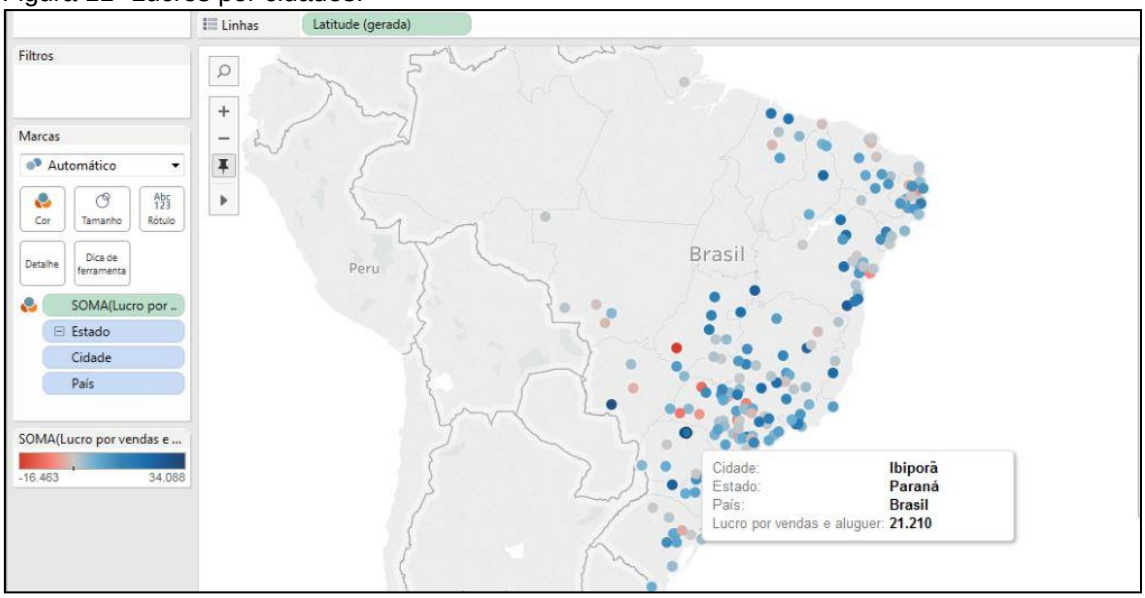
Depois de ter feito a etapa do cadastro de imóveis, realizou-se as consultas no Tableau por meio de visualizações, para sabermos os dados dos lucros e prejuízos em determinadas cidades. Para isso foi feito a importação dos itens cadastrados para o painel do Tableau e automaticamente o software separou o nosso banco em dois painéis, de dimensões e de medidas. Medidas são todos os valores numéricos acumulados, realizando alguns cálculos sobre

eles. Dimensões são informações geográficas por cidades e estados, que foram realizados a venda e alocações dos imóveis.

10.4.1 Visualização da informação dos lucros e prejuízos por cidades.

Para se ter os dados dos lucros e prejuízos por cidades, primeiramente foi feito a configuração das funções geográfica das cidades e dos estados, e em seguida selecionamos o campo lucros por vendas de aluguel por imóveis, em seguida selecionamos as cores no Tableau.

Figura 11- Lucros por cidades.



Fonte: Do autor

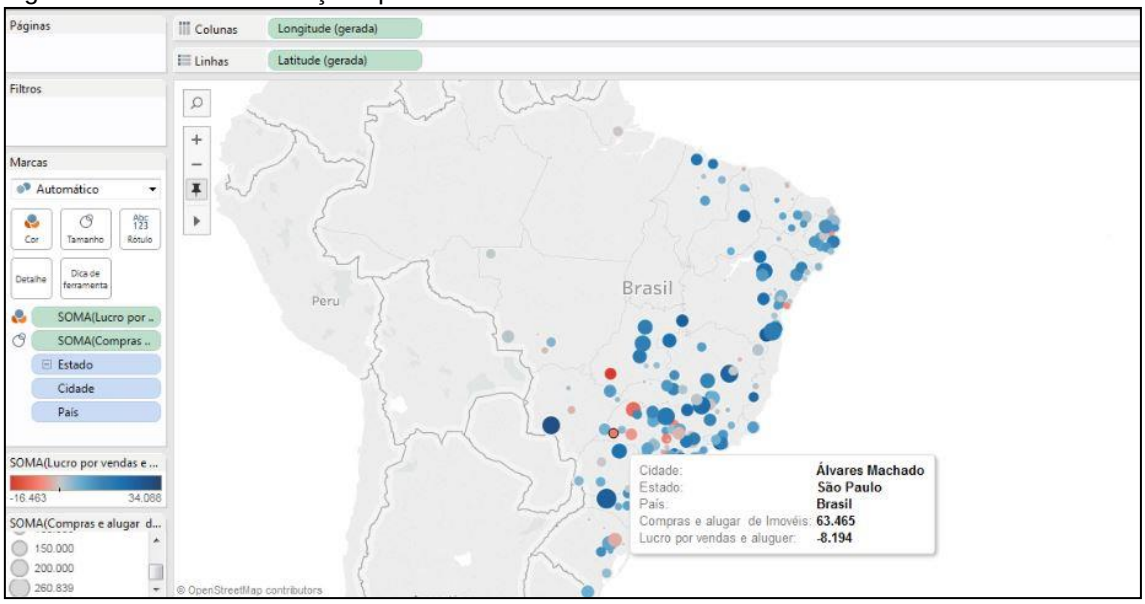
Na figura 12 observa-se os dados obtidos dos lucros por vendas e aluguel de imóveis por cidades, editamos as cores e selecionamos a cor azul como lucro e vermelho como prejuízo. É possível ver na figura que a cidade Ibiporã está marcada em azul simbolizando assim que, é uma cidade que teve lucros positivos.

Temos algumas cidades marcadas com as cores vermelhas, simbolizando que são as cidades com saldos negativos, ou seja, cidades que tiveram menos vendas e alocações.

10.4.2. Extraindo informações das vendas e locações por cidades

Extraímos também da tabela de medidas o campo de compras e vendas de aluguel de imóveis, afim de extrair e visualizar as cidades que tiveram maiores compras e alocações, queremos também saber se as nossas vendas são uniformes nas cidades ou temos cidades mais problemáticas. Para isso selecionou-se o campo do item vendas e compras e arrastamos para a caixa do tamanho do Tableau, essa caixa vai adicionar na nossa visualização as informações das vendas e alocações de imóveis por cada cidade.

Figura 12- Vendas e alocações por cidades



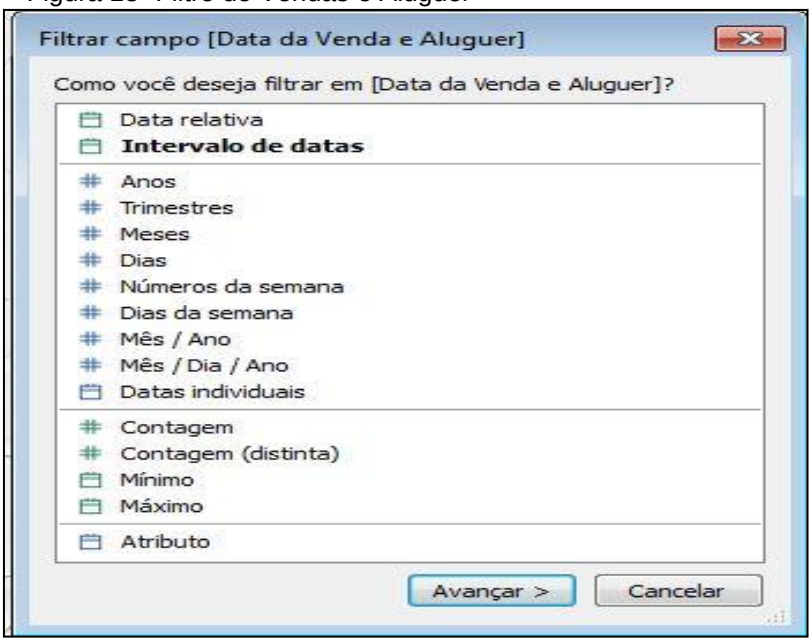
Fonte: Do autor

Observa-se que na figura 13 que as cores tiveram um tamanho maior, isto porque, naquela situação tínhamos as informações dos lucros por cidades, agora temos duas informações em nosso painel de visualização, a informação dos lucros das vendas e alocações por cidades. Selecionamos o curso na cor vermelha e vê-se que a cidade de Álvares Machado doestado de São Paulo é um ponto que teve menos vendas e alocações.

10.4.3 Data das vendas e alocações.

Depois de termos as informações dos lucros, vendas e alocações por cidades, adicionamos também em nosso painel as datas das vendas alocações, para visualizarmos em que data essas vendas e alocações ocorreram assim isto também vai nos permite visualizar as datas em que tivemos os lucros ou prejuízo. Na base de dados criada, apenas existe estruturada o ano como campo de data das vendas. Mais ferramenta Tableau nos permite escolher as datas em vários formatos, como por exemplo por ano, trimestres, meses, dias e por número de semanas, ou seja, ele automaticamente disponibiliza estrutura algumas informações de datas que não existe na base de dados.

Figura 13- Filtro de Vendas e Aluguer



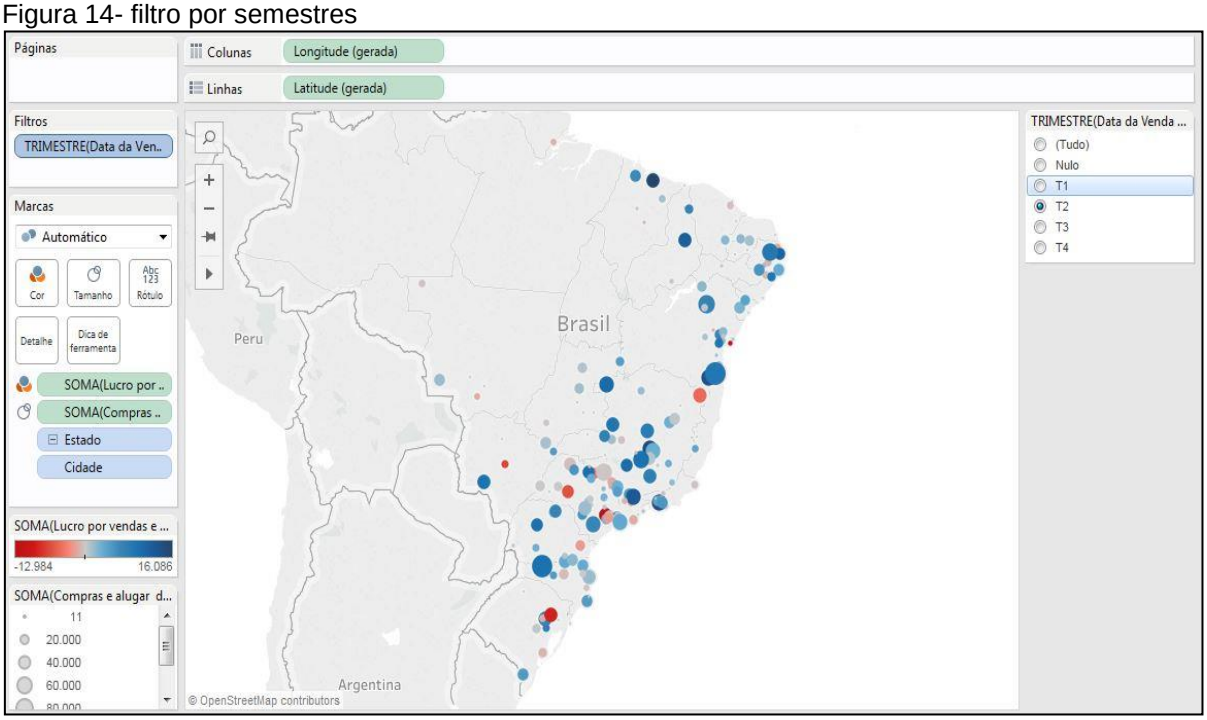
Fonte: Do autor

Selecionamos no painel de dimensões o item data da venda para o nosso painel de visualização, e na planilha filtramos as datas das vendas e locações por trimestres.

Para selecionar as datas de vendas no Tableau, arrastamos da tabela de dimensões o campo da data de venda e adicionamos no filtro. Filtro é uma função que existe no software que tem como finalidade mostrar as opções que

queremos extrair informações, ou seja, em que formato de data desejamos filtrar as informações.

Figura 14- filtro por semestres



Fonte: Do autor

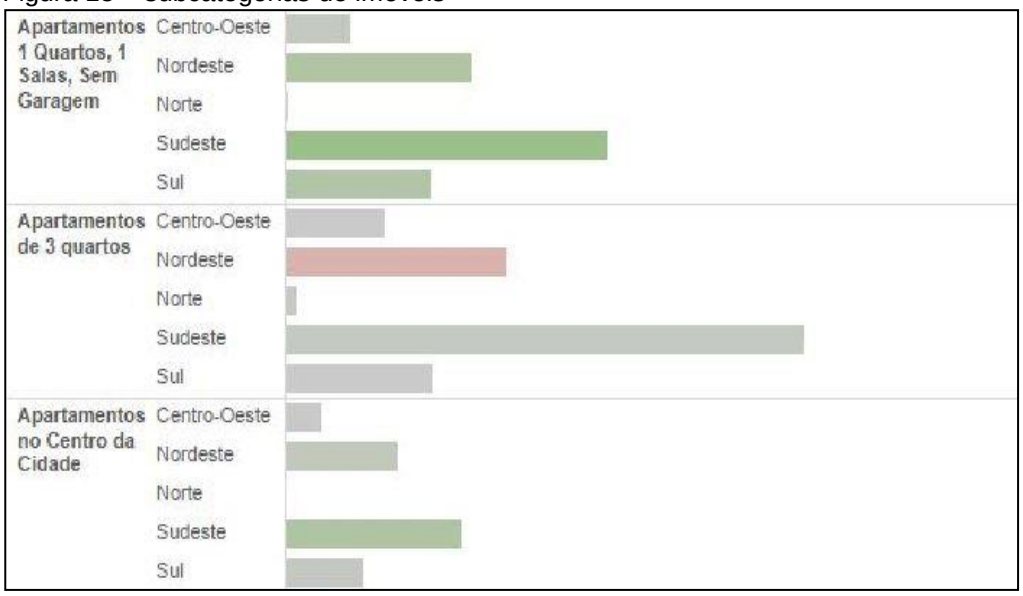
No canto superior direito da figura acima, observa-se o filtro das vendas por semestre na medida em que vamos trocando os números do semestre, os dados presentes na planilha de visualização mostrando os lucros, vendas por cada cidade e dependendo dos semestres selecionado no filtro da data de vendas.

Após a extração e visualização dos lucros das vendas e locações dos imóveis por cidades na planilha de visualização, buscou-se imóveis cadastrados os tipos de subcategorias de imóveis que mais foram procurados, e alocados em determinada regiões.

Na tabela de medidas do software, selecionamos e extraímos o campo de compra de imóveis e aluguel para a planilha, com dois clicks no respectivo campo automaticamente o software mostra os dados do campo na página de visualização, em seguida extrai-se da tabela de dimensões o campo de subcategoria de imóveis para planilha, depois arrastamos o campo região para

a linha do software. Esta linha tem a função de calcular e mostrar na parte da visualização quais regiões tiveram vendas e procuras uniformes.

Figura 15 – subcategorias de imóveis

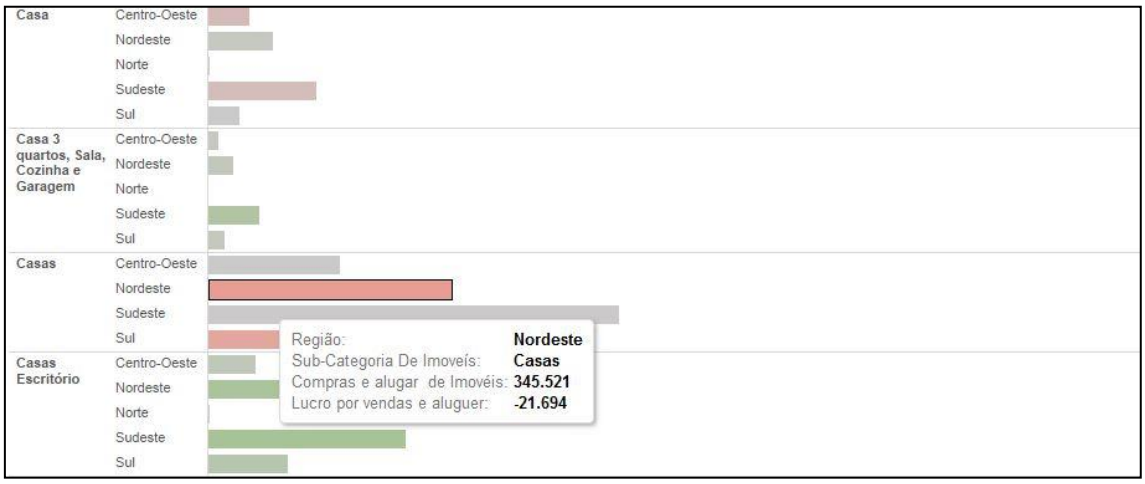


Fonte: Do autor.

Observa-se que na figura 16, temos as subcategorias de imóveis cadastros, e é possível ver quais tipos de imóveis tiveram maior procura em todas as regiões. A subcategoria de imóveis apartamentos, é um tipo de imóveis que tiveram maior número de procura em quase todas as cidades, pode-se ver que, o imóvel apartamento é uma subcategoria que mais teve vendas e locações.

Já a subcategoria de imóveis casa, é uma subcategoria que teve menor procura vendas e locações quase todas as regiões. Pode-se observar na figura abaixo que os lucros por vendas e aluguel de casa teve saldo negativo.

Figura – 16subcategorias casa



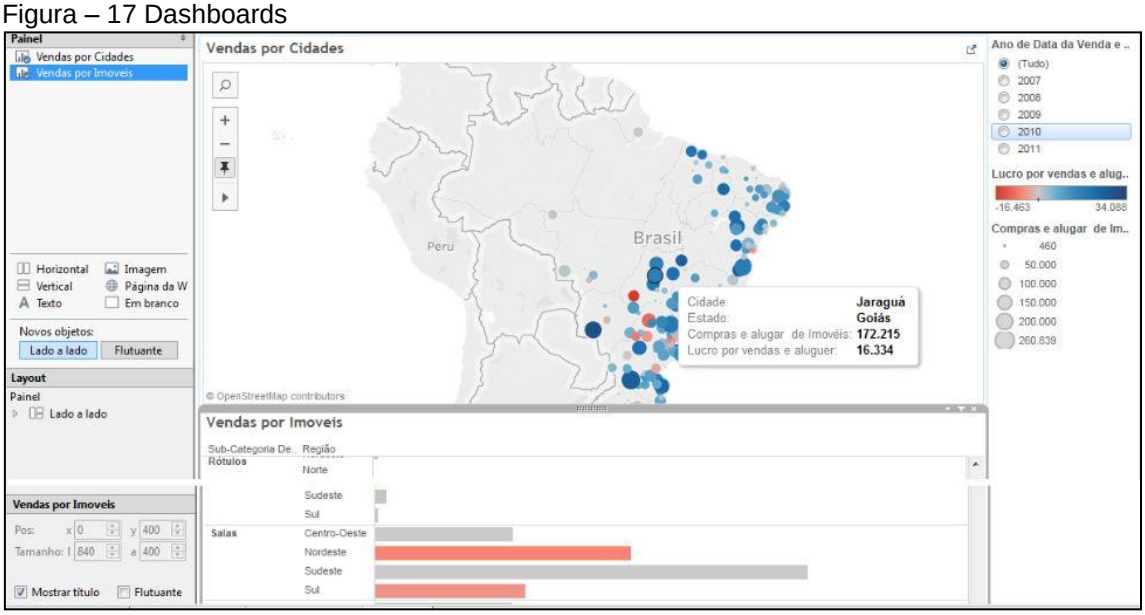
Fonte: Do autor.

10.5 RESULTADOS OBTIDOS

Finalizando a extração da informação pretendida, por meio da visualização de dados no Tableau, criou-se duas planilhas de visualização, a primeira planilha denominada vendas por cidades e a outras vendas por subcategorias de imóveis.

Em seguida criamos um painel para unir as duas planilhas, a planilha de vendas por cidades onde é possível visualizar as informações dos lucros e prejuízos em todas as cidades, e a planilha de subcategorias de imóveis aonde estão as informações dos tipos de imóveis que tiveram maior procura em todas as regiões. Unindo essas duas planilhas no painel do Tableau é possível visualizar todas as informações em uma única vez, ou seja, os painéis que também são conhecidos como *dashboards* são formados por conjuntos de visualizações de todas as informações que estão presentes nas planilhas, os painéis são uma sequência ou conjuntos de planilhas que trabalham juntos para transmitir informações por meio de visualizações. Para se criar um *dashboards* no Tableau é preciso primeiramente ter mais de uma planilha criada, e a união dessas planilhas gerar um *dashboards*.

Figura – 17 Dashboards



Fonte: Do autor

Como *dashboards* criado, é possível ver ao mesmo tempo as informações das duas planilhas criadas, pode-se visualizar as informações dos lucros, vendas e alocações ao mesmo tempo também visualizar as subcategorias de imóveis procurados por regiões. No canto superior direito da figura observa-se que na medida em que selecionamos o ano que se quer visualizar as informações, as das duas planilhas vão se atualizando em virtude do ano selecionado. Tanto a planilha de vendas por subcategorias de imóveis e a planilha da vendas e lucros por cidades automaticamente vai mostrando a informação do ano unificado.

Criado o *dashboards*, é possível salva-lo no servidor e visualizar todas as informações na web, gerenciar, atualizar, visualizar, e prever quais tipos de imóveis podem ser procurados.

A extração do conhecimento por meio da utilização de visualização de dados possibilita descobrir novas correlações, padrões e tendências entre as informações de uma empresa, contidas em uma grande base de dados. Esse processo de visualização de dados no Tableau oferece um diferencial para a organização, pois em tempos de concorrência, ou na exigência da qualidade por parte dos clientes é necessário ter uma ferramenta boa para o auxílio no processo de tomada de decisão.

Através da mineração e visualização dos dados, é possível extrair conhecimento que estão ocultos em um banco de dados, podendo assim, utilizar esses dados como estratégias no negócio.

Foram realizados vários testes de buscas de informações na base de dados criada sem o auxílio do software Tableau, fez-se uma comparação do uso da base de dados verificou-se que há uma perda enorme de tempo na busca dos dados existentes na base dados sem a ferramenta Tableau, a principal característica é a extração dos dados de maneira não-trivial de informações a partir de uma base de dados de grande porte. Informações que são necessariamente implícitas previamente desconhecidas, mais com um potencial util.

Atualmente várias organizações perdem negócios ou tomam um rumo errado por ter tomado decisões erradas, ou seja, decisões que não estão fundamentadas em dados reais obtidos de forma segura. Neste contexto o conhecimento extraído de uma base de dados de forma segura pode fornecer informações importantes para a tomada de decisão.

O diferencial de uma empresa para outra caracteriza-se pela sua capacidade de reagir ás mudanças em um menor período de tempo, para isso uma ou mais decisões são necessárias, e aquelas organizações que possuem a informação certa no momento certo tendo a obter sucesso em seus processos decisórios. A finalidade do uso do software Tableau é de modelar os dados para obter informações que diferencie as ações das empresas e essas consigam se destacar perante à concorrência, desta forma pode-se inferir que a ferramenta Tableau é um diferencial estratégico frente ao mercado cada vez mais aquecido e incerto.

Como contribuição o gestor envolvido neste trabalho, bem como os demais gestores, poderá aproveitar os resultados obtidos para auxiliar a sua reflexão a respeito do uso e do tratamento da informação no seu processo de decisão, minimizando assim, os riscos incorridos em sua tomada de decisão equivocada

11. CONCLUSÃO

Cada dia, um grande volume de dados tem sido gerado por diferentes fontes, como sistemas e dispositivos. Estes dados chamados de Big Data, precisam ser armazenados e processados e podem ser analisados para se ter uma tendência das informações.

O Big Data tornou-se uma ferramenta importante como um papel fundamental na gestão da informação dentro das organizações. A manipulação desses dados e a análise das informações de maneira tradicional tornou-se inviável devido ao grande volume de dados.

Contudo, realizar estas tarefas não é simples, principalmente devido ao volume, velocidade e variedade, visto que, a análise para o Big Data, os sistemas de gerenciamento de banco de dados tradicionais não são adequados para lidar com enorme quantidade de dados.

Este trabalho apresentou aspectos de dados de Big Data, destacando a análise de dados não estruturados. Foi apresentado a infraestrutura do software Tableau para análise de dados não estruturados.

Por fim, foram apresentados alguns resultados importantes para dados não estruturados, tais como processamento, armazenamento, análise, gerenciamento e posteriormente a visualização dados. Apesar de existir algumas ferramentas que atenuam estes desafios, ainda existe muito a fazer, principalmente porque o volume de dados e a velocidade com que estes dados são gerados de forma exponencial. Estes desafios geram oportunidades de pesquisa, de forma que a análise para o Big Data possa ser utilizada pelas empresas, melhorando seus produtos, serviços e a qualidade de vida dos seus clientes.

Durante a realização da pesquisa encontrou-se dificuldades, dentre as quais destaca-se a etapa do tratamento dos dados e a visualização destes dados no software Tableau, visto que, foi usado o Tableau livre e alguns componentes importantes da ferramenta não nos foram liberados por ser pagos e com um custo caro.

Outra dificuldade encontrada, foi na definição do nosso problema no contexto do Big Data visto que, é algo bastante genérico desta forma estabelecer um escopo para abordar o nosso estudo.

Desta forma, pode-se concluir que os objetivos desta pesquisa foram alcançados, uma vez que os estudos forma realizados, e a ferramenta usada correspondeu corretamente no desenvolvimento do trabalho proposto.

Dos resultados obtidos, podem-se sugerir alguns trabalhos futuros como:

a) Adicionar mais funcionalidades na ferramenta para permitir a adição de vários outros métodos e obter dados mais precisos.
b) Em virtude da relevância do tema e do Big Data ser uma realidade, é importante a elaboração de novos estudos de caso. Dessa forma, pode-se avaliar como os gestores de outras organizações são influenciados pelas mesmas variáveis. Assim, será possível identificar um padrão de comportamento geral de efeito no processo decisório.
c) Um outro trabalho futuro seria a descrição completa de um caso de aplicação do Big Data e sua contribuição nos resultados de uma empresa em relação aos seus concorrentes.

REFERÊNCIAS

Abiteboul, S., Buneman, P. and Suciu, D**., Data on the Web**, 1st. Ed. Morgan Kaufmann Publishers, 2008.

ALBERTO FILHO, Luiz. **Aplicação de Técnicas de Mineração de Dados e Textos no Apoio à Tomada de Decisão em Segurança Pública**. 2009. 79 f. Dissertação (Pós-graduação) -Universidade Federal Do Pará, Belém, 2009.

ABREU, Mauricio de (ORG); FRIDMAN, Fania. Cidades Latino Americanas: um debate sobre a formação de núcleos urbanos. Rio de Janeiro, Casa da Palavra, 2010.

AMORIM, S.M.F. Mineração da bolsa de valores: um estudo exploratório. Dissertação de Mestrado. ISPA. Lisboa- Portugal, 2006.

Business Review, edição de outubro de 2012.
FIGUEIRA, Rafael. **Mineração de dados e bancos de dados orientados a objetos.** Rio de Janeiro: UFRJ, 1998 (Dissertação, Mestrado em Ciência da Computação).

Boyd, Danah; Crawford, Kate. (2013) **Six provocations for Big Data**. In: Symposiumon the Dynamics of the Internet and Society. Oxford Internet Institute.

Battenberg Nicolas (2014) **Mining Unstructured Data From Code Reviews**: A Hands.-On Tutorial.

Breitman K. (2012) **Big Data é indispensável ao pré-sal**. Vcongress

Buneman, P. (1997) “Semi-structured data”, In Proc. of ACM Symp. On Principles of Database Systems.

Deutsch, A., Fernandez, M., Florescu, D., Levy, A. and Suciu, D. (1999) “A query language for xml”, In International World Wide Web Conference.

DRUCKER, Peter F**. Administração: tarefas, responsabilidades, práticas**. V.1. São Paulo: Pioneira, 2002.

Date, C. J**. Introdução a Sistemas de Banco de Dados**. Rio de Janeiro: Elsevier Editora, 2004.

ELMASRI, Ramez; NAVATHE, Shamkant B. **SISTEMAS DE BANCO DE DADOS**.6. ed. São Paulo: Addison Wesley, 2011.
FAYOL, Henri. Administração industrial e geral: previsão, organização, comando, coordenação, controle. 10. Ed. São Paulo: Atlas, 1994.

Gupta, A. and Mumick, I. S. (1996) "**What is the data warehousing problem? (Are materialized views the answer?**)", In Proc. of the 22rd VLDB Conference

GOLDSCHMIDT Ronaldo, PASSOS Emmanuel. Data Mining: um guia prático. Rio de Janeiro: Campus, 2005.

Henry, Mintzberg, (2011). Criando organizações eficazes. São Paulo: Atlas.

ISACA®, Big Data – **Impactos e Benefícios**. 2013. Disponível em: http://www.isaca.org/KnowledgeCenter/Research/Documents/BigData_whp_Por_0413.pdf. Acesso em: 19 maio. 2015.
ISACA®, Privacidade & Big Data. 2013. Disponível em: http://www.isaca.org/Knowledge-Center/Research/Documents/Privacy-and-BigData_whp_Portuguese_0913.pdf. Acesso em 17 maio.2015.

ISACA (2013) **Big Data: Impactos e Benefícios**, ISACA. Em: http://www.isaca.org/Knowledge-

IBM (2011) **what is Big Data?** Bringing Big Data to enterprise, www.ibm.com/software/data/bigdata, recuperado em13 novembro de 2015.

MCAFEE, A; BRYNJOLFSSON, E. Big Data: **The Management Revolution**. Harvard.

MacAfee, A &Brynjolfsson, E (2012). Big Data: **The Management Revolution**. *Harvard*
Business Review, edição de outubro de 2012.

MANYIKA, J. e CHUI, M. e BROWN, B., et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. Disponível em: acesso em 13/03/16.

MATTOS, P., A tecnologia -. 10 eds. - São Paulo - ABDA, 2011.

MACHADO, Aydano P. Mineração de texto em redes sociais aplicada à educação a distância. Colabora – Revista digital da CVA – Ricesu, vol. 6, n. 23, jul, 2010.

MOXON, Bruce. **Defining Data Mining. DBMS**, Data Warehouse Supplement, August 1991.

Petry, A. Vida Digital: **O Berço do Big Data.** Revista Veja, São Paulo, Maio. P.71-81,**2013.**

PEREIRA, JULIO Cesar R. **Analise de Dados Qualitativos: Estratégias Metodologias para as Ciências da Saúde, Humanas e Sócias**. 3 eds. 1 reimpr. São Paulo: Editora da Universidade de São Paulo. 2014.

Pan, L., Chen, L., Hui, J., & Wu, J. (july de 2011). **QoS-Based Distributed Service**
Selection in Large-Scale Web Services. Services Computing (SCC), 2011 IEEEInternational Conference on, (pp. 725-726)

Rodrigo Carro (2014) **Big Data transforma Saúde e Educação**. Disponível em: http://brasileconomico.ig.com.br/negocios/2014-05-27/big-data-transforma-saude-e-educacao.html Acesso Maio.19. 2015.

PALADINI, Edson. Pacheco. Gestão da Qualidade, Teoria e Prática. São Paulo, Atlas, 2009.

Rose Business Technologies. Disponível em: http://www.rosebt.com/1/archives/07- 2012/1.html. Acesso em: 21 maio de 2015.

Schneider, R. D. **Hadoop For Dummies, Special Edition**. Mississauga, CAN: John Wiley & Sons Canada, 2012. 41 p.

Silberschatz, A. et al. **Sistema de Banco de Dados**. 5. ed. São Paulo: Elsevier, 2006.

TANKARD, Colin**, Big data security, Network security**, Jul. 2012.

TAURION, Cezar. Você realmente sabe o que é o big data? Blog da IBM, 30 abril2012.Disponivel:https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/voce_realmente_sabe_o_que_e_big_data?lang=en. Acesso em 12 de jun. 2015.

TAYLOR, Frederick Wislow. Princípios de administração científica. 8 ed. São Paulo: Atlas, 1990.
WITTEN, Ian H.; FRANK, Eibe. **Data Mining**: Practical Machine Learning Tools and

ROSSOUW, L. Big Data – **Grandes oportunidades.** Gen Re – Risk Insights, vol. 16, nº 2, 2012.

Ramakrishnan, R.; Gehrke, J. **Sistemas de gerenciamento de banco de dados**. São Paulo: McGraw-Hill, 2008.

VIEIRA, M. F. Gerenciamento de projetos de tecnologia da informação. 2.ed. Rio de Janeiro: Campus, 2007.

APÊNDICE A – ARTIGO

Metodologia para mineração de dados não estruturados no software Tableau (estudo de caso) com big data

Álvaro Ambriz Domingos[1], Paracelso de Oliveira Caldas[2]

[1]Acadêmico do Curso de Ciência da Computação

Universidade do Extremo Sul Catarinense (UNESC) – Criciúma, SC – Brasil

[2]Professor do Curso de Ciência da Computação

Universidade do Extremo Sul Catarinense (UNESC) – Criciúma, SC – Brasil

Abstract *The storage of information in digital format has grown a lot in recent years. Not only people who are responsible for producing data, industrial, operational, and electronic equipment have also become great generators of records such as social networks, servers, smartphone devices, microcomputers scattered in the most varied contexts among an affinity of applications. It was noticed that most of this data is data not structured, and data that cannot be parsed and stored by the traditional database method. The term Big Data refers to all this amount of data that is in the order of structured, semi-structured and unstructured data. The objective of this articles was to carry out a case study and a process using Tableau software for unstructured data mining with Big Data, extracting the information through visualization intermediates*

Resumo. *O armazenamento das informações em formato digital cresceu muito nos últimos anos. Não apenas pessoas que são responsáveis por produzir dados, os equipamentos industrias, operacionais, electrónicos também se tornaram grandes geradores de registros como exemplos as redes sócias, servidores, aparelhos de smartphones, microcomputadores espalhados nos mais variados contextos entre uma afinidade de aplicações. Percebeu-se que maior parte desses dados são dados não estruturado, ou seja, dados que não podem ser analisados e armazenado pelo método de banco de dados tradicionais. O termo Big Data se refere a toda esta quantidade de dados que se encontram na ordem dos dados estruturados, semiestruturados e não estruturados. Este artigo teve como o objetivo realizar um estudo de caso, e um processo usando o software Tableau para mineração de dados não estruturados com Big Data, extraindo as informações por intermédios de visualizações.*

1. Introdução

Atualmente, vivemos por uma era que tem sido marcada por enormes quantidades de dados. Dados esses que tem sido gerado por instituições como empresas, universidades, hospitais, a internet entre outros.

Muitos desses dados são dados não estruturados. Entende-se por dados não estruturados aqueles que são provenientes de documento do formato doc, docx, pdf, XML, HTML, redes sócias e outras infinidades de formatos.

Desta forma, são extremamente relevantes as atenções e os estudos desses tipos de dados visto que eles têm uma representatividade enorme além de terem significados importantes nas decisões que envolvem a sociedade em sua totalidade.

Este artigo teve como objetivo criar um processo para mineração de dados não estruturados no software Tableau, extraindo as informações de uma base de dados por meios de visualizações.

2. Banco de dados

Os sistemas de banco de dados surgiram como opção aos métodos antigos no gerenciamento de dados em computadores. Visto que nos sistemas antigos os dados eram armazenados em arquivos do sistema operacional, que era responsável em criar ambiente para o acesso, manipulação e controle dos arquivos. Para cada interação do sistema operacional entre o sistema e o arquivo era necessário que um programa específico do sistema operacional fosse executado que fornecia aquela função (SILBERSCHATZ, 2006).

A expressão banco de dados é um termo que originou do inglês Data Banks, que eventualmente foi trocado pela palavra Data bases- Bases de dados por possuir significado, mas apropriado (SILVA, 2005).

Um banco de dados é um sistema de armazenamento de dados, que faz interação com os usuários podendo permitir diversas execuções como: inserir registros, consultar, excluir, editar, atualizar e várias outras. Ou seja, é um sistema de coleção de dados organizados, integrados, que formam uma representação de dados que pode ser utilizada por todas as aplicações relevantes sem duplicação de dados.

3. Dados não estruturados

Nos últimos anos vem crescendo o aumento no volume dos dados digitais, criados e armazenados quer seja por pessoas particulares, por empresas que são a maior fonte de dados. A história mostra-nos que as empresas sempre utilizaram os sistemas de computação e os bancos de dados para armazenar maior partes dos dados corporativos em formatos estruturado, como tabelas relacionais, arquivos de formato fixo.

Porém, atualmente grandes partes dos dados são armazenados em documentos criados por ferramentas como: Office Excel, Microsoft Word. E com o avanço da digitalização de documentos, produção de vídeos, formatos de áudio, o conjunto de dados não estruturados cresceu (MICROSOFT, 2011).

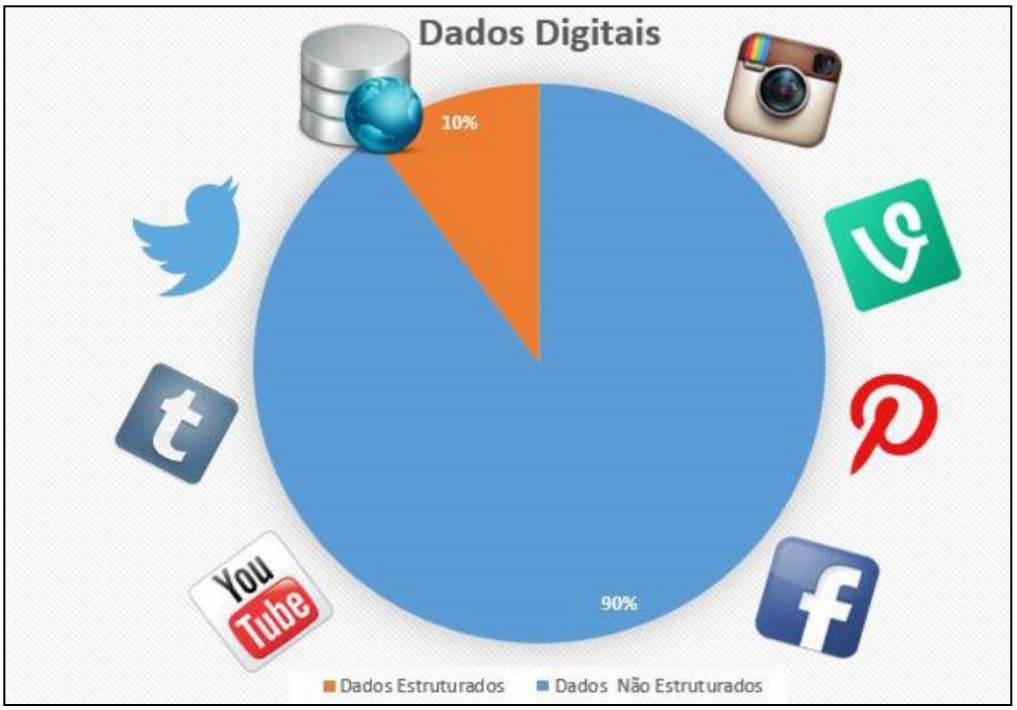
De acordo com a pesquisa feita a estimativas, de 90% dos dados armazenados são mantidos fora do banco de dados relacionais (GUIA DAMA, 2012).

Da maioria dos dados gerados nos últimos anos, apenas 10% destes mesmo dados são dados estruturados. Os outros 90% são dados não estruturados e geralmente se

encontram na sua grande parte nas redes sócias como exemplo do Twitter, Instagram, Facebook e várias outras (IBM, 2011).

A figura abaixo mostra detalhadamente o percentual dos dados não estruturados em relação a outros formatos de dados.

Figura 2 - Dados não estruturados



Fonte: Taurion (2015) ADPTADO.

Segundo Navathe (2005), dados não estruturados são formas para descrever os dados que não estão contidos numa base de dados ou qualquer um outro tipo de estrutura de dados. E geralmente os dados não estruturados podem ser textuais ou não textuais, textuais como e-mails, apresentações como o Power Point, documentos de Word, e dados não estruturados não textuais como imagens JPEG, arquivos de áudio MP3 e arquivos de vídeos em flash, são dados que podem ter qualquer tipo de formato ou sequência, que não seguem nenhuma regra e não são previsíveis.

Os dados não estruturados são conjuntos de dados que seguem uma forma não ordenada de itens como páginas, planilhas, tabela de banco de dados ou outros conjuntos de dados lineares.

O termo conjunto de dados não estruturados está associado aos dados que geralmente não inclui um modelo de dados pré-definido, e que não combina com tabelas relacionais, e geralmente são pesados textos que podem incluir números, datas, fotos, e que se tornam difíceis de serem identificados utilizando bancos de dados convencionais (BETTENBURG, 2014).

Um dos maiores desafios com os dados não estruturados é que eles sejam contextualizados para serem entendidos e utilizados. Alguns pesquisadores afirmam que pesquisar e analisar dados não estruturados são como procurar uma agulha num palheiro (BACCHELLI, 2013).

Com o volume dos dados não estruturado aumentando tão rapidamente, muitas empresas recorrem as soluções tecnológicas para ajudá-los a gerenciar, melhorar e armazená-los.

As empresas acreditam que seus armazenamentos de dados não estruturados envolvem informações que podem ajudá-los a tomar melhores decisões dos seus negócios e projetos (GUERROUJ, 2012).

Para o auxílio do problema, as empresas voltam-se para diferentes soluções de software projetados para procurar dados não estruturados e também extrair informações importantes. E a principal vantagem do uso dessas ferramentas é a capacidade de recolher informações importantes que podem ajudar a empresa a ter sucesso

(IBM, 2011).

4. Big Data

A quantidade de informações que são geradas por diversas fontes como, por exemplo, Web, as redes de sensores e vários outros meios de comunicação, estão na ordem de centenas dados não estruturados. E com esta quantidade de dados não estruturados, novos grandes desafios surgem na forma de armazenamento, manipulação e no processamento de consultas em muitas áreas da computação, especialmente na área de bases de dados, mineração dos dados e recuperação das informações (VIEIRA, 2012).

O conceito do Big Data consiste em dados que excedem a capacidade do armazenamento de banco de dados de sistemas convencionais (DUMBILL, 2012).

Ela também pode ser definida como uma tecnologia que descreve quantidade de dados estruturados, dados semi-estruturados e não estruturados, que tem a capacidade de explorar esses dados para obter melhores informações para a empresa, ressaltando que há um desafio no que se refere a gestão e a análises desses dados (LOHR, 2012).

O Big Data pode ser caracterizado por 5Vs. Primeiro volume, o volume é a dimensão mais comum no que concernem os conceitos do Big Data, visto que o termo volume vem tendo maior atenção pela forma com que os dados vêm aumentando. Cada dia são gerados volumes de dados pela sociedade (TAURION,2012)

Primeiro: Big Data surgiu para tratar volumes extremos de dados estruturados, semi-estruturados e dados não estruturados. É difícil analisar esse volume de dados por sistemas que não tem o auxílio do Big Data, visto que é um desafio armazenar esse grande volume de dados com as ferramentas tradicionais (TANKARD, 2012).

Segundo: velocidade, Big Data, aproveita o conhecimento desenvolvido associado a evolução tecnológica, devolvendo o resultado das consultas extremamente

mais rápido em comparação os tradicionais, visto que é necessário que as transmissões das informações sejam feitas a uma velocidade bastante dinâmica, porque muitas das vezes precisamos agir em tempo real, exigindo um processamento que acompanhe esta velocidade (TAURION, 2012).

Terceiro: variedade, Callebaut (2012) afirma que quando se fala em Big Data, fala-se em diferentes tipos de dados que podem ser logs de servidores, de imagens, textos, documentos digitalizados, e outros conteúdos gerados pelas redes sócias. A existência desses dados estruturados, semi-estruturados e não estruturados são variedade de dados. Big Data opera esses variados formatos de dados com a precisão e a variedade de forma análoga ao tratamento estruturado.

Quarto: veracidade é preciso destacar o conteúdo mais necessário em meio a tanta informação, Big Data proporciona confiança no tratamento das informações garantindo o máximo de consistência possível, além de garantir a separação das informações mais importantes para compreender melhor os clientes e o seus comportamentos (IMB, 2011).

Quinto: valor, o uso do Big Data torna-se importante quando todos os esforços são direcionados para gerar valor, que é a informação da análise e sua conversão em material que serão aplicados nas decisões de uma certa empresa.

Big Data de certa forma é responsável por ganhos nas tomadas de decisões caracterizando assim o valor da informação obtida (TAURION, 2012).

A compreensão desses conceitos e das suas aplicações irá resultar em uma transformação no mundo dos negócios, conhecido como uma economia mais inteligente e analítica.

Segundo Boyd e Crawford (2011) saber lidar com o surgimento de uma era do Big Data é fundamental para questionar as incertezas e as mudanças rápidas no nosso processo decisório, pois, as decisões atuais terão considerável impacto no futuro, razões que levam as organizações aplicarem está tecnologia e porque vem apresentando resultados satisfatórios.

Aplicando os 5vs do Big Data em determinadas áreas permite maior eficiência na previsão de tendências do mercado, o conhecimento de atitudes para fidelizar com os seus clientes e melhorar os processos internos da empresa para atingir o sucesso. Visto que a ideia central do conceito Big Data é a tomada de decisão em tempo real sobre uma gama de dados (TANKARD,2012).

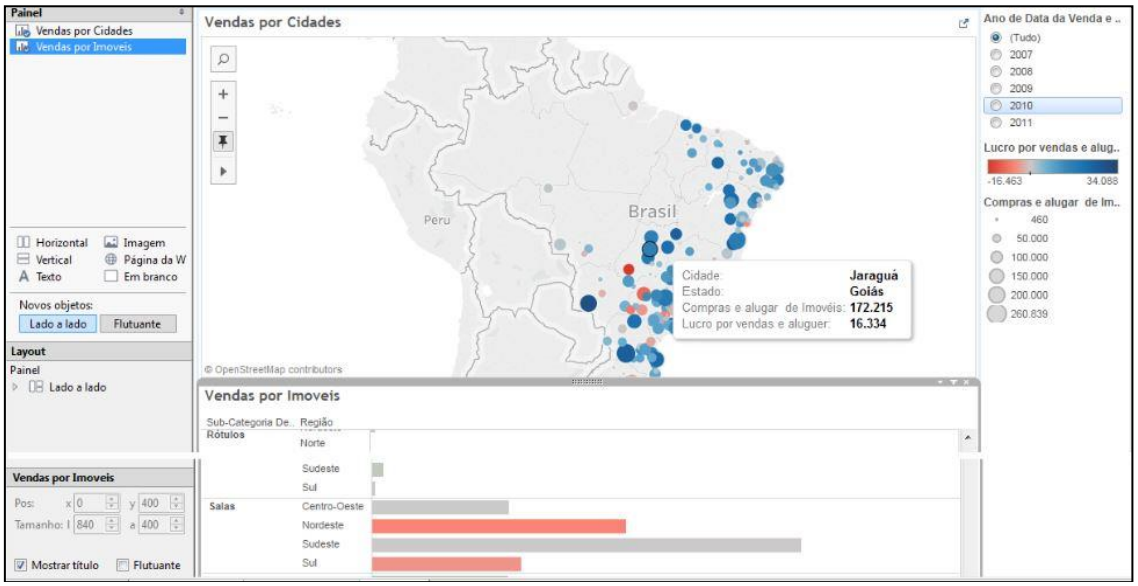
5. Trabalho proposto

O presente trabalho consistiu em realizar uma metodologia para mineração de dados em bases de dados não estruturados no software Tableau. Neste projeto foram analisadas técnicas de mineração de dados, e também foi feito um estudo da ferramenta Tableau usadas para visualização e recuperação dos dados. Após estudos e analisadas as técnicas de mineração de dados foi selecionada uma técnica melhor que se adequa com a implementação deste estudo. Também foi desenvolvido um estudo de caso, na empresa Duda Imóveis para a coletas de dados e informações afim de se ter uma ideia de como é estruturado um banco de dados de uma empresa do setor imobiliário e quais

informações ela contém. Após o desenvolvimento do estudo criou-se uma base de dados para a demonstração da mineração em uma base de dados não estruturados com o auxílio do software Tableau para a visualização dos dados.

5.1 Resultados Obtidos

A extração do conhecimento por meio da utilização de visualização de dados possibilita descobrir novas correlações, padrões e tendências entre as informações de uma empresa, contidas em uma grande base de dados. Esse processo de visualização de dados no Tableau oferece um diferencial para a organização, pois em tempos de concorrência, ou na exigência da qualidade por parte dos clientes é necessário ter uma ferramenta boa para o auxílio no processo de tomada de decisão.



Como dashboard*s* criado, é possível ver ao mesmo tempo as informações das duas planilhas criadas, pode-se visualizar as informações dos lucros, vendas e alocações ao mesmo tempo também visualizar as subcategorias de imóveis procurados por regiões. No canto superior direito da figura observa-se que na medida em que selecionamos o ano que se quer visualizar as informações, as das duas planilhas vão se atualizando em virtude do ano selecionado. Tanto a planilha de vendas por subcategorias de imóveis e a planilha da vendas e lucros por cidades automaticamente vai mostrando a informação do ano unificado.

6 Conclusão

O Big Data tornou-se uma ferramenta importante como um papel fundamental na gestão da informação dentro das organizações. A manipulação desses dados e a análise das informações de maneira tradicional tornou-se inviável devido ao grande volume de dados.

Contudo, realizar estas tarefas não é simples, principalmente devido ao volume, velocidade e variedade, visto que, a análise para o Big Data, os sistemas de gerenciamento de banco de dados tradicionais não são adequados para lidar com enorme quantidade de dados.

Este trabalho apresentou aspectos de dados de Big Data, destacando a análise de dados não estruturados. Foi apresentado a infraestrutura do software Tableau para análise de dados não estruturados.

Por fim, foram apresentados alguns resultados importantes para dados não estruturados, tais como processamento, armazenamento, análise, gerenciamento e posteriormente a visualização dados. Apesar de existir algumas ferramentas que atenuam estes desafios, ainda existe muito a fazer, principalmente porque o volume de dados e a velocidade com que estes dados são gerados de forma exponencial. Estes desafios geram oportunidades de pesquisa, de forma que a análise para o Big Data possa ser utilizada pelas empresas, melhorando seus produtos, serviços e a qualidade de vida dos seus clientes.

Referencias

Abiteboul, S., Buneman, P. and Suciu, D**., Data on the Web**, 1st. Ed. Morgan Kaufmann Publishers, 2008.

ALBERTO FILHO, Luiz. **Aplicação de Técnicas de Mineração de Dados e Textos no Apoio à Tomada de Decisão em Segurança Pública**. 2009. 79 f. Dissertação (Pós-graduação) -Universidade Federal Do Pará, Belém, 2009.

ABREU, Mauricio de (ORG); FRIDMAN, Fania. Cidades Latino Americanas: um debate sobre a formação de núcleos urbanos. Rio de Janeiro, Casa da Palavra, 2010.

MacAfee, A &Brynjolfsson, E (2012). Big Data: **The Management Revolution**. *Harvard*

Business Review, edição de outubro de 2012

Ramakrishnan, R.; Gehrke, J. **Sistemas de gerenciamento de banco de dados**. São Paulo: McGraw-Hill, 2008.

TANKARD, Colin, Big data security, Network security, Jul. 2012.

Silberschatz, A. et al. **Sistema de Banco de Dados**. 5. ed. São Paulo: Elsevier, 2006.

VIEIRA, M. F. Gerenciamento de projetos de tecnologia da informação. 2.ed. Rio de Janeiro: Campus, 2007.

www.dudaimoveis.com.br
contato@dudaimoveis.com.br

DUDA IMÓVEIS
Mais garantia, maior retorno.

CARTA DE AUTORIZAÇÃO

Eu Sander Topanotti Fortuna, Gerente de T.I da empresa Duda Imoveis, tenho ciência e autorizo a realização da pesquisa intitulada METODOLOGIA PARA MINERAÇÃO DE DADOS NÃO ESTRUTURADOS NO SOFTWARE TABLEAU (ESTUDO DE CASO) COM BIG DATA sob responsabilidade do pesquisador Álvaro Ambriz Domingos Para isto, serão disponibilizados ao pesquisador Alvaro Ambriz Domingos o ambiente físico e tecnológicos da empresa, a documentação e os dados necessários para a pesquisa.

(Criciúma), (15/08/2016).

Sander Topanotti Fortuna

(nome completo do responsável e cargo ocupado no local onde a pesquisa será realizada)