**NÍCOLAS HENRIQUE VIEIRA TOLEDO**

**SEBASTIÃO BATISTA DE ANDRADE NETO**

**ESTATÍSTICAS POLÍTICAS**

**EM TEMPO REAL COM BIG DATA**

**UNIVERSIDADE DO VALE DO SAPUCAI**

**POUSO ALEGRE**

**2015**

**NÍCOLAS HENRIQUE VIEIRA TOLEDO**

**SEBASTIÃO BATISTA DE ANDRADE NETO**

**ESTATÍSTICAS POLÍTICAS**

**EM TEMPO REAL 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

Nos dias atuais, vê-se um grande volume de dados circulando por todo o globo diariamente. Para mais fácil entendimento, foi dado o nome de “Big Data” a esse grande volume de dados. Porém a questão é: “O que fazer com Big Data?”.

De fato, esse volume de dados ocupa uma grande demanda TI de 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.

Como a disputa à presidência estava acirrada, marqueteiros perceberam a importância do imenso volume de dados gerados pelas redes sociais e bombardearam os eleitores com informações quase quem em tempo real dos candidatos, além é claro, de acusações ao concorrente. Redes Sociais tem uma vantagem consideravelmente interessante: Elas encurtam a relação com o eleitor.

Muitos usuários demonstram o contentamento com seus representantes através de redes sociais, pelo simples fato delas serem uma forma fácil de expressão. Geralmente, a população tente 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.

Pelo simples fato dessa grande quantidade de dados conter as opiniões mais diretas da população, estamos propondo um novo modelo de análise de opinião pública no Brasil: Estatísticas políticas em tempo real, usando Big Data.

O contexto da proposta é relativamente simples: Descobrir como a população enxerga as atitudes do governo e seus representantes, em tempo real.

Com essa análise de opinião pública, os políticos poderão saber se estão cumprindo ou falhando em seu dever.

Portanto, esse sistema de análise estatística é especificamente desenvolvido para auxiliar os nossos representantes a descobrir as carências da população, melhorando assim a qualidade de vida do povo.

## Objetivo Geral

Desenvolver um site 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 grande quantidade de dados relacionados a política de redes sociais através de palavras-chave;
* Coletar os dados;
* Analisar os dados de forma eficiente e rápida;
* Apresentá-los em tempo real em um site.

## Justificativas

Dentre tantos temas, foi escolhida a política por ser um fator presente na vida de todos os brasileiros. A intenção é ajudar os políticos a ajudar a população, informando a eles onde e como a população é carente.

Em relação opinião pública, foram escolhidas as redes sociais por serem atualmente a forma mais ampla de protesto político no Brasil.

E, em relação, à tecnologia, foi escolhido o Big Data por ser um assunto de atual relevância e diferencial competitivo no mercado mundial.

# FUNDAMENTAÇÃO TEÓRICA

## Big Data

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 grandes quantidades de dados para a geração de resultados importantes que, em volumes menores, dificilmente seriam alcançados. (Alecrim; Emerson, 2013).

## Banco de dados NoSQL

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 grandes volumes 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. (Java Magazine, 2013).

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.

## Hadoop

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. (Hovhannes; Avoyan, 2013).

## 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.

# QUADRO METODOLÓGICO

## Tipo de Pesquisa

A pesquisa será de forma aplicada.

## Contexto de pesquisa

A página irá ser disponibilizada na internet para que qualquer pessoa possa visualizar como está a aprovação da população em relação ao candidato.

## Participantes

Este projeto terá a participação de Nícolas Henrique Vieira Toledo, Sebastião Batista de Andrade Neto sobre orientação do professor Ednardo David Segura.

## Produção de Dados

Reuniões quinzenais 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 página web contendo informações em tempo real conectadas diretamente ao Big Data.
* Testes.
* Demonstrar, exemplificando o projeto, 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 |

# REFERÊNCIAS

SCUP IDEAS, O Voto Conectado. <http://scup.com/ideas/eleicoes-2014>, acessado em 05 de fevereiro de 2015.

CARTACAPITAL, Política. <http://www.cartacapital.com.br/politica/redes-sociais-ganham-importancia-mas-ainda-nao-definem-eleicao-5440.html>, acessado em 06 de fevereiro de 2015.

INFO WESTER, O que é Big Data?. <http://www.infowester.com/big-data.php>, acessado em 07 de fevereiro de 2015.

IMASTERS, Big Data e Hadoop – o que é tudo isso?. <http://imasters.com.br/tecnologia/redes-e-servidores/big-data-e-hadoop-o-que-e-tudo-isso/>, acessado em 07 de fevereiro de 2015.

TAURION, C. Big Data. Editora Brasport: Rio de Janeiro, 2013. Disponível apenas em formato de livro eletrônico.

PICHATELLI, M . Big Data: Sete Passos iniciais. <http://b2bmagazine.consumidormoderno.uol.com.br/index.php/negocios/item/2768-big-data-sete-passos-iniciais>, acessado em 07 de fevereiro de2015.