**NÍCOLAS HENRIQUE VIEIRA TOLEDO**

**SEBASTIÃO BATISTA DE ANDRADE NETO**

**ESTATÍSTICAS DE REDES SOCIAIS**

**COM BIG DATA**

**UNIVERSIDADE DO VALE DO SAPUCAI**

**POUSO ALEGRE**

**2015**

**NÍCOLAS HENRIQUE VIEIRA TOLEDO**

**SEBASTIÃO BATISTA DE ANDRADE NETO**

**ESTATÍSTICAS DE REDES SOCIAIS**

**COM BIG DATA**

Trabalho de Conclusão de Curso Sistemas de Informação da Universidade do Vale do Sapucaí como requisito para obtenção de título de bacharel em SISTEMAS DE INFORMAÇÃO.

Orientador: Prof. Ednardo David Segura

**UNIVERSIDADE DO VALE DO SAPUCAI**

**POUSO ALEGRE**

**2015**

**LISTA DE FIGURAS**

Figura 1 - Os 3 v’s que definem o *Big Data* ........................................................................... 18

Figura 2 - *Data Mining* como um passo no processo da descoberta do conhecimento .......... 19

Figura 3 - O ecossistema *JavaScript* ....................................................................................... 28

Figura 4 - Diagrama demonstrando a fronteira da *Amazon Web Services e* as tecnologias empregadas neste Projeto ........................................................................................................ 31

**LISTA DE TABELAS**

Tabela 1 - Cronograma do Primeiro Semestre de 2015 .......................................................... 39

Tabela 2 - Cronograma do Segundo Semestre de 2015 .......................................................... 40

Tabela 3 - Orçamento do Projeto ............................................................................................ 41

**LISTA DE ABREVIATURAS E SILGAS**

ASCII - *American Standard Code for Information Interchange*

API - Application Programming Interface

CERN - *European Organization for Nuclear Research*

CRUD - *Create, Read, Update and Delete*.

CSS - *Cascading Style Sheets*

EUA - Estados Unidos da América

FAI - Faculdade de Administração e Informática

HP - Hewlett-Packard

HTML - *Hyper Text Markup Language*

HTTP - *Hyper Text Transfer Protocol*

IaaS - *Infraestructure as a Service*

IBM - *International Business Machines*

INATEL - Instituto Nacional de Telecomunicações

INPETECC - Instituto de Pesquisa e Treinamento em Tecnologia

I/O - *Input/Output*

JAR - Java *Archive*

JSON - *JavaScript Object Notation*

JVM - Java *Virtual Machine*

MG - Minas Gerais

NoSQL - *Not Only* SQL

PaaS - *Plataform as a Service*

RAM - *Random Access Memory*

REST - *Representational State Transfer*

SaaS - *Software as a Service*

SP - São Paulo

SQL - Structured Query Language

TI - Tecnologia da Informação

UNIFEI - Universidade Federal de Itajubá

UNIVÁS - Universidade do Vale do Sapucaí

URI - *Uniform Resource Indentifier*

W3C - *World Wide Web Consortium*

WWW - *World Wide Web*

XHTML - *Extensible Hyper Text Markup Language*

XML - *Extensible Markup Linguage*

**SUMÁRIO**

[1 INTRODUÇÃO 9](#_Toc422268904)

[2 OBJETIVOS 12](#_Toc422268905)

[2.1 Objetivo Geral 12](#_Toc422268906)

[2.2 Objetivos Específicos 12](#_Toc422268907)

[3 JUSTIFICATIVA 13](#_Toc422268908)

[4 QUADRO TEÓRICO 14](#_Toc422268909)

[*4.1* *Big Data* 14](#_Toc422268910)

[*4.2* *Data Mining* 18](#_Toc422268911)

[*4.2.1* *Data Mining* em Redes Sociais 20](#_Toc422268912)

[*4.3* REST API 21](#_Toc422268913)

[4.3.1 Twitter4j 22](#_Toc422268914)

[4.3.2 Facebook4j 22](#_Toc422268915)

[*4.4* *Storm* 23](#_Toc422268916)

[4.5 Banco de dados NoSQL 24](#_Toc422268917)

[4.5.1 MongoDB 26](#_Toc422268918)

[4.6 Node.JS 27](#_Toc422268919)

[*4.6.1* *Express* 29](#_Toc422268920)

[*4.7* *Cloud Computing* 30](#_Toc422268921)

[*4.7.1* *Amazon Web Services* 31](#_Toc422268922)

[*4.8* *Open Web Plataform* 32](#_Toc422268923)

[4.8.1 HTML5 32](#_Toc422268924)

[4.8.2 CSS3 33](#_Toc422268925)

[*4.8.3* *JavaScript* 34](#_Toc422268926)

[*4.9* *Bootstrap* 34](#_Toc422268927)

[5 QUADRO METODOLÓGICO 35](#_Toc422268928)

[5.1 Tipo de Pesquisa 35](#_Toc422268929)

[5.2 Contexto de pesquisa 35](#_Toc422268930)

[5.3 Participantes 36](#_Toc422268931)

[5.4 Instrumentos 37](#_Toc422268932)

[5.5 Procedimentos 37](#_Toc422268933)

[5.6 Cronograma 38](#_Toc422268934)

[5.6.1 Primeiro Semestre de 2015 39](#_Toc422268935)

[5.6.2 Segundo Semestre de 2015 40](#_Toc422268936)

[5.6 Orçamento 41](#_Toc422268937)

[REFERÊNCIAS 42](#_Toc422268938)

# INTRODUÇÃO

Atualmente, um grande volume de dados é visto circulando por todo o globo diariamente. Dia após dia, mais dados são criados por usuários pelas mais diversas razões. Eles podem ser desde uma simples foto de viajem postada em uma rede social, até uma página de protesto governamental da web. As possibilidades são gigantes. O problema, é que a maioria desses dados não está armazenada em uma base de dados estruturada. Desse modo, eles são considerados dados não estruturados (ANDRIOLE, 2015).

Os dados não estruturados são provenientes de documentos nos mais diversos formatos, tais como textos, imagens, vídeos, páginas web, dentre outros formatos conhecidos de arquivo (ANDRIOLE, 2015). Porém, há um questionamento crucial a ser feito: Como armazenar de maneira confiável, escalável e segura uma quantidade de dados tão variada e que cresce de maneira tão rápida?

Com o surgimento deste problema, nasceu o conceito de *Big Data*. O *Big Data*, de maneira geral, consiste em um armazenamento de massas de dados não estruturados, de modo que estes sejam coletados, organizados, processados e apresentados de maneira mais rápida e segura possível (ALECRIM, 2015).

LaVelle et al (2010), afirmam que a tendência do *Big Data* está crescendo entre as organizações. As informações estratégicas são coletadas através de canais digitais não estruturados, como: redes sociais, aplicativos de *smartphones*, e tantos outros dispositivos emergentes baseados na internet. Devido a sua grande quantidade, estes dados devem ser armazenados em uma base de dados consistente e confiável.

O *Big Data* apresenta soluções realmente inovadoras, substituindo os tradicionais métodos de coleta e armazenamento de dados de todas as organizações que desejam obter um diferencial competitivo no mercado. Contudo, é necessário também mudar a cultura da organização para que o *Big Data* apresente o resultado esperado.

McAfee e Brynjolfsson (2012), concluíram que as ferramentas e filosofias relacionadas ao *Big Data* se disseminaram, mudando assim as ideias em longo prazo de gestores sobre a experiência, natureza, e a prática de gestão. Desse modo, líderes capacitados de todos os setores do mercado estarão usando o *Big Data* com a devida finalidade para qual ele foi concebido: uma revolução na gestão. Como em qualquer outra grande mudança de negócio, os desafios de tornar o *Big Data* operante em uma organização podem ser relativamente complexos em alguns casos, no entanto, é uma transição que é indispensável para os executivos nos dias atuais.

Devido ao grande número de usuários, conectados através dos mais diversos dispositivos, redes sociais como o *Facebook* e o *Twitter* geram quantidades enormes de dados por dia. Esses dados são de extrema importância, se forem analisados visando um diferencial competitivo. Desse modo, a coleta e análise dessa massa de dados se tornam indispensáveis (FRANÇA et al, 2014).

Nagarajan, Sheth e Velmurugan (2011), afirmam que o compartilhamento de opiniões entre população tornou-se uma atividade comum nos dias atuais. Isso se deve ao crescimento rápido na popularidade das redes sociais, bem como o e de aparelhos constantemente conectados à internet. Isto deu um acesso sem precedentes aos os dados de uma população, bem como capacidade de realizar análises dos dados sociais e o desenvolvimento aplicações socialmente inteligentes, seja no sentido de entrega de conteúdo direcionado, gestão de crises, organização de revoluções ou a promoção do desenvolvimento social nos países em desenvolvimento.

Analisando em um cenário nacional, mais precisamente no âmbito político, os presidenciáveis Dilma Rousseff e Aécio Neves, juntos, tiveram cerca de 1.585.369 de menções no *Facebook* e 2.715.438 de *tweets* nas eleições do ano de 2014 (SCUP IDEAS, 2015).

Geralmente, a população tende a se expressar com maior intensidade quando ocorrem eventos de grandes proporções nacionais, como uma eleição presidencial. A população também se expressa com certa intensidade no ano da posse oficial, onde as expectativas relacionadas ao candidato são criadas. Como as redes sociais são um meio de expressão de fácil acesso, quantidades enormes de opiniões, sugestões, críticas, e análises políticas são feitas diariamente (SILVA; FERREIRA JUNIOR, 2013).

Diante desta situação, a vantagem competitiva poderá ser alcançada através do modelo de análise de dados em redes sociais, resultando em estatísticas. Modelo esse, que necessita da aplicação do conceito de *Big Data*.

Com essa análise, os políticos poderão compreender de forma mais clara e objetiva as carências da população, encurtando assim, a sua relação com a população. Desse modo, pode-se dizer que as “duas faces” são beneficiadas: Os políticos obtêm a vantagem competitiva em relação aos demais candidatos, e a população se vê melhor atendida pelo poder público, ambos devido à exposição de dados feita pela análise do *Big Data*.

Desse modo, devido ao grande interesse gerado pelas tecnologias e conceitos que envolvem o *Big Data*, a presente pesquisa tem como objetivo geral o desenvolvimento de uma aplicação web que demonstre as opiniões dos usuários do *Twitter* relacionadas à política brasileira.

Para alcançar o objetivo geral desta pesquisa, foram colocados os seguintes objetivos específicos:

* Coletar da rede social *Twitter* dados relacionados à política brasileira com base em palavras-chave cadastradas pelos usuários da aplicação web;
* Processar os dados coletados através do *Apache* *Storm*, posteriormente salvando as informações importantes em uma base de dados;
* Apresentar as informações na aplicação web.

A proposta da pesquisa é desenvolver uma aplicação web que seja capaz de demonstrar para o usuário a opinião pública em relação ao cenário político brasileiro, por meio da robustez do *Big Data* e da grande quantidade de dados do *Twitter*. Em outras palavras, uma aplicação que seja eficiente o bastante para demonstrar para o usuário como está a real opinião pública em relação aos políticos brasileiros no *Twitter*.

Em âmbito social, a pesquisa se torna relevante na área política, pela qual serão geradas informações para análise. Uma linha de contribuição é mostrar aos políticos onde a população se torna mais carente, através da análise dos *tweets*. Outra linha é mostrar à população como está o desempenho de seus representantes perante a sociedade, através da exposição dos *tweets* pela aplicação.

No âmbito tecnológico, foi escolhida a rede social *Twitter* pelo fato de seus usuários gerarem altos volumes de dados diariamente. A análise correta destes dados pode trazer diferenciais competitivos para uma organização. O conceito de *Big Data* foi escolhido por ser um assunto de atual relevância e diferencial competitivo no mercado mundial. Combinando estes dois cenários, a contribuição será a integração entre as duas partes, servindo de “exemplo prático” para qualquer pessoa com o conhecimento tecnológico que queira fazer este tipo de integração.

Na perspectiva acadêmica, a pesquisa se torna relevante pois ainda não foi abordado no curso. Assim, a mesma pode ser arquivada na biblioteca a fim de servir de material de consulta para alunos do curso de sistemas de informação.