**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**

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

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, é visto um grande volume de dados 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 armazenado 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), na pesquisa realizada com executivos de negócios, a fim de entender os desafios e oportunidades associadas ao uso da “análise de negócios”, afirma 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), no artigo sobre como o *Big Data* pode mudar o desempenho da empresa, 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), no artigo de como construir uma aplicação voltada à inteligência social, 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, 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*.

# OBJETIVOS

Nos objetivos serão descritos respectivamente o objetivo geral e os objetivos específicos do projeto, demonstrando quais são as intenções do desenvolvimento do projeto.

## Objetivo Geral

Desenvolver uma aplicação Web que demonstre informações de opiniões políticas dos usuários de redes sociais.

## Objetivos Específicos

Para alcançar o objetivo geral, serão usados como base os seguintes objetivos específicos:

* Coletar uma massa de dados relacionados à política das redes sociais com base em palavras-chave;
* Processar os dados coletados através do Apache *Storm*, posteriormente salvando as informações importantes em uma base de dados;
* Apresentar as informações em uma aplicação Web.

# JUSTIFICATIVA

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 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 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 a fim de servir de material de consulta para alunos do curso de sistemas de informação.

# QUADRO TEÓRICO

No quadro teórico serão descritos o histórico e os conceitos que são relacionados com o tema, bem como as tecnologias que serão aplicadas no desenvolvimento do projeto.

## *Big Data*

A história de como os dados se tornaram tão vastos, gerando o que hoje conhecemos como *Big Data*, vai muito além do atual estado estabelecido. Há nove anos atrás já houve tentativas de quantificar a taxa de crescimento no volume de dados ou o que ficou conhecido como a “explosão de informação” (PRESS, 2013).

Aqui será descrito um resumo como tudo aconteceu desde 1944, quando Freemont Rider, em sua obra “The Scholar and the Future of the Research Library”[[1]](#footnote-1), estimaria que “As bibliotecas universitárias americanas dobravam o tamanho a cada dezesseis anos” (RIDER, 1944, apud PRESS, 2013). Ele especulava que em 2040, a Biblioteca de Yale teria cerca de 200.000.000 de volumes e precisaria de cerca de seis mil pessoas trabalhando no serviço de catálogo (PRESS, 2013).

Hal B. Becker, em sua publicação “Can users really absorb data at today’s rates? Tomorrow’s?”[[2]](#footnote-2), de 1986, estimava que a densidade de caracteres alcançada por Guttenberg era de aproximadamente de 500 caracteres por polegada cubica. Por volta do ano 2000, as memórias RAM armazenariam 1.25X10¹¹ bytes[[3]](#footnote-3) por polegada cubica (PRESS, 2013).

Em 1997 Michael Cox e David Ellsworth publicaram “Application-controlled demand paging for out-of-core visualization”[[4]](#footnote-4), no qual eles começaram o artigo afirmando que o tamanho extenso dos conjuntos de dados penalizam as capacidades da memória principal, disco local e até discos remotos, gerando assim um desafio interessante para os sistemas computacionais. A esse desafio, eles deram o nome de “problema do Big Data”. De acordo com eles, quando conjuntos de dados não cabem mais na memória principal, ou quando os dados não cabem nem nos discos locais, a solução mais comum é adquirir mais recursos (PRESS, 2013).

A publicação de 1999 “*Visually exploring gigabyte data sets in real time”*[[5]](#footnote-5), de Steve Bryson, David Kenwright, Michael Cox, David Ellsworth e Robert Haimes, dizia que computadores poderosos são uma dádiva em diversos campos de investigação, porém em contrapartida, suas operações computacionais rápidas geram quantidades massivas de dados (PRESS, 2013).

Eles também afirmavam que tempos atrás, alguns megabytes[[6]](#footnote-6) de conjuntos de dados eram considerados grandes. Entretanto, na época já eram encontrados conjuntos de dados na casa dos 300 gigabytes[[7]](#footnote-7). Porém se torna um desafio entender os resultados dos dados de computações de alto desempenho. É simplesmente difícil olhar para todos os números. Além do mais, o objetivo da computação é o conhecimento, e não números (PRESS, 2013).

Na publicação “*How Much Information?*”[[8]](#footnote-8), de Peter Lyman e Hal R. Varian feita em 2000, foi apontado que em 1999 o mundo produziu cerca de 1.5 exabytes[[9]](#footnote-9) de informação original, ou cerca de 250 megabytes para cada homem, mulher e criança na terra. Uma vasta quantidade de informação é criada e armazenada por indivíduos (democratização dos dados), e isso cresce velozmente. Chamando essa descoberta de “A dominância do Digital”, Lyman e Varian diziam que maioria da informação textual já “nascia digital” e em poucos anos, a sentença também se tornaria verdadeira para imagens (PRESS, 2013).

Também em 2000, Francis X. Diebold apresentou um documento intitulado “*Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting*”[[10]](#footnote-10), onde ele afirma que “Recentemente, muita ciência boa, seja física, biológica ou social, foi forçada a confrontar, além de obter benefícios, do fenômeno “*Big Data*”” (DIEBOLD, 2000, apud PRESS, 2013).

*Big Data* refere-se como a explosão na quantidade (e as vezes na qualidade) da informação relevante disponível, em grande parte, o resultado de avanços recentes e inéditas em gravação de dados e tecnologia de armazenamento. (PRESS, 2013).

Em 2001, Doug Laney, um analista no Meta Group, publicou uma nota de pesquisa intitulada “*3D Data Management: Controlling Data Volume, Velocity, and Variety.*”[[11]](#footnote-11) Uma década depois, os “3v's” se tornaram as três dimensões que definem o *Big Data*, mesmo que nas notas de Laney o termo em si não aparece (PRESS, 2013).

Em 2007 John F Gantz, David Reinsel e outros pesquisadores lançam um livro, intitulado “*The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010*”[[12]](#footnote-12). Nele, era afirmado que em 2006, o mundo produziu 161 exabytes de dados, além de prever que entre 2006 e 2010, a informação adicionada anualmente ao universo digital aumentaria para 988 exabytes, ou dobrando a cada 18 meses. De acordo com os lançamentos de 2010 e 2012 do mesmo estudo, a quantidade de dados digitais ultrapassou esta previsão, atingindo 1.227 exabytes em 2010, e crescendo a 2.837 exabytes em 2012 (PRESS, 2013).

Em 2008 Randal E. Bryant, Randy H. Katz, e Edward D. Lazowska publicaram “*Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society*”[[13]](#footnote-13).

Para Bryant; Katz e Lazowska (2008, p.1 e seg. apud Press 2013):

Assim como os motores de busca têm transformado a forma de acessar as informações, outras formas de *Big Data* podem e vão transformar as atividades das empresas, pesquisadores científicos, médicos e operações de defesa e inteligência da nossa nação[...] *Big Data* é, talvez, a maior inovação na computação na última década. Nós apenas começamos a ver o seu potencial para coletar, organizar e processar dados em todas as esferas da vida. Um modesto investimento por parte do governo federal poderia acelerar consideravelmente o seu desenvolvimento e implementação.

Em 2010 Kenneth Cukier informa em seu relatório “*Data, data everywhere*”[[14]](#footnote-14) que o mundo contém uma inimaginável quantidade de informação digital que está crescendo cada vez mais mais rapidamente, e que o efeito pode ser visto em todos os lugares. Cientistas e engenheiros de computação designaram termo para o fenômeno: “*Big Data*” (PRESS, 2013).

Na publicação de 2011 “*Big data: The next frontier for innovation, competition, and productivity*”[[15]](#footnote-15) feita por James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, e Angela Byers, foi estimado que até 2009, quase todos os setores da economia americana tinham, pelo menos, uma média de 200 terabytes[[16]](#footnote-16) de dados armazenados por empresa (o dobro do tamanho do *Data Warehouse*[[17]](#footnote-17) da varejista americana *Wal-Mart* em 1999). No total, o estudo estimou que 7,4 exabytes de novos dados foram armazenados pelas empresas e 6,8 exabytes por consumidores em 2010 (PRESS, 2013).

Na publicação de 2012 “*Critical Questions for Big Data in* Information, Communications, and Society**”**[[18]](#footnote-18), Danah Boyd e Kate Crawford definem *Big Data* como um fenômeno cultural, tecnológico e acadêmico que é formado da interação entre tecnologia (capacidade computacional e precisão algorítmica para coletar, analisar, unir e comparar dados), análise (identificação padrões, a fim de fazer reivindicações econômicas, sociais, técnicas e legais) e mitologia (crença de que grandes conjuntos de dados oferecem uma forma mais elevada de inteligência e conhecimento que podem gerar perspectivas que antes eram impossíveis, com verdade, objetividade e precisão) (PRESS, 2013).

De acordo com Soubra (2012), existem três propriedades que são capazes de definir a expansão do Big Data: Volume, Variedade e Velocidade. A junção destes é chamada de “os 3 v’s do Big Data”. Cada um deles pode ser definido da seguinte forma:

* Volume: Quantidades de dados realmente grandes, que tem crescimento exponencial;
* Variedade: O grande volume de dados se deve ao fato da existência de uma grande variedade de informações. Os dados podem ser classificados em estruturados, semi-estruturados e não estruturados.
* Velocidade: O tratamento dos dados deve ser feito em tempo hábil. O volume de informações, com uma variedade de dados diversa, deve ser processado e analisado no menor tempo possível.

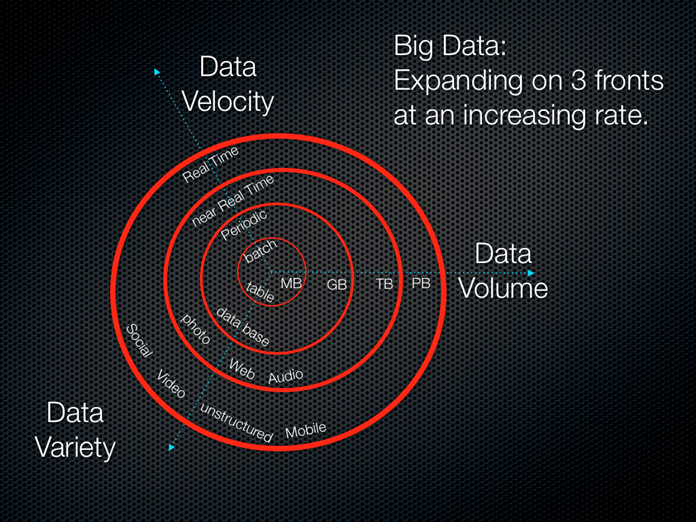


Figura 1 - Os 3 v’s que definem o *Big Data*. Fonte: Data Science Central (2012)

A cada dia que se passa o volume de dados cresce de forma exponencial. Desse modo, se vê necessária a existência de uma abordagem que tenha a capacidade de armazenar de maneira segura uma quantidade de dados crescente em tamanho e variadade. Além disso, existe a necessidade de processar e apresentar essas informações em tempo hábil para a tomada de decisões. Para o atendimento de todas essas necessidades, foi criado o conceito de *Big Data* (ALECRIM, 2015).

## *Data Mining*

Para Han, Kamber e Pei (2012), *Data Mining* (ou mineração de dados) pode ser definida de diversas formas. O termo “mineração” é constantemente usado para caracterizar o processo em que se encontra um conjunto de “pepitas” de uma grande quantidade de matéria-prima. Assim, nasceu o termo *Data Mining*, que em suma, representa a extração de conhecimento a partir da análise de dados.

De acordo com Han, Kamber e Pei (2012), o processo de descoberta do conhecimento segue a sequência iterativa dos seguintes passos:

* Limpeza dos dados: Remoção de lixo e dados inconsistentes;
* Integração dos dados: Combinação de múltiplas fontes de dados;
* Seleção dos dados: Os dados relevantes para a tarefa são obtidos do banco de dados;
* Transformação dos dados: Os dados são transformados e consolidados em formas apropriadas para a mineração por operações de resumo ou operações de agregação;
* Mineração de dados: Processo onde são aplicados métodos inteligentes para a extração de padrões de dados;
* Avaliação dos padrões: Identificação dos padrões verdadeiramente interessantes que representam conhecimento baseados em medidas de interesse;
* Apresentação do conhecimento: Técnicas de visualização e representação do conhecimento são usadas para apresentar o conhecimento extraído para os usuários.

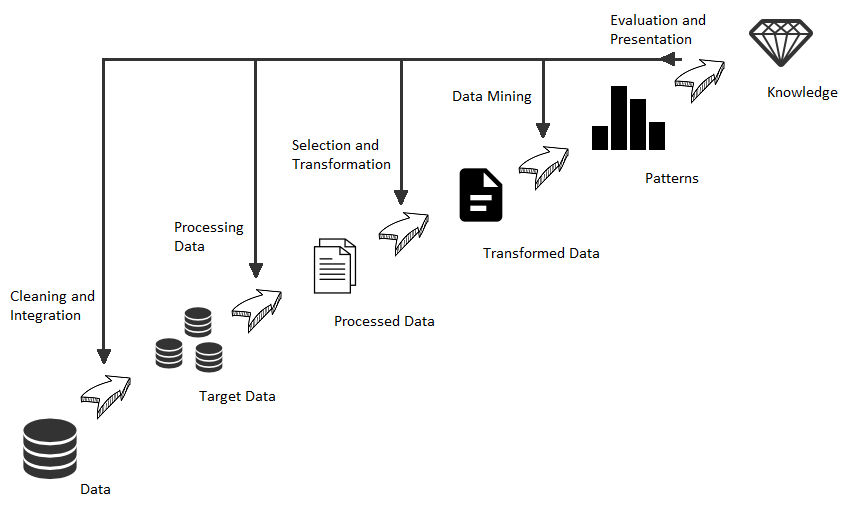


Figura 2 - *Data Mining* como um passo no processo da descoberta do conhecimento. Fonte: Elaborado pelos autores.

A mineração de dados é o processo de descobrir padrões e conhecimento de grandes quantidades de dados. As fontes de dados podem incluir bancos de dados, *data warehouses*, a Web e outros repositórios de informações ou dados que serão transmitidos para o sistema de forma dinâmica (HAN, KAMBER e PEI, 2012).

### *Data Mining* em Redes Sociais

De acordo com Russel (2013), algumas das redes sociais eram consideradas inicialmente como uma simples “moda”. Posteriormente, passaram a ser consideradas como um “*mainstream*”[[19]](#footnote-19). Atualmente, chegaram ao reconhecimento de uma espécie de “grife”, chegando a mudar a maneira de vida da população dentro e fora da Web.

Para Russel (2013), de um modo geral, as redes sociais, juntamente com *data mining*,análises, e técnicas de visualização para explorar dados buscam responder as seguintes perguntas:

* “Quem conhece quem, e quem são os contatos comuns em suas redes sociais?”;
* “Com que frequência pessoas específicas se comunicam com outras?”;
* “Qual rede social gera mais valor para um nicho específico?”;
* “Como a geografia afeta suas conexões sociais em um mundo online?”
* “Quem são as pessoas mais influentes/procuradas em uma rede social?”;
* “Sobre o que as pessoas estão conversando? (É valioso?)”;
* “Quem são as pessoas interessadas, com base na linguagem humana, e o que eles utilizam no mundo digital?”.

A resposta a essas perguntas muitas vezes deram oportunidades valiosas e lucrativas a empresários, cientistas sociais e outros profissionais, promovendo assim, o entendimento de problemas e a descoberta de soluções (RUSSEL, 2013).

## REST API

Para Massé (2011), *Representational State Transfer* (REST) é o nome de uma arquitetura Web que pode ser implementada usando diferentes tecnologias, além de instanciá-la utilizando diferentes valores para suas propriedades abstratas. O estilo arquitetural REST é comumente aplicado ao desenvolvimento de API’s[[20]](#footnote-20) para *Web Services*[[21]](#footnote-21). Uma API em conformidade com a arquitetura REST é uma REST API.

Ainda de acordo com Massé (2011), algumas das melhores práticas para o design REST API estão implícitas no padrão HTTP, enquanto outras abordagens “pseudo-padrão” surgiram ao longo dos anos. No entanto, nos dias atuais recomenda-se continuar a buscar as respostas para uma série de questões, tais como:

* “Quando um *path* URI deve ser nomeado com substantivo no plural?”;
* “Qual método de requisição deve ser usado para atualizar o estado de recurso?”;
* “Como fazer para mapear operações não-CRUD aos URI’s?”;
* “Qual é o código de status de resposta HTTP mais apropriado para um determinado cenário?”;
* “Como gerenciar as versões de representação do estado de um recurso?”;
* “Como estruturar um *hyperlink* em JSON?”.

O conhecimento dos princípios do REST trazem benefícios no desenvolvimento de aplicações Web, bem como na transformação da API em um *Web Server*[[22]](#footnote-22) robusto (MASSÉ, 2011).

### Twitter4j

Para Twitter4j (2015), o *Twitter4j* é uma biblioteca Java não-oficial para o consumo do *Web Service* do Twitter. A principal característica da *Twitter4j* é a facilidade de integração entre a aplicação Java e o *Twitter Service*.

De acordo com Twitter4j (2015), são características da biblioteca *Twitter4j*:

* Funciona em qualquer plataforma Java na versão 5 ou posterior;
* Plataformas *Android* e *Google* *App Engine* incluídas;
* Dependência zero - Não requer jar’s[[23]](#footnote-23) adicionais;
* Suporte a OAuth[[24]](#footnote-24) integrado;
* Suporte a gzip[[25]](#footnote-25);
* Compatível com a versão 1.1 do Twitter API.

A biblioteca *Twitter4J* possui uma série de funcionalidades simples de serem utilizadas, resultando em uma implementação rápida e sem maiores dificuldades (TWITTER4J, 2015).

### Facebook4j

Para Facebook4j (2015), o *Facebook4j* é uma biblioteca Java não-oficial para o Facebook *Graph* API[[26]](#footnote-26). A biblioteca usa a mesma lógica interna do *Twitter4j* para lidar com pedidos HTTP *request/response*.

De acordo com Facebook4j (2015), são características da biblioteca *Facebook4j*:

* Funciona em qualquer plataforma Java na versão 5 ou posterior;
* Dependência zero - Não requer jar’s adicionais;
* Suporte a OAuth integrado.

A principal funcionalidade da biblioteca *Facebook4J*, além de coletar dados do Facebook, é a sua fácil implementação, diminuindo assim a complexidade da mineração dos dados (FACEBOOK4J, 2015).

## *Storm*

De acordo com Leibiusky, Eisbruch e Simonassi (2012), o *Storm* é um sistema distribuído, confiável e tolerante a falhas desenvolvido para processar fluxos de dados. Sua estrutura consiste em diferentes tipos de componentes. Desta forma o trabalho é dividido entre os componentes, cada qual é responsável por uma tarefa de processamento simples e específica. O fluxo de entrada do *cluster*[[27]](#footnote-27) do *Storm* é tratado por um componente chamado *spout*.

Segundo Apache (2015), um *spout* pode ser denominado como a fonte de fluxos em uma topologia. Geralmente, eles irão ler tuplas (sequências de objetos) de uma fonte externa (uma API de rede social, por exemplo), posterirormente encaminhando-os para a topologia.

Continuando com Leibiusky, Eisbruch e Simonassi (2012), o *spout* por sua vez direciona os dados para um componente denominado *bolt*, o qual irá transformá-los de alguma maneira. Um *bolt* pode tanto persistir os dados em algum tipo de armazenamento, quanto direcioná-lo para outro *bolt*. Para melhor entendimento, pode-se imaginar o *cluster* do *Storm* como uma corrente de *bolts*, cada qual executa algum tipo de transformação nos dados expostos pelo *spout*.

Leibiusky, Eisbruch e Simonassi (2012) afirmam que, dentre propriedades do *Storm*, as principais a serem citadas são:

* Simples de programar: A complexidade é reduzida;
* Suporte a várias linguagens de programação: Maior facilidade desenvolver em linguagens baseadas na JVM[[28]](#footnote-28), no entanto qualquer linguagem é suportada desde que se implemente uma pequena biblioteca intermediária;
* Tolerante a falhas: O *cluster* do *Storm* cuida dos trabalhos, realocando as tarefas quando necessário;
* Escalável: O *Storm* pode reatribuir as tarefas para as novas máquinas, caso seja necessário;
* Confiável: Todas as mensagens são obrigatoriamente processadas ao menos uma vez. Em caso de erros, a mensagem pode ser processada novamente. Assim, tem-se a garantia de nunca perder uma mensagem;
* Rápido: Velocidade é um dos fatores-chave do *Storm*;
* Transitivo: Pode-se obter exatamente a mesma semântica de mensagens para praticamente qualquer tipo de computação.

O *Storm* torna-se uma solução eficiente, rápida, confiável e de baixo custo para operações relacionadas à *Big Data*. Grandes massas de dados tem que ser processadas no menor tempo possível, com a garantia que nenhum dado se perca. A ferramenta possibilita este tipo de solução, através de seu design e suas funcionalidades (LEIBIUSKY, EISBRUCH e SIMONASSI, 2012).

## Banco de dados NoSQL

Segundo Sadalage e Fowler (2013), inicialmente desenvolvido por Carlo Strozzi, esse tipo de banco de dados tem como característica o armazenamento de suas tabelas como arquivos do tipo ASCII[[29]](#footnote-29). Cada uma de suas tuplas é representada por uma linha, que tem os seus campos separados por guias.

Para Sadalage e Fowler (2012), o nome “NoSQL” vem do fato de não ser usado SQL[[30]](#footnote-30) como uma linguagem de consulta aos seus registros. Ao invés disso, o banco de dados NoSQL é manipulado através de *shell scripts*[[31]](#footnote-31) que podem ser combinados nos *pipelines* do UNIX[[32]](#footnote-32).

Ainda de acordo com Sadalage e Fowler (2012), o uso do NoSQL que é reconhecido em dias atuais, remonta a um encontro que ocorreu em 11 de junho de 2009 em São Francisco, EUA, organizado pelo desenvolvedor de software Johan Oskarsson. Um exemplo que foi desenvolvido usando BigTable[[33]](#footnote-33) e Dynamo[[34]](#footnote-34) havia inspirado diversos projetos a experimentar um armazenamento de dados alternativo, e discussões a respeito do NoSQL haviam se tornado uma característica das melhores conferências de software em torno desse tempo.

De acordo DevMedia (2013), o movimento NoSQL promove soluções de armazenamento de dados não relacionais. Suas diversas ferramentas resolvem problemas relacionados a grandes volumes de dados, execução de consultas com baixa latência e modelos flexíveis de armazenamento de dados (como XML[[35]](#footnote-35) ou JSON[[36]](#footnote-36)).

O objetivo do NoSQL não é de substituir os bancos de dados relacionais, mas sim apresentar algumas soluções, que em determinados momentos, são mais adequadas e eficientes. Assim, nada se impede de se trabalhar com as tecnologias baseadas em NoSQL e os tradicionais bancos de dados relacionais em um mesmo projeto (DEVMEDIA, 2013).

Para DevMedia (2013), os tipos de bancos de dados NoSQL 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 é escolhido após uma análise crítica do gestor de TI da organização, dependendo da necessidade da mesma (DEVMEDIA, 2013).

### MongoDB

Segundo Chodorow (2013), o MongoDB é um banco de dados NoSQL poderoso, flexível e escalável. Ele combina a capacidade de dimensionmento com características tais como índices secundários, consultas por abrangência, classificação, agregações e índices geoespaciais.

O MongoDB não é um banco de dados relacional, mas sim um banco de dados orientado a documentos. Uma das principais razões para a migração dos tradicionais modelos relacionais é a necessidade de fazer um escalonamento de modo mais fácil (CHODOROW, 2013).

Um banco de dados orientado a documentos substitui o tradicional conceito de uma "linha" com um modelo mais flexível, o "documento". Ao permitir a incorporação de documentos e *arrays*[[37]](#footnote-37), o banco de dados orientado a documento permite representar relações hierárquicas mais complexas com um único registro. Isso se encaixa naturalmente a forma como os desenvolvedores de linguagens orientadas a objetos pensam sobre seus dados (CHODOROW, 2013).

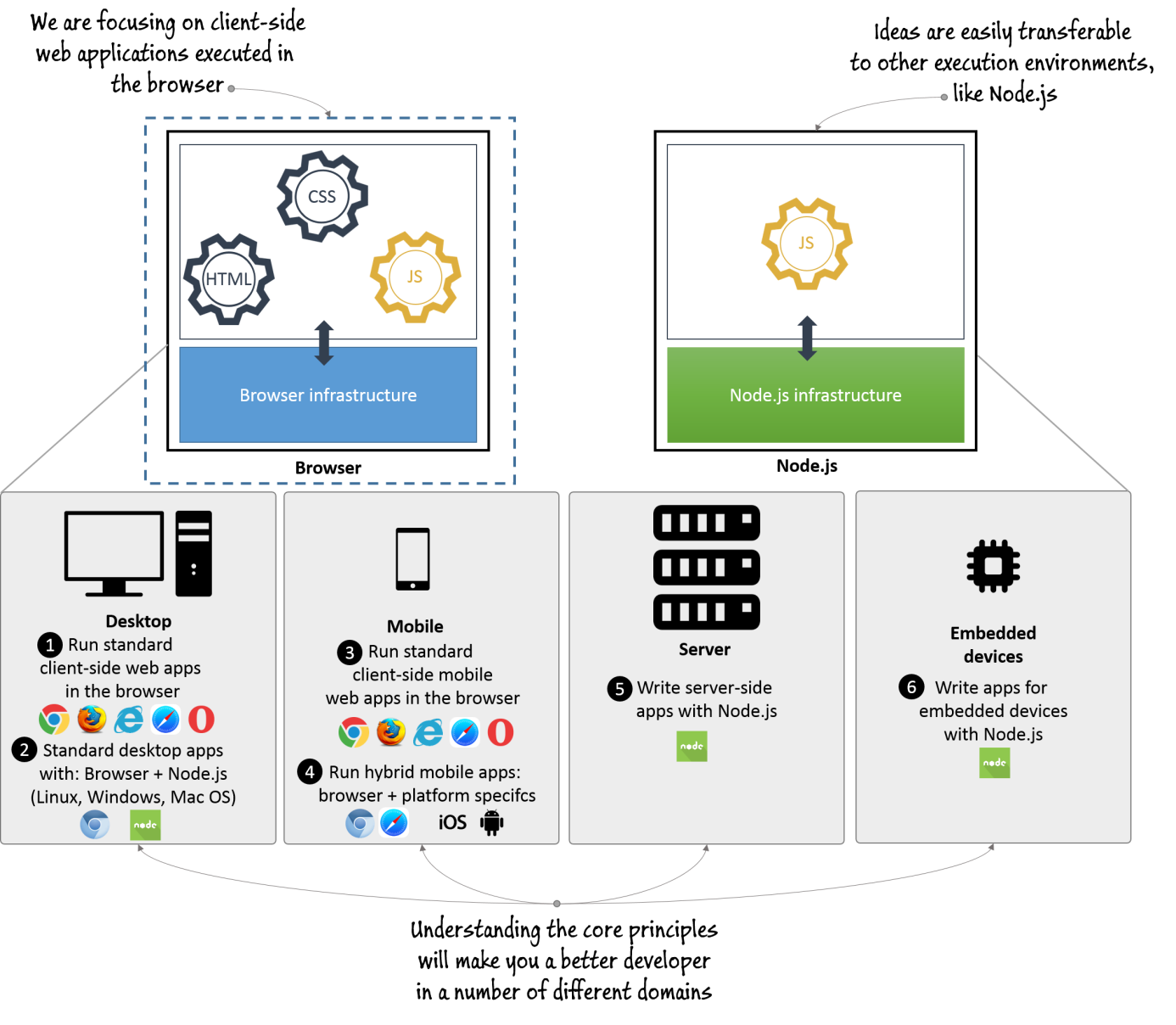
O MongoDB foi projetado para ser escalonável. O seu modelo de dados orientado a documentos o permite separar os dados entre vários servidores de maneira mais fácil. Ele se encarrega automaticamente de balancear a carga de dados através de um *cluster*[[38]](#footnote-38). Assim, ele consegue redistribuir documentos automaticamente e rotear as solicitações dos usuários para as máquinas corretas. Esta característica permite que os desenvolvedores se concentrem no desenvolvimento da aplicação, não do escalonamento (CHODOROW, 2013).

Caso um cluster necessite de mais capacidade, novas máquinas podem ser adicionadas. Automaticamente, o MongoDB irá descobrir como os dados devem ser distribuídos entre elas (CHODOROW, 2013).

## Node.JS

De acordo com Pereira (2013), os sistemas Web desenvolvidos sobre plataforma .NET, Java, PHP, Ruby ou Pythoncompartilham da mesma característica da paralisação de processamento enquanto utilizam um I/O[[39]](#footnote-39) no servidor. Essa paralisação é conhecida com modelo bloqueante (*Blocking-Thread*). O problema é que essas arquiteturas se tornam ineficientes em alguns casos, pois elas mantêm uma fila ociosa enquanto as operações de I/O são executadas, bloqueando totalmente o sistema. Desse modo, o aumento de acessos no sistema geraram *gap’s*[[40]](#footnote-40)mais frequentes, aumentando a necessidade de fazer um *upgrade* de *hardware* nos servidores, o que é financeiramente inviável.

Baseado neste problema que em 2009, Ryan Dahl e mais 14 colaboradores criaram o Node.JS. O Node.JS tem um modelo inovador, sendo sua arquitetura não-bloqueante (*non-blocking thread*), apresentando uma boa performance e baixo consumo de memória (PEREIRA, 2013).



O ecossistema *JavaScript*. Fonte: DZone (2015).

O Node.JS é uma plataforma escalável e de baixo nível, pois a programação é feita diretamente com protocolos de rede e bibliotecas que acessam os recursos do sistema operacional. O *JavaScript* é a sua linguagem de programação. Isso se deu graças à *engine* *JavaScript* V8, utilizada também no navegador Google Chrome (PEREIRA, 2013).

### *Express*

Para Brown (2014), o *Express* é um Web *framework*[[41]](#footnote-41) Node.JS minimalista e flexível que fornece um conjunto robusto de recursos para a criação das mais diversas aplicações Web.

De acordo com Brown (2014), o *Express* tem como características:

* Minimalista: A filosofia do *Express* é proporcionar a camada mínima entre o desenvolvedor e o servidor. Isso significa que ele libera o desenvolvedor para ter a plena expressão de suas ideias, e, ao mesmo tempo, proporciona algo útil para o mesmo.
* Flexível: Como o Express fornece um framework minimalista, fica a cargo do desenvolvedor adicionar ao *Express* diferentes partes de sua funcionalidade quando for necessário, substituindo assim o que não atender as suas necessidades.
* Web *Framework*: Possibilita a construção de páginas Web, sites e aplicações Web. É fácil imaginar que em poucos anos, não existirá mais a distinção entre eles.
* Suporte a aplicações Web de página única: Em vez de um site fazendo requisições de rede à medida que o usuário vai navegando por páginas diferentes, uma aplicação web de página única faz o *download* do site inteiro para o cliente do navegador. O desenvolvimento desse tipo de aplicação é facilitado pelo *Express*.
* Suporte a aplicações Web Multi-páginas e Aplicações Web híbridas: As aplicações Web Multi-páginas usam a abordagem tradicional para o desenvolvimento de websites, enquanto as aplicações Web híbridas usam em conjunto as aplicações Web de página única e as aplicações Web Multi-páginas. O desenvolvimento de ambas é facilitado pelo *Express*.

O *Express* facilita o desenvolvimento de aplicações Web que utilizam o Node.JS como servidor. Ele é minimalista, robusto e flexível, além de ter um eficiente sistema de roteamento, um executável para a geração de aplicações, dentre outras vantagens (NODEBR, 2013).

## *Cloud Computing*

De acordo com Marinescu (2013), o ideal que a computação pode ser livremente distribuída como uma utilidade pública, do mesmo modo que a água e a eletricidade, foi formulado na década de 1960 pelo cientista da computação e visionário John McCarthy, o mesmo que defendeu a lógica matemática na inteligência artificial. Cerca de quatro décadas depois, com a internet já consolidada, a causa da “utilidade computacional” foi resgatada por grandes empresas de TI, tais como Amazon, Apple, Google, HP, IBM, Microsoft e Oracle.

O movimento da computação em nuvem foi um movimento iniciado em algum período em meados da primeira década do novo milênio, cuja ideia principal é que o processamento de informações pode ser feito com mais eficiência em métodos de armazenamento via internet, sendo assim alocados em um “*server farm*”, ou seja, um conjunto de servidores com capacidades computacionais extremamente superiores a de um computador convencional (MARINESCU, 2013).

Uma das principais vantagens da computação em nuvem é que ela oferece serviços computacionais e de armazenamento escaláveis e elásticos. Todos os recursos utilizados para estes serviços são medidos. Com isso, os usuários serão somente cobrados pelos serviços que eles consomem. Aplicações científicas e de engenharia, mineração de dados, financiamento computacional, jogos, redes sociais, e muitas outras atividades computacionais de uso intensivo de dados podem se beneficiar da computação em nuvem (MARINESCU, 2013).

A arquitetura da computação em nuvem oferece três tipos de serviços: Software como serviço (SaaS), Plataforma como serviço (PaaS) e Infraestrutura como Serviço (IaaS). Os serviços baseados em SaaS são mais designados ao usuário final. Os serviços baseados em PaaS são mais voltados a área de desenvolvimento, oferecendo assim poderosas ferramentas ao desenvolvedor. Já os serviços baseados em IaaS oferecem serviços de infraestrutura de TI completos, liberando assim a organização para focar nos negócios (MARINESCU, 2013).

A computação em nuvem já é uma realidade empresarial em dias atuais. O número de organizações que estão adotando o paradigma é crescente (MARINESCU, 2013).

### *Amazon Web Services*

De acordo com Murty (2008), o *Amazon Web Services* é um conjunto de *web services* disponibilizados pela *Amazon*[[42]](#footnote-42) que permitem aos desenvolvedores acessar e desenvolver sobre a sua plataforma tecnológica. Neste conjunto de *web services*, estão inclusos diversos serviços de infraestrutura, os quais têm a finalidade de expansão ou substituição da infraestrutura física tradicional de aplicações Web (MURTY, 2008).

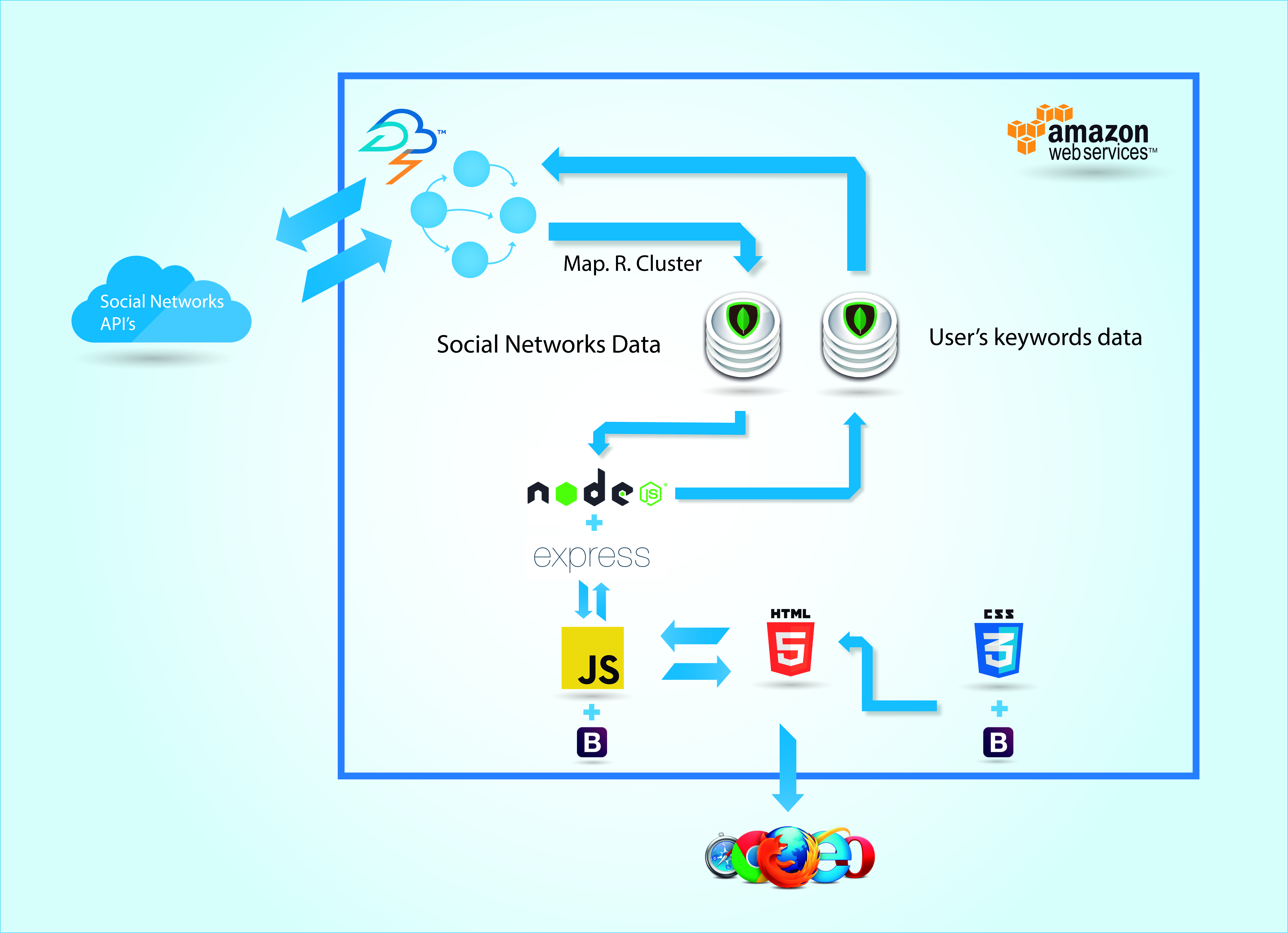


Diagrama demonstrando a fronteira da *Amazon Web Services e* as tecnologias empregadas neste Projeto. Fonte: Elaborado pelos autores.

Estes serviços de infraestrutura fornecem armazenamento, capacidade computacional, um sistema de mensagens, um sistema de pagamento e um banco de dados que pode ser acessado por qualquer pessoa, desde que possua uma conta na Amazon.com e um cartão de crédito. A principal característica, é que os usuários pagam somente pelo que usam (MURTY, 2008).

Esses serviços oferecem uma alternativa ao tradicional desenvolvimento de aplicações sob o hardware físico, pois eles fornecem componentes de infraestrutura escalável, confiável e rentável sem exigir muito esforço, ao contrário da montagem do hardware proprietário. Desse modo, o foco se torna a aplicação em vez da plataforma física (MURTY, 2008).

Outra vantagem desses serviços é a redução significativa do investimento inicial exigido para construir e oferecer um serviço. A partir do momento em que a gestão de infraestrutura está sob a responsabilidade da *Amazon*, o foco se torna o desenvolvimento da aplicação e o crescimento dos negócios. Desse modo, a aplicação web está liberada para competir e ter sucesso com base em suas ideias, inovação e execução (MURTY, 2008).

## *Open Web Plataform*

De acordo com W3C (2015), a *Open Web Plataform* é uma coleção de tecnologias abertas que podem ser implementadas sem a necessidade de qualquer aprovação ou pagamento de taxas de licença.

Seguindo esse conceito, neste tópico serão descritas as tecnologias da *Open Web Plataform* que serão utilizadas neste projeto.

### HTML5

Segundo W3C (2015), a WWW (*World Wide Web*) foi criada em 1989, por Tim Berners-Lee no CERN (*European Organization for Nuclear Research*[[43]](#footnote-43)) em Genebra, na Suíça. Berners-Lee estava trabalhando em uma seção de serviços de computação no CERN, quando ele teve a ideia de reunir e organizar as pesquisas de diversos cientistas do mundo todo. Mais do que somente reunir, ele propôs que os textos poderiam ter *links*[[44]](#footnote-44) entre si, ou seja, durante a leitura de uma pesquisa, seria possível rapidamente exibir parte de outro artigo que contém texto ou diagramas diretamente relevantes.

Como estava tudo muito recente, era necessário ter uma implementação relativamente simples. Seguindo esse conceito, Berners-Lee desenvolveu o protocolo HTTP. Desse modo, o formato de texto escrito para o HTTP foi nomeado de HTML. Entre 1993 e 1997, o HTML iria passar pelas versões HTML+, HTML 2.0, HTML 3.0, HTML 4.0 e HTML 4.0.1, todas abrangendo diversas propostas para enriquecer as possibilidades da linguagem. Em 2009, foi lançado o HTML5, sua atual versão (W3C, 2015).

Ainda de acordo com W3C (2015), o HTML5 tem como uma de suas principais funções a facilidade de manipulação do elemento. Com isso, o desenvolvedor pode modificar as características dos objetos de forma não intrusiva e transparente para o usuário final. Contrariamente as suas versões anteriores, o HTML5 fornece um número de ferramentas para o CSS[[45]](#footnote-45) e o *JavaScript*[[46]](#footnote-46) executarem suas funções de maneira mais eficiente possível.

Continuando com W3C (2015), HTML é uma abreviação de *Hypertext* *Markup* *Language* - Linguagem de Marcação de Hypertexto. A função do HTML é a publicação de conteúdo (texto, imagem, vídeo, áudio e etc) na Web.

### CSS3

Segundo Pereira (2009), o CSS nasceu das complicações de aparência e da diferença de compatibilidade dos navegadores das páginas HTML. A principal função do CSS é separar o conteúdo e formato de um documento de sua apresentação, incluindo elementos como cores, formatos de fontes e layout, além de ser responsável por definir animações e transições. Com isso, foi proporcionada uma maior flexibilidade e controle na especificação de como as características serão exibidas, além de permitir um compartilhamento de formato e reduzir a repetição no conteúdo estrutural de uma página.

Para Mozilla Developer Network (2015), o CSS3 é a mais recente evolução do CSS. Dentre suas novidades, estão cantos arredondados, sombras, gradientes, transições ou animações, bem como novos *layouts* como multiplas colunas, caixa flexível ou *layouts* de grade. As partes experimentais devem ser usadas com cautela, pois tanto sua sintaxe quanto sua semântica podem mudar no futuro.

### *JavaScript*

De acordo com Powers (2009), a linguagem *JavaScript* foi inicialmente destinada a ser uma interface *script* entre uma página Web carregada no cliente do navegador (na época, o Netscape) e a aplicação no servidor. Desde a sua introdução em 1995, *JavaScript* se tornou um componente chave do desenvolvimento Web, além da descoberta seu uso em outros lugares.

Segundo Microsoft (2015), a linguagem *JavaScript* é uma linguagem de *script* interpretada com base em objetos. Sua sintaxe é semelhante a sintaxe da linguagem C.

Continuando com Powers (2009), *JavaScript* é uma linguagem fracamente tipada, ou seja, os tipos de dados das variáveis não são declarados de forma explícita. Na maioria das vezes, o *JavaScript* executa conversões de tipo das varáveis automaticamente.

## *Bootstrap*

De acordo com Magno (2012), o *Bootstrap* é um *framework* *front-end*[[47]](#footnote-47) intuitivo criado para facilitar o desenvolvimento dos elementos de interface em páginas Web. Pode ser também utilizado como um guia para reproduzir de forma consistente os padrões de desenvolvimento consolidados pelo Twitter. O *Bootstrap* também pode ser utilizado para facilitar na padronização e nas melhores práticas de desenvolvimento HTML/CSS e JavaScript, tanto para iniciantes, quanto para desenvolvedores com habilidades mais avançadas que desejam dar um passo além em interações mais complexas.

# QUADRO METODOLÓGICO

No quadro metodológico serão apresentadas as informações e procedimentos necessários para a execução desta pesquisa.

## Tipo de Pesquisa

A pesquisa tem como objetivo a obtenção de novos conhecimentos, tomando por base a utilização de procedimentos científicos. Contribui para a solução dos problemas e processos eventuais nas mais diversas atividades humanas, em ações comunitárias, no processo de formação e outros. Dessa forma, o conhecimento se torna uma ferramenta para o desenvolvimento do ser humano e a pesquisa uma consolidação da ciência (SILVA, 2008).

Para Oliveira (2002, p. 62), “A pesquisa, tanto para efeito científico como profissional, envolve a abertura de horizontes e a apresentação de diretrizes fundamentais, que podem contribuir para o desenvolvimento do conhecimento”.

Abrangendo o contexto de pesquisa, este projeto tomará por base a metodologia de pesquisa aplicada, pois será desenvolvida uma aplicação que fará o processamento, a análise e a demonstração de resultados das informações coletadas, sendo aplicado o conceito de *Big Data*.

De acordo com Marconi; Lakatos (2009, p. 6), “Pesquisa aplicada caracteriza-se por seu interesse prático, isto é, que os resultados sejam aplicados ou utilizados, imediatamente, na solução de problemas que ocorrem na realidade”.

A opção por essa metodologia foi feita pelo motivo dela abranger a fundamentação necessária para ser feita a demonstração das ferramentas e conceitos em torno do *Big Data*, além da análise de dados e suas devidas conclusões.

## Contexto de pesquisa

Do ponto de vista acadêmico, este projeto tem como público alvo qualquer estudante de graduação na área tecnológica com interesse em obter conhecimentos relacionados à *Big Data* e suas ferramentas, além de mineração de dados de redes sociais.

Do ponto de vista social, este projeto tem como público alvo qualquer pessoa brasileira com o direito de voto e com acesso à internet que esteja disposta a acompanhar a opinião pública em relação ao seu representante. A pesquisa abrange também o político brasileiro com acesso à internet, caso haja o interesse do mesmo em acompanhar a sua avaliação perante a população.

Tomando como princípio o software livre, toda parte relacionada ao desenvolvimento da aplicação será disponibilizada em um controlador de versão[[48]](#footnote-48) gratuito na internet, sem qualquer tipo de restrição ao seu acesso. Assim, qualquer pessoa pode visualizar e dar continuidade à evolução do projeto.

## Participantes

Nícolas Henrique Vieira Toledo, aluno do 7º (sétimo) período do curso de Sistemas de Informação da Universidade do Vale do Sapucaí (UNIVÁS) é estagiário em Desenvolvimento de Software no INATEL. Possui formação técnica pela Microcamp de Bragança Paulista - SP. Possui conhecimentos em HTML, CSS, JavaScript, PHP, Java, C#, MySQL, PostgresSQL, SQLServer, 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) é microempreendedor de uma empresa especializada em montagem e manutenção de computadores localizada em Silvianópolis - MG. Possui formação técnica pelo INPETECC de Pouso Alegre - MG. Possui conhecimentos em HTML, CSS, JavaScript, PHP, Java, Python, PostgresSQL, ShellScript, Redes, Linux e inglês intermediário. Neste projeto, atuará como desenvolvedor.

Ednardo David Segura é Professor da Universidade do Vale do Sapucaí (UNIVÁS) e Especialista em Sistemas no INATEL. É 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, MG. Neste projeto, atuará como orientador.

## Instrumentos

Para ser feita a escolha do tipo de instrumento, se faz necessária a observação do que será estudado.

Na realização de uma pesquisa, segundo Oliveira (2002, p. 66):

Depois de definidas as fontes de dados e o tipo de pesquisa, que pode ser de campo ou de laboratório, devemos levantar as técnicas a serem utilizadas para a coleta de dados, destacando-se: questionários, entrevistas, observação, formulários e discussão em grupo.

A coleta de dados será realizada por meio de livros e materiais relacionados com o tema, trabalhos e artigos acadêmicos, páginas na internet, pesquisas bibliográficas e reuniões com pessoas com conhecimentos relacionados com o tema.

Também serão feitas reuniões semanais com o orientador do projeto, com o objetivo de mostrar pesquisas e resultados, definir o foco do projeto até a reunião seguinte, tomar as devidas decisões relacionadas ao desenvolvimento da aplicação e escolher a forma da apresentação das tecnologias abordadas no projeto.

## Procedimentos

Os procedimentos constituem a finalização de apresentação de um projeto de pesquisa. Após ser feito o esclarecimento do problema, será feita a demonstração de como o problema será abordado empiricamente.

Este projeto será divido em 6 etapas:

* Descrever o conceito de *Big Data* e suas ferramentas;
* Levantar os requisitos necessários para o desenvolvimento da aplicação;
* Desenvolver uma aplicação Web contendo informações conectadas diretamente ao *Big Data*;
* Hospedar o *Big Data*, suas ferramentas e a aplicação Web na nuvem;
* Fazer a coleta, processamento e análise dos dados;
* Apresentar os resultados.

Após ser feita a conclusão destas etapas, será feita a demonstração de como as tecnologias envolvidas neste trabalho podem afetar o futuro das organizações.

## Cronograma

O cronograma tem como finalidade auxiliar no gerenciamento do processo desenvolvido no projeto, permitindo assim, a visualização de seu andamento de acordo com o que foi programado. Desse modo, os cronogramas demonstrados nos itens 5.6.1 e 5.6.2 demonstram como serão realizadas as etapas do desenvolvimento da pesquisa.

## Primeiro Semestre de 2015

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |  |

A tabela acima demonstra a programação do primeiro semestre de 2015.

## Segundo Semestre de 2015

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Mês**  **Tarefas** | **JUN** | **JUL** | **AGO** | **SET** | **OUT** | **NOV** | **DEZ** |
| Estudo das Api’s das redes sociais e das tecnologias envolvidas | X |  |  |  |  |  |  |
| Desenvolvimento de protótipos | X |  |  |  |  |  |  |
| Desenvolvimento da Aplicação | X | X | X |  |  |  |  |
| Atualização da pesquisa |  |  | X | X |  |  |  |
| Análise e discussão de resultados |  |  |  | X |  |  |  |
| Pré-banca |  |  |  | X |  |  |  |
| Redação final do TCC |  |  |  |  | X | X |  |
| Defesa pública |  |  |  |  |  | X |  |
| Acertos finais para capa dura |  |  |  |  |  | X | X |
| Entrega da capa dura |  |  |  |  |  |  | X |

A tabela acima demonstra a programação do segundo semestre de 2015.

## Orçamento

A tabela abaixo tem como finalidade demonstrar as despesas deste projeto:

|  |  |  |  |
| --- | --- | --- | --- |
| ORÇAMENTO DETALHADO DO PROJETO | | | |
| 1. RECURSOS MATERIAIS | | | |
| 1.1 Material Permanente: (equipamentos, livros, máquina fotográfica e gravadores, equipamentos de informática, softwares, etc.). | | | |
| Descrição do Material | Quantidade | Valor (unidade – em reais) | Total R$ |
| Livro - MongoDB The Definitive Guide | 1 | 140,32 | 140,32 |
| Livro - Getting Started Whith Storm | 1 | 63,76 | 63,76 |
| Livro - Cloud Computing Theory And Pratice | 1 | 177,01 | 177,01 |
| Livro - NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence | 1 | 103,45 | 103,45 |
| Livro - Learning JavaScript | 1 | 111,61 | 111,61 |
| Livro - Web Development with Node & Express | 1 | 78,00 | 78,00 |
| Livro - Aplicações web real-time com Node.js | 1 | 59,90 | 59,90 |
| Livro - Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB | 1 | 198,00 | 198,00 |
| Livro - Data Mining Concepts and Techniques | 1 | 180,00 | 180,00 |
| Livro - Mining the Social Web | 1 | 114,00 | 114,00 |
| Livro - REST API:Design Rulebook | 1 | 51,00 | 51,00 |
| Subtotal | - | - | 1217,05 |
| 1.2 Material de Consumo: (Papéis necessários para impressões, cartuchos de tinta para impressora, etc.) | | | |
| Descrição do Material | Quantidade | Valor (unidade – em reais) | Total R$ |
| Papéis para impressão | 100 | 0,10 | 10,00 |
| Subtotal | - | - | 10,00 |
| 2. Serviços: (cópias, encadernações, impressos gráficos, despesas de locomoção e estadia, etc.) | | | |
| Descrição do Material | Quantidade | Valor (unidade – em reais) | Total R$ |
| Impressões | 10 | 3,00 | 30,00 |
| Encadernação | 2 | 5,00 | 10,00 |
| Subtotal | - | - | 40,00 |
| 3. RESERVA TÉCNICA/ Despesas Operacionais (10% no total do dispêndio) | | | |
| TOTAL | - | - | 1267,05 |

O orçamento acima foi aprovado por todas as partes envolvidas neste projeto

# REFERÊNCIAS

ALECRIM, E. **O que é Big Data?**. Disponível em: <http://www.infowester.com/big-data.php>. Acessado em: 31/05/2015.

ANDRIOLE, S. **Unstructured Data: The Other Side of Analytics**. Disponível em: <http://www.forbes.com/sites/steveandriole/2015/03/05/the-other-side-of-analytics/2/>. Acessado em 07/05/2015.

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

APACHE MAVEN PROJECT. **Maven**. Disponível em: <https://maven.apache.org/what-is-maven.html>. Acessado em 25/07/2015.

BROWN, E. **Web** **Development with Node & Express**. Sebastopol: O’Reilly, 2014.

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.

FRANÇA, T. C.; FARIA, F. F de; RANGEL, F. M.; FARIAS, C. M. de; OLIVEIRA, J. **Big Social Data: Princípios sobre Coleta, Tratamento e Análise de Dados Sociais**. Disponível em: <http://www.inf.ufpr.br/sbbd-sbsc2014/sbbd/proceedings/artigos/pdfs/127.pdf>. Acessado em 05/05/2015.

HAN, J; KAMBER, M; PEI, J. **Data Mining Concepts and Techniques**. 3. Ed. Waltham: Morgan Kaufmann, 2012.

LAVELLE, S.; LESSER, E.; SHOCKLEY, R.; HOPKINS, M. S.; KRUSCHWITZ, N. **Big Data, Analytics and the Path From Insights to Value**. Disponível em: <http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/>. Acessado em 05/05/2015.

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

MAGNO, A. **Globo Bootstrap**. Disponível em:

<http://blog.alexandremagno.net/2012/08/globo-boostrap/>. Acessado em 19/04/2015.

MARAS, J. **Secrets of the JavaScriptNinja – The JavaScript ecosystem**. <http://java.dzone.com/articles/secrets-javascript-ninja>. Acessado em 13/05/2015.

MARCONI, M. A; LAKATOS, E. M. **Fundamentos de metodologia científica**. 6. ed. São Paulo: Atlas, 2009.

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

MASSÉ, M. **REST API:** Design Rulebook**.** Sebastopol: O’Reilly, 2011.

MCAFEE, A.; BRYNJOLFSSON, E. **Big Data: The Management Revolution**. Disponível em: <http://www.researchgate.net/profile/Erik\_Brynjolfsson2/publication/232279314\_Big\_data\_the\_management\_revolution/links/53ecf40e0cf23733e804e561.pdf>. Acessado em 04/05/2015.

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.

MOZILLA DEVELOPER NETWORK. **CSS3**. Disponível em: <https://developer.mozilla.org/en-US/docs/Web/CSS/CSS3>. Acessado em: 05/06/2015.

MURTY, J. **Programming Amazon Web Services:** S3, EC2, SQS, FPS, and SimpleDB**.** Sebastopol: O’Reilly, 2008.

NAGARAJAN, M.; SHETH A.; VELMURUGAN, S. **Citizen Sensor Data Mining, Social Media Analytics and Development Centric Web Applications**. Disponível em: <http://knoesis.wright.edu/library/download/tr27-sheth1.pdf >. Acessado em 05/05/2015.

NODEBR. **Primeiros passos com Express e Node.js**. Disponível em: <http://nodebr.com/primeiros-passos-com-express-em-node-js/>. Acessado em: 24/05/2015.

OLIVEIRA, S. L. de. **Metodologia científica aplicada ao direito**. São Paulo: Pioneira Thomson Learning, 2002.

OLIVEIRA, S. L. de. **Tratado de metodologia científica**: projetos de pesquisa, TGI, TCC, monografia, dissertação e teses. 2. ed., quarta reimpressão. São Paulo: Pioneira Thomson Learning, 2002.

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.

PEREIRA, C. R. **Aplicações web real-time com Node.js.** São Paulo: Casa do Código, 2013.

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.

RUSSEL, M. A. **Mining the Social Web**. 2. Ed. Sebastopol: O’Reilly, 2013.

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

SCUP IDEAS. **O voto conectado**. Disponível em:

< http://scupideas.scup.com/ideas/eleicoes-2014/>. Acessado em 06/05/2015.

SILVA, L. T. da; FERREIRA JUNIOR, A. B. **Marketing político e sua importância através das mídias sociais**. Disponível em: <http://periodicos.ufpb.br/ojs/index.php/tematica/article/viewFile/21645/11972>. Acessado em 06/05/2015.

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

SOUBRA, D. **The 3Vs that define Big Data**.Disponível em: <http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data>. Acessado em: 14/05/2015.

SVENNERBERG, G. **Beginning Google Maps API 3**. New York: Apress, 2010.

TWITTER4J. **Twitter4j - A Java library for the Twitter API**. Disponível em: <http://twitter4j.org/en/index.html>. Acessado em 30/05/2015.

TWITTER DEVELOPERS. **The Stream APIs Overview**. Disponível em: <https://dev.twitter.com/streaming/overview>. Acessado em 29/08/2015.

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

W3C. **A vocabulary and associated APIs for HTML and XHTML**. Disponível em: <http://www.w3.org/TR/html/introduction.html#introduction>. Acessado em 19/04/2015.

W3C. **A history of HTML**. Disponível em: <http://www.w3.org/People/Raggett/book4/ch02.html>. Acessado em 26/04/2015.

W3C. **Open Web Plataform**. Disponível em: <http://www.w3.org/wiki/Open\_Web\_Platform>. Acessado em 26/05/2015.

1. A escola e o futuro da biblioteca de pesquisa. [↑](#footnote-ref-1)
2. Os usuários podem realmente absorver dados nas taxas de hoje? Amanhã? [↑](#footnote-ref-2)
3. Byte – unidadede medida computacional. [↑](#footnote-ref-3)
4. Paginação por demanda controlada por aplicativo para visualização externa. [↑](#footnote-ref-4)
5. Explorar visualmente conjuntos de dados gigabyte em tempo real. [↑](#footnote-ref-5)
6. Megabyte - Unidadede medida computacional. [↑](#footnote-ref-6)
7. Gigabyte - Unidadede medida computacional. [↑](#footnote-ref-7)
8. Quanta informação? [↑](#footnote-ref-8)
9. Exabyte - Unidadede medida computacional. [↑](#footnote-ref-9)
10. Modelos de fator dinâmico de Big Data para medição e previsão macroeconômica. [↑](#footnote-ref-10)
11. Gestão de dados 3D: controlar o volume de dados, velocidade e variedade. [↑](#footnote-ref-11)
12. A expansão do universo digital: uma previsão de crescimento da informação Worldwide até 2010. [↑](#footnote-ref-12)
13. Computação Big-Data: Criando avanços revolucionários no comércio, ciência e sociedade. [↑](#footnote-ref-13)
14. Dados, dados em toda parte. [↑](#footnote-ref-14)
15. Big data: a próxima fronteira para a inovação, concorrência e produtividade. [↑](#footnote-ref-15)
16. Terabyte – Unidade de medida computacional. [↑](#footnote-ref-16)
17. Termo para definir o conceito de onde ocorre o armazenamento em um banco de dados de informações relativas às atividades de uma organização. [↑](#footnote-ref-17)
18. Questões críticas para Big Data em Informação, Comunicação e Sociedade. [↑](#footnote-ref-18)
19. Termo inglês usado para designar o pensamento ou gosto atual da maioria da população. [↑](#footnote-ref-19)
20. Conjunto de instruções e padrões de programação para acesso a um aplicativo baseado na Web. [↑](#footnote-ref-20)
21. Solução de integração de sistemas, bom como na comunicação entre aplicações diferentes. [↑](#footnote-ref-21)
22. Tecnologia de processamento de requisições HTTP. [↑](#footnote-ref-22)
23. Arquivo compactado usado para distribuir um conjunto de classes ou um aplicativo Java, entre outros. [↑](#footnote-ref-23)
24. Protocolo aberto que permite uma autorização segura com um método simples e padronizado para aplicações Web , *mobile* e *desktop*. [↑](#footnote-ref-24)
25. Abreviação de GNU zip, um Software Livre de compressão sem perda de dados. [↑](#footnote-ref-25)
26. API baseada em HTTP de baixo nível para a obtenção de dados do Facebook. [↑](#footnote-ref-26)
27. Relacionamento entre dois ou mais computadores para que estes trabalhem de maneira conjunta no intuito de processar uma tarefa. [↑](#footnote-ref-27)
28. *Java Virtual Machine* - Máquina Virtual Java. [↑](#footnote-ref-28)
29. *American Standard Code for Information Interchange* - Codificação usada para representar textos em computadores, equipamentos de comunicação, entre outros dispositivos que trabalham com texto. [↑](#footnote-ref-29)
30. *Structured Query Language* - Linguagem de consulta estruturada, usada em bancos de dados. [↑](#footnote-ref-30)
31. Linguagem baseada em *scripts*. [↑](#footnote-ref-31)
32. Sistema operacional. [↑](#footnote-ref-32)
33. Banco de dados criado pelo Google. [↑](#footnote-ref-33)
34. Serviço de banco de dados NoSQL. [↑](#footnote-ref-34)
35. *Extensible Markup Linguage* - Linguagem de Marcação Extensiva. [↑](#footnote-ref-35)
36. *JavaScript Object Notation* - Notação de Objetos *JavaScript*. [↑](#footnote-ref-36)
37. Termo da computação que significa uma lista que armazena uma coleção de elementos. [↑](#footnote-ref-37)
38. Sistema que relaciona dois ou mais computadores para que estes trabalhem de maneira conjunta no intuito de processar uma tarefa. [↑](#footnote-ref-38)
39. Input/Output – Entrada/Saída. [↑](#footnote-ref-39)
40. Gargalos. [↑](#footnote-ref-40)
41. Abstração que une códigos comuns entre vários projetos de software provendo uma funcionalidade genérica. [↑](#footnote-ref-41)
42. Empresa multinacional de comércio eletrônico. [↑](#footnote-ref-42)
43. Organização europeia de pesquisas nucleares. [↑](#footnote-ref-43)
44. Ligações, referências. [↑](#footnote-ref-44)
45. *Cascading Style Sheets*, formata a informação entregue pelo HTML [↑](#footnote-ref-45)
46. Linguagem de programação interpretada pelo navegador. [↑](#footnote-ref-46)
47. Toda parte incluída na apresentação visual de um site. [↑](#footnote-ref-47)
48. S*oftware* com a finalidade de gerenciar diferentes versões no desenvolvimento de um projeto. [↑](#footnote-ref-48)