**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**

Pré-projeto de pesquisa apresentado à disciplina de TCC 1 do Curso de Sistemas de Informação como requisito parcial para obtenção de créditos.

**UNIVERSIDADE DO VALE DO SAPUCAI**

**POUSO ALEGRE**

**2015**

# INTRODUÇÃO

Atualmente, é visto um grande volume de dados circulando por todo o globo diariamente. Para mais fácil entendimento, a ele foi dado o nome de “*Big Data*”. Porém a questão é: “O que fazer com *Big Data*?”.

De fato, esse volume de dados ocupa uma grande demanda nos recursos de TI em uma organização. Porém os benefícios adquiridos podem dar uma severa vantagem competitiva em relação aos concorrentes.

Redes Sociais como o Facebook, por exemplo, podem ser um diferencial em determinadas áreas, devido à grande quantidade de dados gerados diariamente.

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, 05 de fevereiro de 2015).

De fato, a disputa à presidência estava acirrada. Os marqueteiros de ambos os lados perceberam a importância do volume de dados que eram gerados pelas redes sociais. Desse modo, eles atualizavam as informações dos candidatos quase que em tempo real, a fim de obter vantagens relacionadas aos eleitores.

Muitos usuários demonstram o contentamento com seus representantes através de redes sociais. De fato, delas são uma forma fácil de expressão. Geralmente, a população tende a se expressar quando ocorrem eventos de grandes proporções nacionais, como uma eleição presidencial. Com esse imenso volume de dados gerados pelos usuários, nasce o *Big Data*.

Essa grande quantidade de dados contém opiniões públicas de uma forma totalmente direta. O modelo de análise de dados mais indicado para esta ocasião é feito em forma de estatísticas, processando os dados presentes no *Big Data*.

Com essa análise de opinião pública, os políticos poderão compreender de forma mais clara e objetiva as carências da população, encurtando assim, a relação entre o povo e seus representantes.

## Objetivo Geral

Desenvolver uma aplicação web que mostre informações em tempo real de opiniões políticas dos usuários de redes sociais.

## Objetivos Específicos

Os nossos objetivos específicos tem como base:

* Coletar uma massa de dados relacionados a política das redes sociais através de palavras-chave;
* Processar os dados coletados;
* Apresentá-los em tempo real em uma aplicação web.

## Justificativas

No âmbito social, o projeto se torna relevante na área política, a qual serão geradas estatísticas de análise. Uma linha de contribuição é mostrar aos políticos onde a população se torna mais carente, através da análise dos dados. Outra linha é mostrar à população como está o desempenho de seus representantes perante a sociedade, através da exposição dos dados pela aplicação.

No âmbito tecnológico, foram escolhidas as redes sociais pelo fato de seus usuários gerarem altos volumes de dados diariamente. A análise correta destes dados pode trazer dados importantes 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 conhecimento tecnológico que queira fazer este tipo de integração.

Na perspectiva acadêmica, o tema se torna relevante pois ainda não foi abordado no curso. Assim, o projeto pode ser arquivado na biblioteca acadêmica e servir de material de consulta para alunos do curso de sistemas de informação.

# QUADRO TEÓRICO

No quadro teórico serão descritos os conceitos que se aplicam ao tema e as tecnologias que serão utilizadas no desenvolvimento do projeto.

## Big Data

De acordo com Alecrim (2013), a princípio, podemos definir o conceito de Big Data como sendo conjuntos de dados extremamente amplos e que, por este motivo, necessitam de ferramentas especialmente preparadas para lidar com grandes volumes, de forma que toda e qualquer informação nestes meios possa ser encontrada, analisada e aproveitada em tempo hábil.

De maneira mais simplista, a ideia também pode ser compreendida como a análise de uma massa de dados para a geração de resultados relevantes que, em volumes menores, dificilmente seriam alcançados.

## Banco de dados NoSQL

De acordo com a revista Java Magazine (2013), NoSQL é um movimento que promove soluções de armazenamento de dados não relacionais. Ele é composto por diversas ferramentas que, de forma particular e específica, resolvem problemas como tratamento de grande volume de dados, execução de consultas com baixa latência e modelos flexíveis de armazenamento de dados, como documentos XML ou JSON.

NoSQL não têm como objetivo substituir os bancos de dados relacionais, mas apenas propor algumas soluções que em determinados cenários são mais adequadas. Desta forma é possível trabalhar com tecnologias NoSQL e banco de dados relacionais dentro de uma mesma aplicação.

Os tipos de bancos de dados NoSQL que são mais usados são:

* Chave-Valor: Armazenam objetos indexados por chaves, e possibilitam sua busca a partir de sua respectiva chave.
* Documentos: Conjunto de documentos no formato JSON. Os documentos são tratados como objetos únicos, os quais possuem campos com os respectivos valores.
* Coluna: Formado por colunas que contem um conjunto de informações, semelhante a uma tabela.
* Grafos: Formado por dados distribuídos em forma de vértices e arestas, os quais possuem atributos tanto nas arestas quanto nos vértices.

Os bancos de dados NoSQL são implantados quando os bancos de dados relacionais já não estão suportando a demanda de dados, ou seja, estão apresentando lentidão em consultas e processamentos. Um NoSQL pode substituir um banco de dados relacional por completo, ou simplesmente, ser implantado como um “suporte”, realizando as consultas/processamentos em um grande volume de dados onde a velocidade é necessária. O tipo de NoSQL a ser é escolhido após uma análise crítica do gestor de TI da organização, dependendo da necessidade da organização.

## Hadoop

Segundo Avoyan Hovhannes (2013), o Hadoop é uma implementação de código aberto do paradigma de programação Map-Reduce. Map-Reduce é um paradigma de programação introduzido pelo Google para processar e analisar grandes conjuntos de dados. Todos esses programas que são desenvolvidos nesse paradigma realizam o processamento paralelo de conjuntos de dados e podem, portanto, ser executados em servidores sem muito esforço. A razão para a escalabilidade desse paradigma é a natureza intrinsecamente distribuída do funcionamento da solução. Uma grande tarefa é dividida em várias tarefas pequenas que são então executadas em paralelo em máquinas diferentes e então combinadas para chegar à solução da tarefa maior que deu início a tudo.

## Implantação das Tecnologias

Segundo Marcos Pichatelli (2013) O Big Data pode se tornar caro de processar e armazenar se implantado em bancos de dados tradicionais. Para resolver esse problema novas tecnologias usam soluções open source e plataformas de hardware de custo acessível para armazenar os dados de maneira mais eficiente, paralelizar trabalhos e entregar poder de processamento.

Ainda segundo Marcos Pichatelli (2013) alguns passos necessários para se conseguir o pleno potencial de Big Data:

* **Coletar:** O dado é coletado das fontes de informação e distribuído por meio de múltiplos nós, por exemplo em um arquitetura grid, cada um dos quais processa um subconjunto de dados em paralelo.
* **Processar:** O sistema então usa o mesmo paralelismo gerenciado para ter um desempenho computacional mais rápido em cada nó. Depois, cada nó transforma os resultados das pesquisas em informações mais consumíveis para serem usadas tanto pelos seres humanos (em caso de análise) quando pelas máquinas (em caso de interpretação de resultados em larga escala).
* **Gerenciar:** Geralmente o processamento de Big Data é heterogêneo, originado a partir de diferentes sistemas transacionais. Quase todos os dados precisam ser entendidos, definidos, anotados, limpos e auditados para fins de segurança.
* **Medir:** As análises de negócios devem determinar uma métrica e devem ser acompanhadas constantemente. Geralmente as companhias medem o quanto um dado pode ser integrado/relacionado com um comportamento de consumo ou registro histórico; e como essa integração ou correção aumenta ou diminui com o tempo.
* **Consumir:** O resultado da análise dos dados deve atender a demanda original. Por exemplo, se o resultado for de algumas centenas de terabytes de interações em redes sociais, ele pode demonstrar como seus clientes compram produtos complementares. Então, deve haver regras de como os dados de mídias sociais são acessados e atualizados. O mesmo serve para o acesso de dados máquina-a-máquina (M2M).
* **Armazenar:** Como a tendência “data-as-a-service” ainda toma forma, cada vez mais os dados permanecem em um único lugar, enquanto os programas de acesso a essas informações se movem. Mesmo que os dados sejam armazenados para o curto prazo de processamento em lote ou para o longo prazo de retenção, as soluções de armazenamento devem ser deliberadamente dirigidas.
* **Governar:** A governança de dados engloba as políticas e fiscalização de informações por meio de uma perspectiva de negócios. Como definido, a governança de dados se aplica a cada um dos seis estágios de entrega de Big Data.

Realmente, o Big Data não é uma solução barata e de fácil implantação. É necessário haver um estudo de caso, para saber se a organização realmente precisa dessa tecnologia. Além do mais, é necessário ter profissionais especializados na área para ocorrer a implantação, e obter a vantagem competitiva desejada.

# QUADRO METODOLÓGICO

## Tipo de Pesquisa

A pesquisa objetiva a produção de novos conhecimentos por meio da utilização de procedimentos científicos. Contribui para o trato dos problemas e processos do dia a dia nas mais diversas atividades humanas, no ambiente de trabalho, nas ações comunitárias, no processo de formação e outros. O conhecimento torna-se uma premissa para o desenvolvimento no ser humano e a pesquisa como a consolidação da ciência. (SILVA, 2008).

Abrangendo este contexto, a metodologia de pesquisa deste projeto será de forma aplicada, pois a aplicação deverá fazer o processamento, análise e a demonstração de informações coletadas.

De acordo com Gil (2006), a pesquisa aplicada “objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Envolve verdades e interesses locais”.

A opção por essa metodologia foi feita por ela ser adequada para o desenvolvimento de uma aplicação que demonstre as vantagens e as funcionalidades das ferramentas que estão relacionadas com *Big Data*.

## Contexto de pesquisa

O contexto de pesquisa deste projeto tem como público alvo qualquer pessoa brasileira com o direito de voto que esteja disposta a acompanhar as atividades de seu representante. A pesquisa abrange também o próprio político brasileiro, caso haja o interesse do mesmo em acompanhar a sua avaliação perante a população.

Tomando como princípio software livre, toda parte relacionada ao desenvolvimento da aplicação será disponibilizada na internet. Assim, o acesso se torna livre e sem nenhum tipo de impedimento.

## Participantes

Nícolas Henrique Vieira Toledo, aluno do 7º (sétimo) período do curso de Sistemas de Informação da Universidade do Vale do Sapucaí (UNIVÁS). Possui formação técnica pela Microcamp de Bragança Paulista, estado de São Paulo. Possui conhecimentos em HTML, CSS, JavaScript, PHP, Java, C#, MySQL, PostgresSQL, NodeJS, ShellScript, Redes, Linux e inglês intermediário. Neste projeto, atuará como desenvolvedor.

Sebastião Batista de Andrade Neto, aluno do 7º (sétimo) período do curso de Sistemas de Informação da Universidade do Vale do Sapucaí (UNIVÁS). Possui formação técnica INPETECC de Pouso Alegre, Minas Gerais. Possui conhecimentos em HTML, CSS, JavaScript, PHP, Java, Python, PostgresSQL, ShellScript, Redes, Linux e inglês intermediário. Neste projeto, atuará como desenvolvedor.

Ednardo David Segura, Professor e Analista Web Developer. Pós Graduado em Engenharia Web pela Universidade Federal de Itajubá (UNIFEI), Bacharel em Ciência da Computação pela Faculdade de Administração e Informática (FAI) e Técnico em Informática pelo INPETTECC de Pouso Alegre, Minas Gerais. Neste projeto, atuará como orientador.

## Produção de Dados

Reuniões semanais com o orientador do projeto, com o objetivo de mostrar pesquisas e resultados, além da definição de foco do projeto até a reunião seguinte.

## Procedimentos

* Descrever o Big Data e suas ferramentas.
* Analisar o desempenho de vários tipos de banco de dados, para escolher o que melhor atende as necessidades do projeto.
* Desenvolver uma aplicação web contendo informações em tempo real conectadas diretamente ao Big Data.
* Demonstrar como as tecnologias desenvolvidas podem afetar o futuro das organizações.

## Cronograma

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Mês**  **Tarefas** | **DEZ** | **JAN** | **FEV** | **MAR** | **ABR** | **MAI** | **JUN** |
| Orientação do Pré-projeto | X |  |  |  |  |  |  |
| Formulação do Pré-projeto |  | X |  |  |  |  |  |
| Pesquisas dos itens do pré – projeto |  | X |  |  |  |  |  |
| Fechamento do pré-projeto |  |  | X |  |  |  |  |
| Entrega do pré-projeto |  |  | X |  |  |  |  |
| Orientação da Introdução, Objetivos e Justificativa |  |  | X |  |  |  |  |
| Formulação da Introdução, Objetivos e Justificativas |  |  | X |  |  |  |  |
| Fechamento da Introdução, Objetivos e Justificativas |  |  | X |  |  |  |  |
| Entrega da Introdução, Objetivos e Justificativas |  |  | X |  |  |  |  |
| Orientação do Quadro Teórico |  |  |  | X |  |  |  |
| Formulação do Quadro Teórico |  |  |  | X |  |  |  |
| Entrega do Quadro Teórico |  |  |  | X |  |  |  |
| Orientação do Quadro Metodológico |  |  |  |  | X |  |  |
| Formulação do Quadro Metodológico |  |  |  |  | X |  |  |
| Entrega do Quadro Metodológico |  |  |  |  | X |  |  |
| Revisão do projeto para entrega |  |  |  |  |  | X |  |
| Entrega dos projetos para qualificação |  |  |  |  |  |  | X |
| Bancas de qualificação de Projetos |  |  |  |  |  |  | X |
| Orientações finais dos projetos |  |  |  |  |  |  | X |

## Orçamento

|  |  |  |  |
| --- | --- | --- | --- |
| ORÇAMENTO DETALHADO DO PROJETO | | | |
| 1. RECURSOS MATERIAIS | | | |
| 1.1 Material Permanente: (equipamentos, livros, softwares, equipamentos de informática, etc.). | | | |
| Descrição do Material | Quantidade | Valor (unidade – em reais) | Total R$ |
| Livro 1 | 1 | 100,15 | 500,65 |
| Livro 2 | 1 |  |  |
| Livro 3 | 1 |  |  |
| Livro 4 | 1 |  |  |
| Livro 5 | 1 |  |  |
| Subtotal |  |  | 500,65 |
| 1.2 Material de Consumo: (Papéis necessários para impressões, cartuchos de tinta para impressora, pastas, etc.) | | | |
| Descrição do Material | Quantidade | Valor (unidade – em reais) | Total R$ |
| Impressões | 2 | 2,00 | 4,00 |
| Subtotal |  |  | 4,00 |
|  |  |  |  |
| TOTAL |  |  | Total |
|  |  | 504,65 |

# REFERÊNCIAS

APACHE. **Apache Storm**. Disponível em: <https://storm.apache.org/documentation/Concepts.html>. Acessado em 06/03/2015

CHORODOW, K. **MongoDB The Definitive Guide**. 2. Ed. Sebastopol: O’Reilly, 2013.

DEVMEDIA, **O que é NoSQL?**. Disponível em: <http://www.devmedia.com.br/o-que-e-nosql-java-magazine-86/18777>. Acessado em 03/03/2015.

GIL, Antônio C. **Métodos e técnicas de pesquisa social**. 5. ed. São Paulo: Atlas, 2006. **\_\_\_\_\_. Como elaborar projetos de pesquisa**. 3. ed. São Paul: Atlas, 1996.

INFO ESCOLA. **HTML**. Disponível em: <http://www.infoescola.com/informatica/html/>. Acessado em 07/03/2015.

JUNIOR, E. **HTML5 - Parte 1 – história e elementos de estrutura**. Disponível em: <http://elemarjr.net/2010/10/12/html-5-parte-1-histria-e-elementos-de-estrutura/>. Acessado em 07/03/2015

LEIBIUSKY, J.; EISBRUCH, G.; SIMONASSI, D. **Getting Started Whith Storm.** Sebastopol: O’Reilly, 2012.

MARINESCU, D. C. **Cloud Computing Theory And Pratice.** Waltham: Elsevier Inc., 2013.

MICROSOFT, **Java Script princípios básicos**. Disponível em: <https://msdn.microsoft.com/pt-br/library/ie/6974wx4d%28v=vs.94%29.aspx>. Acessado em 21/02/2015.

PEREIRA, A. P. **O que é CSS?**.Disponível em: <http://www.tecmundo.com.br/programacao/2705-o-que-e-css-.htm>. Acessado em 07/03/2015.

POWERS, S. **Learning JavaScript**. 2. Ed. Sebastopol: O’Reilly, 2009.

PRESS, G. **A Very Short History Of Big Data**. Disponível em: <http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/>. Acessado em 04/02/2015.

SADALAGE, P. J; FOWLER, M. **NoSQL Distilled:** A Brief Guide to the Emerging World of Polyglot Persistence**.** Crawfordsville: Pearson Education Inc, 2013.

SILVA, Renata. **Apostila de metodologia científica**. Brusque: ASSEVIM- Associação Educacional do Vale do Itajaí-Mirim, fev. 2008. (mimeo).

UOL. **História do JQuery**. Disponível em <http://jqueryeter.xpg.uol.com.br/historiajquery.html>. Acessado em 07/03/2015

UTTERBACK, B. **O que é o Bootstrap?** **– Verdades e mitos: Parte 1 de 2**. Disponível em <https://www.prestashop.com/blog/pt/2014/03/06/o-que-e-o-bootstrap-verdades-e-mitos-parte-1-de-2/>. Acessado em 07/03/2015.

W3C, **Visão Geral do HTML5**. Disponível em: <http://www.w3c.br/cursos/html5/conteudo/capitulo1.html>. Acessado em 21/02/2015.