**TRABALHO DE GRADUAÇÃO**

# VISUALIZAÇÃO ANALÍTICA DO SENTIMENTO NEGATIVO EM OPINIÕES ACERCA DE SERVIÇOS DE TELECOMUNICAÇÕES EMITIDAS POR USUÁRIOS DO TWITTER

**Leandro Claudino**

**Pablo Piorno Baltoré**

**Brasília, Julho de 2015**

**UNIVERSIDADE DE BRASILIA**

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASILIA

Faculdade de Tecnologia

TRABALHO DE GRADUAÇÃO

# VISUALIZAÇÃO ANALÍTICA DO SENTIMENTO NEGATIVO EM OPINIÕES ACERCA DE SERVIÇOS DE TELECOMUNICAÇÕES EMITIDAS POR USUÁRIOS DO TWITTER

**Leandro Claudino**

**Pablo Piorno Baltoré**

*Relatório submetido ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Engenheiro de Redes de Comunicações*

**Banca Examinadora**

Prof. Dr. Rafael Timóteo de Sousa Jr, ENE/UnB
*Orientador*

Dr. Fábio Mesquita Buiati, PPGEE/UnB
*Examinador Interno*

Prof. MSc. Fábio Lúcio Lopes Mendonça
*Examinador Externo*

**Dedicatória(s)**

*A Deus, sempre bom e fiel e a minha família, base do que sou.*

*Leandro Claudino*

*Aos familiares e amigos que me apoiaram e estiveram sempre ao meu lado, aos professores e mestres por tudo e a Deus.*

*Pablo Piorno Baltoré*

**Agradecimentos**

*Agradeço à minha família que sempre me apoiou e incentivou, desde antes da minha entrada na UnB até os dias de hoje. Ao meu pai e minha madrasta, Paulo e Dacy, que com suas firmes cobranças e seu amor, me guiaram para um grande crescimento como pessoa e como profissional. E a minha amada mãe, Sandra, que mesmo afastada pela distância, é o meu incentivo diário para vencer os obstáculos encontrados.*

*Ao meu grande amigo, Pablo, por ter sempre me apoiado com muita irmandade, por todo companheirismo nessa graduação e nesse projeto, me ensinando muito com toda sua humildade e otimismo. E claro, sem esquecer jamais, dos momentos memoráveis e cheios de felicidade vividos na UnB ao longo desses anos todos.*

*Ao Bruno, parceiro inestimável de projeto e de vida, por suas brilhantes ideias e proatividade nesse trabalho e em todos os semestres de UnB.*

*À Anna Carolina, por toda sua ajuda, disposição, simpatia, alegria e eficiência ao longo desses vastos semestres.*

*Aos professores, que são os principais responsáveis pelo meu crescimento acadêmico e profissional e também ao meu orientador, Rafael de Souza, por seu auxílio para a melhor execução desse trabalho.*

*Agradeço também ao Pódion, em especial ao Ismael e à Marlise, por toda confiança em meu trabalho, que me fez crescer como profissional e a amar a arte de ensinar/aprender.*

*Leandro Claudino*

*Agradeço primeiramente a minha família que sempre me apoiou com amor e carinho. Aos meus amigos que tornaram mais agradáveis cada momento de estudo ou trabalho, levando alegria e motivação para o cotidiano acadêmico. A todos os professores que se dedicaram e tiveram a paciência de ensinar a caminhar profissionalmente para quem sou hoje. Aos meus companheiros de labuta que souberam me apoiar nas horas de desespero e noites sem dormir. E a Rejane que foi minha companheira e desde o começo do projeto, me incentivou a fazer tudo da melhor forma possível. Em especial, gostaria de agradecer à Anna Carolina que sempre nos prestou apoio pedagógico, filosófico, emocional, psicológico, físico e técnico e orientou em todos os campos do conhecimento.*

*Pablo Piorno Baltoré*

**RESUMO**

Este trabalho objetiva realizar o tratamento de grande volume de dados extraídos da rede social Twitter, de modo a revelar padrões na quantidade de reclamações dos usuários de serviços relacionados a telecomunicações.

Cria-se um ambiente de visualização em tempo real dos resultados obtidos a partir de um sistema web com uma interface acolhedora. Desenvolve-se com o software Gephi gráficos relacionados a grandes volumes de dados que auxiliam a interpretação humana e a visão global da informação gerada a partir do estudo de Big Data e mineração de dados em redes sociais.

A partir dos dados coletados no Twitter, foi gerado um estudo sobre a percepção por parte dos usuários dos serviços de telecomunicações das principais empresas atuantes no mercado brasileiro. Este estudo ilustra quantitativamente e de forma comprovável, através de um tratamento de inteligência artificial, uma análise do possível sentimento de insatisfação vivenciado por alguns usuários, não sendo, no entanto, capaz de acertar na totalidade dos casos sua percepção.

**ABSTRACT**

This study aims to carry out the treatment of large volume of data extracted from the social network Twitter, to reveal patterns in the number of complaints from users of services related to telecommunications.

It creates a display environment in real time the results obtained from a web system with a warm interface. It develops with the graphics software Gephi related to large volumes of data that help human interpretation and the global vision of information generated from the study of Big Data and data mining on social networks.

From the data collected in Twitter, a study was generated on the perception of the users of telecommunications services of the leading companies in the Brazilian market. This study illustrates quantitatively and verifiable manner through a treatment artificial intelligence, analysis of the possible feeling of dissatisfaction experienced by some users, which does not, however, able to hit in all cases perception.

# SUMÁRIO

# LISTA DE FIGURAS

# LISTA DE SÍMBOLOS

**Siglas**

| | |
|---|---|
| API | Aplication Programming Interface |
| BI | Bussiness Inteligence |
| CPU | Central Processing Unit |
| DB | Data Base |
| DN | DataNode |
| E/S | Entrada e saída |
| EC2 | Amazon Elastic Computer Cloud |
| GB | Gigabyte |
| GPS | Global Positioning System |
| HDD | Hard Drive Disk |
| HDFS | Hadoop Distributed File System |
| HiveQL | Hive Query Language |
| HTTP | Hypertext Transfer Protocol |
| IBM | International Business Machine |
| IP | Internet Protocol |
| JSON | Javascript Object Notation |
| KB | Kilobyte |
| MB | Megabyte |
| NCSA | National Center for Supercomputing Applications |
| NoSQL | Not-only SQL |
| OLAP | On-line Analytical Processing |
| ONU | Organização das Nações Unidas |

| | |
|---|---|
| PDA | Personal Digital Assistant |
| PHP | Hypertext Preprocessor |
| RAM | Random Access Memory |
| RESTful | Representational State Transfer |
| SAC | Serviço de atendimento ao consumidor |
| SNN | Secondary NameNode |
| SQL | Structured Query Language |
| SSD | Solid State Disk |
| TCP | Transmission Control Protocol |
| TI | Tecnologia da Informação |
| TT | Task-Tracker |
| WEB | Word Wide Web |
| XML | Extensible Markup Language |

# 1 INTRODUÇÃO

Grandes corporações devem possuir meios para identificação de falhas em seus produtos e serviços. Não é desconhecido o fato que há uma corrida de quem oferece o melhor serviço. A solução de problemas com maior rapidez e agilidade é considerada um fator primordial nesta corrida.

Atualmente as empresas utilizam de meios tradicionais para descobrir falhas, onde basicamente tem se por esperar o cliente efetuar a reclamação. Um exemplo básico disso é a ligação para determinada operadora para relato de possível problema. Para uma corporação ser totalmente proativa em termos de identificação de falhas, ela deve investir muito em equipamentos de gerência, muitas vezes ineficientes. Desse modo, ela consegue aferir possíveis problemas em determinado serviço, sem que o cliente entre em contato, ou mais, que nem saiba que existe um problema.

Com o advento da Internet e das redes sociais, os clientes começaram a buscar novos meios para comunicar falhas, e, além disso, reclamar da empresa ou do serviço prestado. Ou seja, os usuários viram nas redes sociais um ambiente livre que podem publicar uma reclamação — ou elogio — de serviços que estão contratando. Com isso, podemos utilizar dessas informações públicas para traçar e mapear possíveis falhas. Quanto mais informação estiver disponível, mais segmentado pode ser esse mapeamento e o mais importante, mais próximas da realidade podem ser essas informações.

## 1.1 MOTIVAÇÃO

Este estudo objetiva contribuir com a visualização dos dados obtidos a partir de um estudo previamente realizado descrito nos trabalhos relacionados através de uma ferramenta experimental que identifique possíveis problemas nos serviços prestados pelas operadoras de telecomunicação no Brasil, a partir de uma análise utilizando inteligência artificial, em tempo real, baseado no monitoramento da rede social Twitter.

Neste trabalho, consideramos a dificuldade em extrair informações de um ambiente de armazenamento de dados escalável e eficiente após a coleta de grandes massas de dados. Como obter informações valiosas e principalmente apresentá-las graficamente para acelerar o entendimento daqueles que possuem acesso a essa informação. Ou seja, como visualizar de forma clara e suscinta dados de Big Data.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo Geral

Criar um ambiente de visualização gráfica que expresse a opinião dos usuários da rede social *Twitter* acerca das principais operadoras de telefonia, internet e TV por assinatura do Brasil e analisar padrões observados na relação de consumo desses serviços de telecomunicações.

### 1.2.2 Objetivos Específicos

- O primeiro objetivo específico é definir as ferramentas necessárias para a obtenção de gráficos de Big Data, fazendo uma revisão teórica abordando as principais tecnologias vigentes que facilitaram a realização deste trabalho.

- O segundo objetivo específico tem por intento implementar as funções de mapeamento e redução nos dados anteriormente obtidos, agregando valor e estruturando esse grande volume de dados.

- O terceiro objetivo específico visa a criação de um sistema web capaz de exportar dados provenientes de um ambiente de dados escalável e eficiente. E ainda utilizar o software Gephi para a confecção de novos gráficos com recurso visual mais aprimorado.

## 1.3 METODOLOGIA

O trabalho foi realizado a partir de uma pesquisa bibliográfica e estudo da temática de interesse. Após o estudo das técnicas almejadas para realização deste projeto, partimos para a parte prática que foi dividida em quatro segmentos: segmentação dos dados; criação do sistema web; utilização do Gephi e por fim, análise dos resultados.

Este trabalho se relaciona com o projeto final de graduação do curso de Engenharia de Redes de Comunicação da Universidade de Brasília escrito pelo aluno Bruno Monteiro Pimentel de Alencar no primeiro semestre do ano de dois mil e quinze, dando continuidade no que diz respeito aos dados obtidos através de seu trabalho de captura, coleta e tratamento de dados da rede social Twitter. Estes dois projetos correlacionados foram desenvolvidos a partir do mesmo escopo original.

Em seu trabalho, o aluno Bruno desenvolveu um ambiente de captura de tweets dos usuários de empresas de telecomunicações atuantes no mercado e ainda, em parceria conosco, desenvolveu um sistema de tratamento dos dados coletados, segmentando-os e classificando-os de acordo com os requisitos estipulados para os nossos trabalhos. Essa parceria resultou também em um sistema web que realiza o tratamento dos dados, atualizando os gráficos em tempo real. Fator este decisivo para concluirmos o presente trabalho dentro das expectativas e objetivos traçados, considerando-o bem sucedido.

## 1.4 JUSTIFICATIVA

Com o crescente uso dos dispositivos móveis conectados à internet e com o advento das redes sociais, um novo canal de comunicação entre empresas e clientes tem ganhado importância para as organizações. Nessa perspectiva, entender como coletar informações de valor significativo a partir das redes sociais torna-se o grande desafio do futuro, podendo significar um diferencial entre as empresas que puderem realizar essa tarefa de forma bem sucedida. Além disso, o volume de dados crescente com a era digital torna esse desafio mais complicado. As grandes empresas, portanto, tem despendido recursos financeiros e humanos na tentativa de dominar tecnologia

capaz de facilitar a análise das redes sociais e auxiliar na compreensão das demandas de seus clientes.

Ao desenvolver esse trabalho, temos a pretensão de contribuir com o desenvolvimento de ferramentas que possam propiciar uma análise quantitativa de dados provenientes das redes sociais que possibilite às organizações se comunicarem de forma mais íntima e informal com os usuários de seus produtos ou serviços.

## 1.5 ORGANIZAÇÃO DO TRABALHO

O primeiro capítulo é uma introdução, que expõe os motivos que respaldam esse estudo e descreve os objetivos almejados no projeto. Apresenta uma síntese sobre os assuntos ratados servindo de guia de fácil acesso àqueles que procuram literatura relacionada com o tema.

No próximo capítulo, será mostrada uma breve revisão teórica sobre todas as tecnologias utilizadas no desenvolvimento desse trabalho, servindo de base conceitual para o andamento da pesquisa realizada.

O terceiro capítulo trata do ambiente de visualização edificado em torno dos dados obtidos da rede social Twitter, previamente coletados e tratados no sentido de segmenta-los segundo a orientação dos motivadores desse projeto.

No quarto capítulo mostramos as ferramentas utilizadas, como hardwares e softwares, na implementação do projeto. Esse capítulo serve de guia para futuros interessados em desenvolver parcial ou inteiramente tecnologias para análise dos dados provenientes de redes sociais.

No capítulo seguinte, apresentamos nossa interpretação sobre os resultados obtidos levando à conclusão no sexto capítulo.

# 2 CONCEITOS E FUNDAMENTAÇÃO TEÓRICA

*Uma breve revisão teórica sobre o que é o Big Data, sua classificação e as tecnologias aplicáveis a esse tipo de dado.*

## 2.1 BIG DATA

A era digital que vivemos nos dias de hoje introduz o desafio de desenvolvermos a capacidade de processar, armazenar e lidar com volumes cada vez maiores de dados. No âmbito empresarial, solucionar este desafio pode ser o diferencial em tornar empresas bem sucedidas.

Com o frequente uso de redes sociais e um número cada vez maior de dispositivos conectados à internet, principalmente os dispositivos móveis, os clientes das empresas e os usuários de diversos serviços passaram a produzir relevante informação sobre fatores que podem influenciar na percepção que uma empresa tem a respeitos de seus produtos, negócios, consumidores e muito mais. Por exemplo, os consumidores tem procurado registrar suas insatisfações não só nos órgãos responsáveis por fiscalização e regulamentação, mas também em redes sociais como o Facebook e principalmente o Twitter e as empresas que assumiram uma postura de se relacionar com os seus clientes através das redes sociais se tornaram mais bem vistas por seus usuários por estreitarem a relação com eles.

Não obstante a isso, as organizações também tem participado do aumento na geração de dados com seus próprios requisitos internos. A quantidade de dados que as empresas geram em resposta a seus processos internos é cada vez maior, são logs transacionais, medições de sensores, planilhas de gastos e receitas, apresentações, pesquisas internas, treinamentos, entre outras tantas fontes de dados. Com isso, a era do Big Data já chega de forma irreversível, tornando a informação cada vez mais valorosa para as empresas e para o desenvolvimento das ações comerciais. A necessidade de adaptação a esse cenário dinâmico se tornou, portanto, ponto chave no ramo empresarial, abrindo uma grande oportunidade de faturamento no ramo de tecnologia de informação visto que as empresas que exploram essa nova fonte de informação advinda dos clientes saem em vantagem em relação as que não exploram ainda tais fontes de dados.

Big Data engloba o tratamento de um grande volume de dados, que muitas vezes se encontra desestruturado ou até mesmo em um formato dificilmente entendido pelas máquinas, como por exemplo, em linguagem humana, e que precisam ser processados em uma velocidade muito aquém da velocidade disponível computacionalmente para que a informação gerada nessa análise não se torne obsoleta.

### 2.1.1 Os 4 V’s

Alguns autores, como [Berman, J. J. 2013], tratam Big Data como um conjunto de dados que possuem 3 V’s: Volume, Variedade e Velocidade. Este assunto causa certa divergência de opiniões visto que outros autores, como [Sathi, A. 2012] citam

ainda um quarto V: Veracidade ou Valor. Esta segunda definição foi a escolhida para o embasamento teórico deste trabalho e o conceito de cada V será apresentado a seguir. A figura 1 mostra o conceito de Big Data sustentado pelos 4 V's.

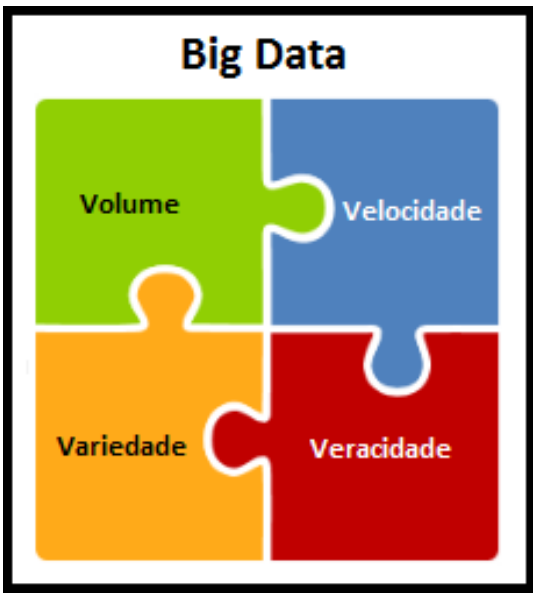


**Figura 1: Os 4 V's do Big Data**

### 2.1.1.1 Volume

O volume, em Big Data, representa a quantidade de bytes a serem tratados. Essa medida vem crescendo de forma exorbitante devido ao uso desenfreado de tecnologias, tanto internas quanto externas, nas organizações. De acordo com Fortune [Fortune 2012, p. 163], nós criamos até o ano de 2003, 5 exabytes de dados digitais. Já em 2011, a mesma quantidade era criada a cada dois dias. Em 2013, o período esperado era de apenas 10 minutos.

O volume de dados produzidos tende a continuar sua escalada crescente, pois a cada dia surgem novas tecnologias que se utilizam internet como novos serviços em nichos mercadológicos. Por exemplo, a quantidade de dados provenientes da localização de dispositivos móveis tem sido cada vez maior devido a novos aplicativos que são tendências na atualidade, como serviços de aluguel de carros executivos ou rastreamentos de animais de estimação.

Segundo a assessoria da empresa de telecomunicações Oi, em matéria publicada pelo portal www.convergenciadigital.uol.com.br, no dia 14/07/2014, foi registrado que:

"nos 31 dias de competição, a Oi registrou 74 terabytes de dados trafegados nas redes de mídia e informática providas para a FIFA e utilizadas pelos cerca de 20 mil profissionais de mídia de 113 países credenciados para cobrir o evento. O volume de dados foi o triplo do registrado na Copa do Mundo da FIFA África do Sul 2010 (cerca de 25 terabytes) e equivale a mais de 80 milhões de fotos em resolução normal ou cerca de 20 milhões de fotos em alta resolução.(...) O maior volume de tráfego de dados (3,2 terabytes) foi registrado no dia 23/06, quando aconteceram outros quatro jogos da fase de grupos (Brasil x Camarões, Austrália X Espanha, Chile x Holanda, e Croácia x México)."

Há uma década, as organizações normalmente contavam seu armazenamento de dados para análise de infraestrutura em terabytes. Eles já graduaram para aplicações que requerem armazenamento em petabytes [Sathi 2012]. Essa enorme quantidade de dados necessita um processamento diferenciado, baseado em computação distribuída ou em nuvem. Assumindo que esta grande quantidade de dados seja superior ao que os bancos de dados convencionais conseguem trabalhar, o processamento dos dados se reduz basicamente a arquiteturas de processos paralelos ou soluções de Apache Hadoop desenvolvido para paralelizar os processamento de dados em diversos nós, com objetivo de aumentar o poder computacional e diminuir a latência [Hurwitz et al. 2013, p. 112].

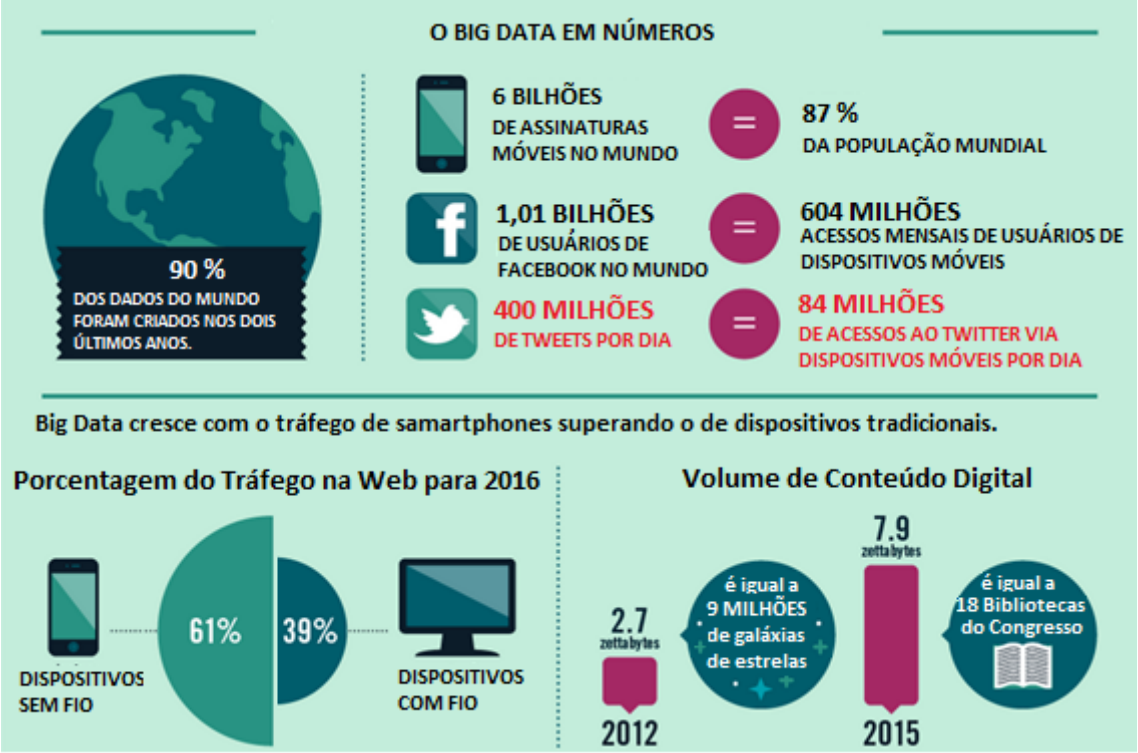


**Figura 2: O Big Data em números (Modificado)**

### 2.1.1.2 Variedade

Grande parte dos dados obtidos pelas fontes de informações supracitadas vem desestruturada, isso significa que eles não se encontram tabelados de forma clara que máquinas possam ler como em bancos de dados comuns. O formato desses dados não estruturados ou semiestruturados representa outro desafio computacional que dever ser enfrentado com o uso de tecnologias de Big Data.

São considerados dados estruturados os que vêm em formatos específicos, que podem ser facilmente tabelados em um banco de dados convencional. São exemplos de dados estruturados os nomes, datas de nascimento, endereços, operações e pontos de fidelidade. Já os dados semiestruturados ou não estruturados são aqueles que não se apresentam facilmente interpretada por computadores, como por exemplo, imagens, opiniões de produtos, valor emocional de seres humanos, ou seja, toda aquela informação que não é clara e depende, portanto, da interpretação humana. Ensinar uma máquina a coletar, entender e interpretar essas informações é o desafio da variedade em Big Data.

#### 2.1.1.3 Velocidade

Há dois aspectos referentes à velocidade no mundo do Big Data. Um que trata do fluxo de dados, ou seja, da vazão e o outro representa a latência[Sathi 2012]. Muitos softwares que buscam novos padrões de informações em grandes volumes de dados não conseguem processar em tempo real as informações que possuem ciclos menores de execução por causa de sua demora na manipulação das informações. Quando se finda a execução dos procedimentos de análise o resultado acaba se tornando sem valor e vira descarte por estar desatualizado [PETRY; VILICIC, 2013, p.74-75].

#### 2.1.1.4 Veracidade

Muitos dos dados são provenientes de fontes não confiáveis, no qual, podem sofrer diversos tipos de problemas. Um deles é de precisão, na qual os dados confrontados nada tem a haver com o objetivo ou público desejado. A veracidade representa ambos a credibilidade dos dados como a adequação para dado público. Segundo Sathi [Sathi 2012], a coerência dos dados para certo audiência não necessariamente sirva para outra. Devemos pensar na adequação e quanto a veracidade pode ser compartilhada com o público específico.

Ao contrário dos dados internos cuidadosamente regulados, a maioria Big Data vem de fontes fora de nosso controle e, portanto, sofre com exatidão ou precisão significativa problemas. Veracidade representa tanto a credibilidade da fonte de dados, bem como a adequação dos dados para o público-alvo. Vamos começar com a fonte de credibilidade. Se uma organização estavam a recolher produto informações de terceiros e oferecê-lo aos seus empregados de contact center para apoiar as consultas dos clientes, os dados teriam de ser rastreados para a exatidão fonte e credibilidade. Caso contrário, os contact centers podem acabar recomendando ofertas competitivas que podem marginalizar ofertas e reduzir as oportunidades de receita.

Um monte de respostas de mídia social em campanhas poderiam estar vindo de um pequeno número de empregados últimos insatisfeitos ou pessoas empregadas pela concorrência, a postar comentários negativos. Por exemplo, vamos supor que "like" em um produto significa clientes satisfeitos. E se o "como" foi colocado por um terceiro? Temos também de pensar sobre platéia adequação e quanta verdade pode ser compartilhados com um público específico. A veracidade dos dados criados dentro de uma organização pode ser assumido como sendo, pelo menos, bem intencionado. No entanto, alguns dos dados internos podem não estar disponíveis para uma comunicação mais ampla. Por exemplo, se o cliente serviço forneceu insumos para engenharia de deficiências de produtos como visto na pontos de contato do cliente, esses dados devem ser compartilhados selectivo, com um know necessidade-to-base. Outros dados podem ser compartilhados apenas com os clientes que têm contratos válidos ou outros pré-requisitos.

### 2.1.2 Personagens do Big Data e os Tipos de Dados

Big Data é formado por dois fatores distintos, os personagens e os tipos de dados. Esses fatores são complementares e tornam o conceito de Big Data desafiador. Os personagens do Big Data tratam das fontes de informação enquanto os tipos de dados tratam do tipo de informação fornecida e da forma como esta é obtida.

Conforme os dispositivos móveis se tornaram mais acessíveis, as pessoas deixaram de ser apenas consumidores passando a serem também parte de uma cadeia de produtos. Isso pode até soar de forma estranha, mas nas redes sociais, os consumidores são na verdade o produto. Informações sobre os usuários de redes sociais famosas como Facebook ou Twitter são vendidas para grandes empresas. Outro produto de consumo nas redes é o acesso aos usuários por meio de propagandas pagas, ou por meio de postagens alavancadas por recursos financeiros. A prospecção de clientes, no entanto, está longe de ser o único interesse das empresas nas redes sociais, grande parte das organizações já operam nas redes sociais de forma ativa, com o uso das referidas propagandas, de forma reativa, respondendo às reclamações de consumidores, e muitas vezes, solucionando diversos problemas apresentados e de forma passiva, coletando informações sobre os consumidores ou possíveis mercados para uso futuro.

Outros personagens do Big Data são as organizações públicas e privadas. Ambas participam da massificação dos dados digitais, contribuindo para esse fenômeno mundial da atualidade chamado de Big Data. A figura 3 ajuda a entender o papel de cada personagem do Big Data e suas respectivas contribuições para o cenário de grande volume de dados e de informação disponível.

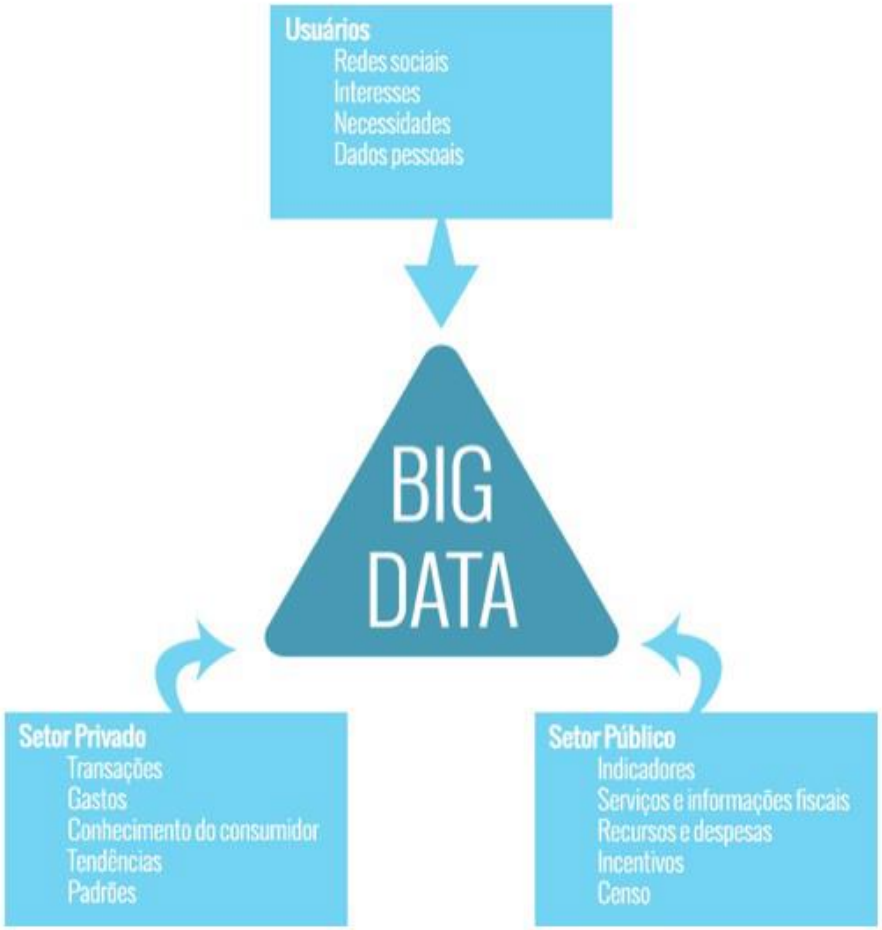


**Figura 3: Personagens do Big Data**

Já quanto aos tipos de dados, vamos definir algumas classificações segundo óticas diferentes. Podemos classificar os dados, ou melhor dizendo, cada conjunto de

dados de acordo com algumas característcas comuns: quanto à origem dos dados, os dados podem ser classificados como dados de máquina ou como dados humanos; quanto à forma, os dados podem ser estruturados, semiestruturados ou ainda desestruturados. A seguir, serão detalhadas essas quatro classificações dos dados.

- Dados de máquinas: são dados gerados como resultado de algum processo realizado por computadores, sem a intervenção direta de seres humanos. Esse tipo de dados pode ser usado pelas máquinas para controle, recuperação ou análise por máquinas e computadores e pode ainda servir de objeto para análise humana como controles de auditoria ou análise de sistemas.

- Dados humanos: é a nova fonte de informação de maior valia para as organizações, tanto as públicas quanto as privadas, e é o objeto principal de estudo deste trabalho. Trata-se dos dados gerados devido à interação humana com novas tecnologias e ajudam a entender o ambiente onde se encontra o indivíduo. Informações sobre idade, localização, gostos, gênero, opinião, posicionamento político, relacionamentos pessoais e de consumo, entre tantas outras informações, são encontradas com facilidade nas redes sociais. O interesse comercial por essa grande variedade de informação está na propecção de novos clientes e não obstante a isso, na manutenção do mercado consumidor existente. Ao se utilizarem das redes sociais, as pessoas utilizam uma linguagem praticamente nova, distante da norma culta, cheia de abreviações ou repetições excessivas de letras que denotam ênfase e com uso de figuras como emoticons. Ou seja, ao tentar suprir a ineficiência de expressar sentimentos na linguagem escrita, os usuários apelam para o uso de recursos estilísticos presentes somente na linguagem humana por meio digital. Isso torna a interpretação de mensagens simples para humanos e extremamente complicada para máquinas.

- Dados estruturados: é o tipo de dados menos frequente dentro do enorme volume de dados gerados todos os dias, mas ao mesmo tempo, é o tipo de dados mais aproveitado até então. Especialistas no campo de análise de dados concordam que esse tipo de dados compõem cerca de 20% de todo os dados que são gerados no mundo, e são com eles que normalmente lidamos [HURWITZ et al., 2013]. Os dados estruturados, como o nome indica, são dados que obedecem uma certa estrutura, ou seja, são conjuntos de dados que são produzidos já obedecendo uma estrutura determinada que pode facilmente ser entendida pelo computador por se encaixar naturalmente na estrutura de qualquer banco de dados. Esses dados são agrupados em tabelas e matrizes relacionais em campos de tamanho fixo, que facultam o processamento e interpretação dos dados pelas máquinas.

- Dados semiestruturados ou desestruturados: segundo Berman, 2013, são os dados compostos por conteúdo que não esteja organizado em forma de matrizes de atributos e valores. Os dados desestruturados, são dados obtidos sem uma organização matricial ou tabelada. Essa desorganização representa um enorme desafio computacional, pois a informação mais importante pode estar intrínseca à interpretação humana daqueles dados, por exemplo, uma imagem é entendida facilmente por um ser humano, mas para uma máquina que tem a limitada visão binária da mesma imagem pode ser uma complicadíssima missão. Dentre os dados tidos como desestruturados, podemos citar imagens, vídeos, dados de sensores e, como motivação principal para esse trabalho, interação humana. Esse tipo de dados

representa a maior parte dos dados gerados, como vimos no tópico anterior, aproximadamente 80% do total de dados produzidos no mundo digital.


**Figura 4: Exemplos de Dados Desestruturados**

Alguns autores ainda citam que a maioria desses dados desestruturados é pelo menos semiestruturados. Segundo Franks [FRANKS, 2012], dados semiestruturados ou multiestruturados são aqueles que possuem um fluxo lógico e um formato que pode ser entendido, porém não intuitivo para os humanos. E ainda ler um dado semiestruturado para analisa-lo não é simples como especificar um tamanho fixo de formato. Parar ler é necessário empregar regras complexas que determinam dinamicamente como proceder depois de ler cada pedaço de informação [FRANKS, 2012]. Um exemplo de um dado semiestruturado são os logs provindos de sistemas WEB:


**Figura 5: Exemplos de Dados Semiestruturados**

## 2.2 DATA MINING

O Data Mining, ou Mineração de Dados, é um processo de extração das informações advinda de um enorme banco de dados sem o seu prévio conhecimento para que através do que foi colhido decisões sejam tomadas. Essa metodologia será utilizada em diversas áreas que informações (conhecimento), como as empresas, as indústrias e as instituições de pesquisa. Data Mining pode ser definido como um processo automatizado de captura e análise de enormes conjuntos de dados para obter um significado onde poderá ser descrito características já consolidadas ou então para "prever" tendências para o futuro.

Com o crescimento do armazenamento de informações, há a necessidade de buscar conhecimento para aproveitar a riqueza contida neste conjunto de dados. A mineração de dados ingressa no cenário como um dos principais processos na transformação de uma grande quantidade de dados operacionais em conhecimento. Segundo Han e Kamber [HAN; KAMBER, 2006], mineração de dados pode ser entendida como o "resultado na evolução natural da tecnologia da informação", e em uma perspectiva de data warehouse, uma fase de processamento analítico online (OLAP – On-line analytical processing). Porém, vai muito além do âmbito restrito de estilo sumarizado de processamento analítico, a mineração de dados provê técnicas avançadas de análise de dados.

São vários os casos de uso do Data Mining em várias empresas como Wal-Mart, Bank of America, Google e  nas áreas de Telecomunicações, Edeucação e Saúde. Um caso interessante e bem divulgado foi o da cadeia americana Wal-Mart, que identificou um hábito curioso de seus consumidores.  Ao procurar eventuais relações entre os volume de vendas e os dias da semana, em dados passados, a aplicação do Data Mining apontou que, às Sextas-Feiras, as vendas de cerveja cresciam nas mesma proporção das vendas de fraldas de bebês. Qual a relação de cerveja com fralda? Crianças bebendo cerveja? Não. Uma investigação mais detalhada mostrou que quando os pais iam comprar as fraldas para suas crianças, eles já aproveitavam para abastecer seu estoque de cerveja para o final de semana.Em casos como esse, fica claro que a obtenção de informações desse tipo são muito relevantes para a estratégia em que uma empresa pode tomar com relação aos hábitos de seus clientes, melhorando vendas, atendimento, marketing e satisfação de seus usuários, por exemplo.

Com o crescimento do armazenamento de informações, há a necessidade de buscar conhecimento para aproveitar a riqueza contida neste conjunto de dados. A mineração de dados ingressa no cenário como um dos principais processos na transformação de uma grande quantidade de dados operacionais em conhecimento.

Para usufruir dessa riqueza é necessário organizar de uma forma que seja possível identificar padrões e analisar com ajuda de filtros, agrupando os dados numa variedade de modos [O'NEIL; QUASS, 1997].

Na mineração de dados temos a própria extração dos dados a serem estudados e deles tenta-se encontrar padrões e regularidades através de ferramentas (programas). Quando se têm padrões fortes pode-se dizer que boas predições serão adquiridas, porém podemos encontrar alguns problemas, como, por exemplo, a maior parte dos padrões podem não ser interessantes, esses padrões podem não ser exatos e ainda tais dados podem estar truncados.

Podemos então, classificar de uma maneira geral, as análises possíveis sobre certo conjunto de dados em algumas partes.

Na primeira parte temos a Amostragem que engloba a parte de Detecção de Desvios e a Análise de Desvios. Nessa análise tem um objetivo simples: encontrar comportamentos que fogem muito a situação comum, fazendo com que a confiabilidade da amostragem e dos resultados obtidos aumente. Para isso ela conta com a tarefa de detecção de desvios onde são encontradas informações que não irão obedecer ao comportamento do modelo de dados em geral. Tais dados dissonantes pode receber algum tratamento em especial ou simplesmente serem ignorados/descartados antes do início da mineração.

Semelhante a isso temos a Análise de Desvios, porém o que distingue é que a medida de comparação é que irá definir se o dado salta do normal comportamento estabelecido no modelo em estudo. Exemplificando essa tarefa, podemos fazer a associação dela à fatura de cartão de crédito gerada para um determinado mês. Se ela fugir do padrão de uso para um determinado usuário em quesitos como compra, valor gasto e o tipo de produto, isso pode ser considerado um indicio de fraude, clonagem do cartão ou qualquer outra ocorrência associada.

Na segunda parte olha-se para as Tarefas Descritivas, onde uma varredura será feita para buscar um estabelecimento de associações, relações, descrição e caracterização do modelo, bem como coletar informações relevantes com difícil visualização. Ela pode ser iniciada sem que haja, obrigatoriamente, uma ideia ou hipótese já formada anteriormente.

Para analisar melhor essa parte, temos algumas tarefas que compõem tais Tarefas Descritivas. Para começar, temos a Classificação, cujo objetivo se encontra na organização de categorias com os dados em classes já anteriormente definidas em conformidade com a similaridade de certa característica dos dados. Pode-se imaginar em comparação a esse cenário uma farmácia, onde seus produtos são organizados medicamentos industrializados ou manipulados, de usos terapêuticos, objetos e instrumentos de higiene, toalete e perfumaria.

Em seguida, vem a parte das Associações, que vem a identificar grupos de ações que ocorrem em conjunto ou de alguma forma condicionada. Dessa forma encontramos a associações e os relacionamentos entre os itens. Esses resultados são expressos em forma de regras de associação, de forma que se um certo conjunto de itens A ocorre e o conjunto B é uma base dados para o conjunto A, pode significar que B tenderá a ocorrer. Uma quantidade de regras de associação poderá ser gerada com a analise de associações em um banco de dados, sendo que algumas podem não ter um valor tão significativo de importância, pois não seriam de frequente ocorrência nos dados em questão. Com isso alguns parâmetros devem ser criados para determinar quais regras são interessantes, não ocasionando uma perda de tempo com informações de baixa importância. Semelhante a essa tarefa, temos outra que é a de Agrupamento, cuja diferença se dá, pois na classificação, as classes são definidas de uma forma prévia, já que no agrupamento as classes serão definidas junto com a tarefa de forma que o estabelecimento do conjunto de atributos é que deverão direcionar tal categorização.

Conforme a similaridade desses atributos direcionadores é que os grupos serão formados. Temos uma tarefa também que é a Descrição, que como o próprio nome diz,

fará uma descrição em texto de um conjunto de singularidades vistas com maior frequência para um determinado evento. É comumente utilizada para traçar perfis de comportamento. No caso do cartão de crédito, podemos observar que a maior ocorrência dos envolvidos nesses casos de fraudes são homens, entre 25 e 40 anos, já com um possível curso superior e com bom nível de instrução, por exemplo. Logo o Descrição já vem a fornecer uma informação bem mais valiosa sobre o conjunto de dados em questão. Nas tarefas de Detecção de Sequencias há o estabelecimento de relacionamentos de tempo entre fatos. Um exemplo prático é de pessoas que compram um celular, mesmo que ao adquirir o produto tenham já comprado uma capa protetora para o mesmo, voltara logo para comprar outra, ou para colocar uma película protetora no mesmo. Por último, temos a tarefa de Segmentação que objetiva subdividir um conjunto de dados em menores conjuntos de forma a agrupá-los de acordo com alguma característica. Pode-se pensar como exemplo em uma segmentação dos consumidores por região e sexo e depois buscar associações nesses dados. Com isso, poderíamos achar possíveis diferenças de compras entres regiões e entre homens e mulheres.

Já para a terceira e ultima parte, temos a parte mais valorosa das tarefas. Isso por que nessa parte teremos as tarefas que retornam resultados de previsão que podem mudar rumos de negócios e tomadas de decisões em grandes organizações, por exemplo. As Tarefas de Prognóstico visam através de uma análise retornar um valor ou possível comportamento ou até mesmo estimar valores desconhecidos, embasando-se nãos dados colhidos na análise descritiva.

Nela, duas ordens de tarefas se apresentam. A primeira é a Estimação que estabelecerá valores desconhecidos a partir de valores conhecidos. Ao analisar, por exemplo, o padrão de vida, a forma de despesas e a idade de uma pessoa, podemos estimar a quantidade de filhos daquela pessoa, que ate então seria um número desconhecido. Já na segunda, a Predição é o processo de previamente determinar um valor em um ponto futuro baseando-se em valores conhecidos. De acordo com o nível de escolaridade, o emprego em exercício e qual o ramo de trabalho, poderá ser estipulado qual será a evolução salarial em certo período de tempo.

Resumindo, podemos ver na tabela a seguir, quais são as tarefas principais para o Data Mining:

| TAREFAS | DESCRIÇÃO |
|---|---|
| Classificação | Categoriza dados a partir de um modelo de certo tipo que possa ser aplicado nos dados especificados |
| Associação | Determina quais itens podem ocorrer ao mesmo tempo |
| Sumarização | Através de métodos específicos, encontra uma descrição resumida para um subconjunto de dados |
| Estimativa | Definirá um valor para certa variável sem conhecimento |
| Segmentação | Partição de dados diferentes em subconjuntos de dados (mais) homogêneos |

**Tabela 1: Tarefas Principais para o Data Mining**

## 2.3 DATA WAREHOUSE

O Data Warehouse (armazaém de dados, ou depósito de dados no Brasil) é uma arquitetura de banco de dados volumoso no qual armazena-se informações de uma organização de modo consolidado. Nascida na década de 80 como um conceito acadêmico, ganhou destaque no mundo empresarial com o crescimento do volume de dados tendo como promessa analisar todos os dados da organização, além disso, favorece os relatórios e facilita a tomada de decisões.

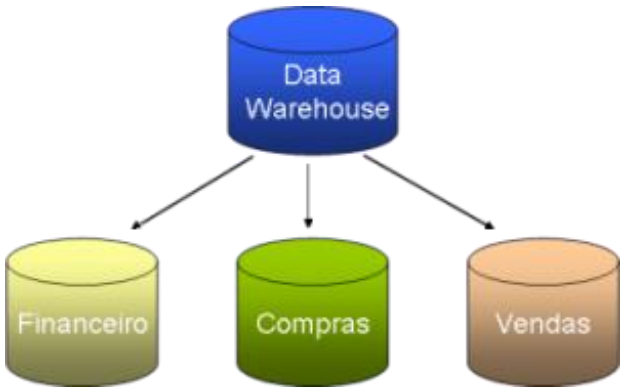


**Figura 6: Utilidades do Data Warehouse**

Os dados armazenados nos Data Warehouses são provenientes de sistemas indepentendes que produzem dados operacionais: sistema financeiro, de clientes, de pagamentos, de vendas, e assim por diante. Estes dados possuem grande riqueza de informação e são utilizados no dia a dia das operações dos sistemas de informática por isso é importante que a informação seja acessada rapidamente para a análise. Através

desssa análise possibilitada pelas séries históricas de dados armazanados em um depósito de dados desses, decisões para tomadas de ações presentes podem ser feitas e também uma previsão de certos eventos futuros. Por definição, os dados em um Data Warehouse são voláteis, ou seja, aqueles dados que não mudam, exceto quando correções em dados previamente já inseridos no banco necessitam de correção. Os dados ficam acessíveis apenas para leitura, não podendo serem modificados.

Uma ferramenta clássica para uma boa exploração de Data Warehouse é a Online Analytical Processing (OLAP), que permite manipular e analisar um certo volume de dados sob variáveis perspectivas.

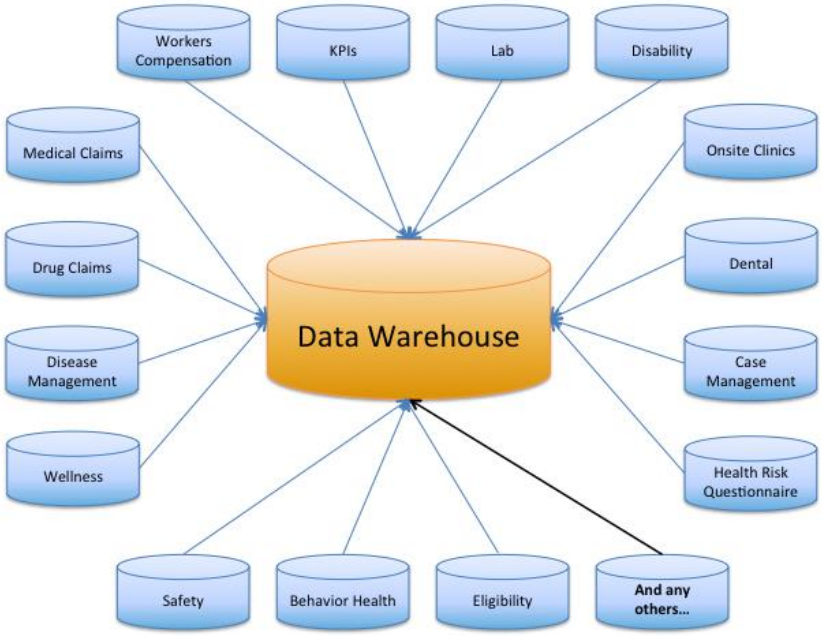


**Figura 7: Variáveis do Data Warehouse**

Foi ficando cada vez mais claro que sem uma fonte de dados central integrados, pode ser extremamente difícil traçar uma visão significativa e, portanto, tomar medidas eficazes.

No entanto, o Data Warehouse veio perdendo campo no que se diz respeito às tomadas de decisões e previsões de eventos devido a um aumento significativo da quantidade de dados, e a necessidade de respostas praticamente em tempo real, foi necessária a adaptação das arquiteturas, surgindo a arquitetura de Big Data.

## 2.4 HADOOP E HDFS

Informação é algo absolutamente importante nos dias atuais, uma vez que as grandes decisões são tomadas a partir dela. Os sistemas de armazenamento de dados, por sua vez, possuem restrições, assim como um arquivo qualquer, e por isso a sua atualização é inevitável e basicamente obrigatória, afinal, tudo evolui, inclusive a maneira de se gerenciar dados.

Quando se enfrenta um grande volume de dados, as abordagens tradicionais de gerenciamento e armazenamento de dados, como data storage e data warehouse, se tornam obsoletas e ineficientes, pois não permitem o processamento dos dados de forma a extrair informação suficiente para suprir às necessidades impostas pela globalização. Diante desse desafio, uma nova tecnologia precisou ser criada, para assegurar o suporte das novas demandas de arquivamento e processamento. Neste cenário, surge uma nova ferramenta capaz de suportar o imenso volume de dados de forma rápida, segura e com um preço acessível. Esse é o projeto Hadoop, o qual se destaca como a mais eficiente solução para a situação atual, na qual o processamento de informação e seu devido armazenamento precisam evoluir constantemente em termos de volume, variedade e velocidade.

O projeto Hadoop foi criado e desenvolvido a partir de 2006 pela empresa Yahoo!, mais precisamente por Doug Cutting, o qual ocupava cargo de executivo à época. O seu objetivo primordial, era encontrar valor na massiva quantidade de dados extraídos de sua ferramenta de busca na web. Posteriormente, este se tornou um software aberto gerenciado pela Apache Software Foundation, o que o torna muito mais barato e competitivo que outros softwares, visto que hoje existem diversas empresas que procuram desenvolver produtos com objetivo semelhante.

O projeto Hadoop é tido como o mais importante projeto da Apache, juntamente com outros softwares desenhados a partir do mesmo para as funções mais específicas do gerenciamento de Big Data, como Hive(Data Warehouse distribuído), HDFS (Sistema de arquivos distribuídos do Hadoop), HBASE (Banco de dados distribuído) , entre outros. O sucesso do Apache Hadoop foi confirmado ao ser usado por outras grandes empresas, entre as quais podem ser citadas Adobe, Ebay, Facebook, Google, IBM, Spotfy, Twitter e a própia Yahoo! entre muitas outras. A lista completa das empresas que se utilizam dessa tecnologia pode ser encontrada no sítio eletrônico [http://wiki.apache.org/hadoop/PoweredBy]. A partir desta constatação , podemos verificar que o projeto Hadoop realmente inovou na capacidade de processamento e que possibilitou não apenas o próprio reconhecimento como uma ferramenta essencial ao funcionamento dos atuais sistemas de processamento de Big Data, mas também a criação de outros projetos setorizados que formam a visão geral da capacidade do Hadoop e que podem ser usados em acordo com outros programas conforme a necessidade de uma computação distribuída.

Hadoop é, portanto, um framework para processamento distribuído e particionado que surgiu a partir da necessidade de se suprir várias demandas de aspectos inerentes ao suporte a Big Data. Entre as lacunas verificadas anteriormente, pode-se perceber o preenchimento destas através da oferta de um serviço de processamento distribuído em uma nuvem de computadores, ou cluster, com altíssima escalabilidade, capaz de transformar em informações os grandes volumes de dados desestruturados obtidos a partir de diversas fontes - no caso desse trabalho, os dados coletados através da rede social Twitter, que são do tipo desestruturados e semi estruturados – e ainda garantir resiliência, ou seja, a capacidade de se adaptar a novos cenários inerentes a falhas de sistema, como por exemplo a interrupção no serviço de energia ou um erro advindo do hardware sem prejuízo do andamento do projeto.

Percebe-se até então que esta é de fato uma solução bastante interessante para um mundo tão interligado e que precisa evoluir a cada segundo, mas essa não é a única necessidade atual. Além de ofertar a disseminação de informação de qualidade através

de uma plataforma rápida e segura, existe um outro aspecto absolutamente importante, inclusive por impactar na viabilidade do projeto, o custo. E por este motivo, o Hadoop é tão efetivo, pois além de ser eficiente no que se propõe, ele ainda demanda recursos acessíveis, o que o torna ainda mais interessante. A plataforma Hadoop, por trabalhar com computação distribuída, não necessita de máquinas ultra poderosas, nem de super processadores, tampouco de memórias primárias ou secundárias de ordem animalesca. O uso de máquinas commodities torna viável o projeto de gerenciamento de Big Data para praticamente qualquer empresa, pois a computação é completamente distribuída aproveitando os recursos disponíveis. Evidente que quanto maior for o processamento das máquinas e suas memórias primárias, menor será o tempo gasto no processamento do grande volume de dados relacionado ao Big Data, no entanto, o fato de o processamento poder ser realizado, e bem realizado, por commodities torna o projeto Hadoop uma solução barata e portanto, extremamente viável para empresas de qualquer porte. Este trabalho exemplifica como empresas de segmentos variados podem aproveitar a tecnologia disponível para extrair das redes sociais valiosíssima informação a respeito da imagem da empresa em relação ao seus clientes: como seus clientes veem a empresa; quais suas maiores queixas e quais seus maiores elogios.

Grande parte da dificuldade do processamento de Big Data está no armazenamento dos dados na memória, principalmente no que diz respeito ao tempo de leitura e escrita destes em discos. O HDFS – Hadoop Distributed File System – é o componente do Hadoop responsável por este armazenamento de grandes volumes de dados em disco. Baseado na arquitetura master-slave, HDFS provê escalabilidade e disponibilidade para o processamento de Big Data a partir da construção de réplicas dos blocos tantas vezes quanto for necessário a partir de sua configuração, a qual pode ser facilmente alterada , inclusive para ajustar o número de réplicas às necessidades de processamento de forma adaptativa. O número de cópias do mesmo bloco é escolhido baseado no tradeoff entre processamento e segurança da informação, envolvendo disponibilidade e garantia contra falhas de software. Quanto maior o número de cópias, maior será a segurança, visto que o mesmo bloco estará armazenado em mais máquinas, e maior será o processamento necessário já que o mesmo bloco será processado várias vezes, aumentando o tempo caso seja mantido o número e a qualidade das máquinas. O HDFS, assim como o Hadoop, garante a escalabilidade do processo de armazenamento tanto para suprir uma falha ou ausência momentânea de uma das máquinas, como para se adaptar a um novo recurso disponível, por exemplo, uma máquina nova adicionada à nuvem ou outra que tenha deixado de funcionar e agora pode retornar a atividade.

O funcionamento do HDFS, citado anteriormente, é baseado na arquitetura master-slave e possui, como componentes principais, duas estruturas diferentes: a primeira, cujas responsabilidades são armazenar os metadados dos blocos e gerenciar o tráfego de dados na rede do cluster, é denominada Namenode; a outra, a qual é formada por componentes que realizam o armazenamento e processamento dos blocos em si, são chamadas Datanodes. A seguir estas estruturas serão abordadas de forma detalhada.

#### 2.4.1.1 Namenode

Namenode possui como primeira tarefa receber um arquivo de tamanho considerável e recortá-lo em partes menores, as quais são denominadas blocos. Estes

serão processados e armazenados por Datanodes. O Namenode atua, principalmente, como controlador de tráfego e armazenador de metadados sobre os blocos armazenados em cada Datanode. Em outras palavras, ele guarda o endereço de processamento de cada bloco criado a partir do arquivo principal para reconstruí-lo mais tarde, através do gerenciamento do acesso aos arquivos em relação a escrita, leitura, criação, destruição e replicação de arquivos e gerencia o tráfego na rede distribuindo os blocos a serem processados em cada Datanode. Essa distribuição leva em conta além do tráfego dos blocos na nuvem, a quantidade de CPU livre em cada máquina.

A soma de todos os arquivos presentes no cluster, ou nuvem, formam o chamado namespace. O Namenode é o responsável primário por organizar e gerenciar o namespace através das commodities presentes no cluster de computação distribuída.

Outra tarefa importante realizada pelo Namenode é se adaptar ao número de Datanodes disponíveis a cada instante, podendo adicioná-los ou retirá-los do cluster conforme ocorram demandas ou conforme a disponibilidade destes seja alterada. O gerenciamento se dá através da replicação dos arquivos nos Datanodes e pelo controle de disponibilidade destes. A replicação de arquivos de um mesmo bloco serve para assegurar a integridade do namespace em caso de falha de um Datanode, pois nenhum arquivo seria único e nem estaria presente em um único Datanode. O controle de disponibilidade dos Datanodes é realizado por mensagens de heartbeat, que são mensagens periódicas enviadas pelos Datanodes que carregam informações sobre o seu status de processamento.

Heartbeat é o conjunto de mensagens periódicas enviadas pelos datanodes para o namenode. As mensagens de heartbeat mostram ao namenode quais são os datanodes que estão disponíveis no cluster e qual a sua respectiva capacidade de processamento livre e questionam sobre suas próximas tarefas. O namenode responde as mensagens de heartbeat informando os Datanodes acerca de suas próximas tarefas. O conjunto de mensagens de heartbeat proporciona um feedback de qualidade a respeito da disponibilidade dos blocos e do espaço existente em cada um deles, desta maneira, as tarefas podem ser redistribuídas entre os dispositivos disponíveis a cada momento. Da mesma forma que podem ser acolhidos novos dispositivos no cluster, podem ser facilmente ignorados Datanodes que venham a se tornar indisponíveis por qualquer circunstância. Os Datanodes também enviam uma outra série de mensagens, estas com menor frequência, para os Namenodes: os block reports, que são mensagens contendo as informações sobre quais os blocos disponíveis no Datanode.

O uso dessa variedade de mensagens pelos Datanodes oferece ao Namenode uma visão ampla e globalizada a respeito do processo em todo o cluster, o que permite que o processo de dados seja mais efetivo em qualidade e tempo.

O Namenode possui uma importância fundamental ao guardar informações cruciais do sistema de arquivos processado no cluster. Obviamente, uma informação tão valiosa para o sistema não poderia, de forma alguma, ficar guardada em um único local, pois sujeitaria o sistema a uma falha pontual tornar-se crítica, tornando vulnerável todo o processo. O Namenode mantém todo seu estado em memória principal para obter melhor desempenho, pois esta memória é acessada mais rapidamente que a memória do tipo secundária, no entatanto, essa melhoria na performance pode resultar em um enorme risco no caso de falha, uma vez que a

memória principal não mantém a informação caso seja desligada, ou seja, é uma memória de acesso mais rápido que as secundárias, mas que é volátil enquanto as secundárias não o são. Para Impedir que o sistema se perca em caso de falha no Namenode, duas estruturas garantem a resiliência do sistema de arquivos HDFS, checkpoints e logs transacionais. Os checkpoints são criados pelo Namenode a partir dos checkpoints anteriores e dos logs transacionais que registram toda e qualquer transação que altere o namespace. A combinação dessas duas estruturas, guardadas em memórias secundárias e backups, possibilita a recuperação do sistema de arquivos em caso de falha. O registro das informações referentes a localização dos blocos disponíveis em cada Datanode não é guardada em memória secundária estável, dessa forma, em caso de falha, o Namenode espera pelas mensagens de block reports para organizar novamente a disponibilidade dos blocos nos Datanodes e retomar o processo de onde este foi interrompido.

#### 2.4.1.2 Datanode

O sistema HDFS trabalha com centenas ou milhares de Datanodes, enquanto somente um Namenode é utilizado. Cada Datanode é uma central de armazenamento de blocos obtidos de um arquivo muito grande, da ordem de Gigabytes ou Petabytes, por exemplo. Os Datanodes são os obedientes escravos da configuração master – slave que citamos anteriormente, sendo a mão de obra do trabalhoso processo de armazenamento de Big Data em computação distribuída.

Enquanto o Namenode tem um papel extremamente inteligente em gerenciar um imenso cluster de máquinas e controlar o tráfego da rede, os Datanodes são os responsáveis pelo trabalho massivo em garantir a resiliência do processo e a integridade dos dados.

Os Datanodes, conforme já foi explicado, informam o Namenode, através do heartbeat, o andamento de seus processos internos. Ao receber uma mensagem de heartbeat o Namenode entende que o Datanode que a enviou se mantém ativo, podendo ser demandado de novas tarefas. Na situação contrária, caso um Datanode deixe de enviar mensagens de heartbeat durante um longo período, o Namenode entende que aquele Datanode não está mais disponível no cluster ou que não há conexão entre o Datanode e o Namenode. As novas tarefas passam a ser redistribuídas entre os Datanodes disponíveis sem prejuízo para o processo geral. Quando o heartbeat resurge, ou um novo aparece, ele é adicionado ao cluster de forma completamente transparente para o usuário e para a aplicação. Quando surge um novo heartbeat, o Namenode adiciona o Datanode responsável por ele no cluster sem que o usuário possa sequer perceber essa alteração.

Outra função essencial em um sistema de computação distribuída e desempenhada pelo Datanode é garantir a integridade do sistema de arquivos distribuídos. Como cada bloco de dados é replicado em vários Datanodes conforme configuração previamente estabelecida, um mesmo bloco é processado paralelamente por diversos Datanodes. Cada Datanode processa um bloco independente dos demais que processam o mesmo bloco, dessa forma é possível detectar erros no processamento de blocos através do uso de checksum, que é uma soma verificadora dos dados presentes em um arquivo ou no caso bloco. Se houver divergências no checksum de um mesmo bloco processado em Datanodes diferentes haverá portanto, a indicação de que ocorreu um erro no processo em um dos dois blocos. Como no HDFS

o processamento de cada bloco ocorre em inúmeros nós, podemos descartar os blocos em que houve inconsistência na checagem, indicando que houve erro no processamento do bloco. O uso do checksum possibilita que um bloco seja descartado caso sua checagem seja diferente da de outros blocos iguais processados por vários Datanodes, pois é estatisticamente improvável que ocorra o mesmo erro no processamento do bloco em vários Datanodes tornando a probabilidade de acerto maior que de quando ocorre cada possível erro.

Com os avanços crescentes em Big Data, diversos avanços em diferentes setores ficam cada vez mais visíveis e com isso alavancando o uso de tal tecnologia.

Por meio de um bom entendimento em padrões de dados estruturados e não estruturados, grandes resultados podem ser apresentados gerando valiosas informações que podem melhorar diretamente os resultados de uma empresa ou negócio, fazendo com que ela se destaque de outra empresa. Com isso o Big Data vem sendo a ferramenta para o "entendimento" desses dados coletados para gerar uma mudança nos negócios.

Para tais entendimentos em Big Data são necessárias ferramentas tecnológicas para executar essa promissora tecnologia.

As tecnologias que sustentam Big Data podem ser analisadas sob duas óticas: as envolvidas com analytics, tendo Hadoop e MapReduce como nomes principais e as tecnologias de infraestrutura, que armazenam e processam os petabytes de dados. Neste aspecto, destacam-se os bancos de dados NoSQL (No, significa not only SQL). Por que estas tecnologias? Por que Big Data é a simples constatação prática que o imenso volume de dados gerados a cada dia excede a capacidade das tecnologias atuais de os tratarem adequadamente.

Falamos que as tecnologias atuais de tratamento de dados não são mais adequadas. Por que? Olhando o modelo relacional, proposto pelo pesquisador da IBM, Edgar F. Codd, em 1969. Quando foi proposto, a demanda era acessar dados estruturados, gerados pelos sistemas internos das corporações. Não foi desenhado para dados não estruturados (futurologia na época), nem para volumes na casa dos petabytes de dados (inimaginável na época) muito menos para os zetabytes que logo menos estarão ao nosso alcance. Precisava-se sim de um modelo que categorizasse e normalizasse dados com facilidade. E o modelo relacional foi muito bem sucedido nisso, tanto que é o modelo de dados mais usado atualmente.

Para tratar dados na escala de volume, variedade e velocidade do Big Data precisamos de outros modelos. Surgem os softwares de banco de dados NoSQL, desenhados para tratar imensos volumes de dados estruturados e não estruturados. Existem diversos modelos como sistemas colunares como o Big Table, usado internamente pelo Google (é a base de dados sob o Google App Engine),o modelo Key/value como DynamoDB da Amazon, o modelo "document database" baseado no conceito proposto pelo Lotus Notes da IBM e aplicado em softwares como MongoDB, e o modelo baseado em grafos como o Neo4j. Interessante lembrar que antes do modelo relacional já existia um software de banco dados que lidava com grandes volumes que é o IMS da IBM, modelo hierárquico, criado para suportar o projeto Apollo de conquista da Lua e que ainda hoje é base da maioria das transações financeiras que circulam pelo mundo.

Por outro lado, esta diversidade de alternativas demanda que os líderes dos projetos de Big Data escolham a mais adequada ou mesmo demandem mais de uma opção, de acordo com as necessidades específicas.

Depois da infraestrutura é necessário atenção aos componentes de analytics, pois estes é que transformam os dados em algo de valor para o negócio. Big Data Analytics não signfica eliminar os tradicionais sistemas de BI que existem hoje, mas pelo contrário, devem coexistir.

Aliás, ao lado destas alternativas surgem outras opções, como o uso de appliances, como o Netezza da IBM, que embarcam em um hardware adaptado todos os softwares necessários para criar projetos de Big Data.

### 2.4.2 MapReduce

O modelo de programação paralela para processos largamente distribuídos de grandes volumes de dados chamado de MapReduce foi proposto inicialmente em 2004 pela empresa Google. (Dean and Ghemawat 2008). Houve muitas iniciativas de implementações de tal tipo de programação para várias linguagens, no entanto, a mais conhecida e de grande importância para o seu entendimento nesse trabalho é a que consta no projeto Hadoop.(White 2012)

MapReduce veio com o intuito de tralhar problemas podendo particioná-los ou fragmentá-los em subproblemas através das funções *Map* e *Reduce* em um grande conjunto de dados que podem ser divididos para que, em paralelo, várias funções *Map*, possam ser executadas ao mesmo tempo.

Esse trabalho ele é realizado distribuindo o processamento para várias maquinas para que seja feito em um tempo aceitável, já que muitas vezes falamos de casas muito superiores a terabytes de dados. Tal distribuição será feita de forma que cada máquina processe uma quantidade de dados em paralelo, sendo que os conjuntos de dados designado a cada uma são diferentes entre si. Logo, cada componente será o responsável pela totalidade do processamento de um pequeno grupo de dados ao invés de ser o responsável por processar todos os dados de uma só vez em certa etapa computacional.

De acordo com Hurwitz [HURWITZ et al., 2013], a distribuição do trabalho deve ser realizada em paralelo por 3 razões:

• O processo deve poder se expandir e contrair automaticamente;

• O processo deve continuar mesmo com falhas na rede ou em sistemas individuais;

• Desenvolvedores que utilizam o MapReduce devem ser capazes de desenvolver serviços que são fáceis para outros desenvolvedores, até porque esta abordagem deve ser independente de onde os dados estão e aonde que serão processados.

A função básica de uma aplicação MapReduce consiste na divisão e no processamento através das funções *Map* e *Reduce*. Na função *Map* teremos o primeiro contato, ocorrendo a transformação das informações, fazendo um processamento em paralelo em diversas máquinas do *cluster* computacional de cada registro desses dados.

Já para a função *Reduce* será a responsável por fornecer um resultado final da aplicação executada.

O mapeamento das informações se tornou o coração da tecnologia de processamento de dados. Nisso, a função *Map* utiliza-se dos documentos de entrada para mapear pares (tuplas) < *chave, valor* >. No mapeamento, é então aplicada uma função em cada elemento de um conjunto de dados e tem como produto um novo conjunto sem modificar os dados originais. Esse novo conjunto é particionado e cada partição é agrupada e ordenada pela chave (tupla).

Já a função de redução, tem como entrada a saída das funções de *Map* e reduz o conjunto de dados. É executado uma vez para cada chave, em sequência ordenada, com o conjunto de valores que compartilham a mesma chave. No final do processo, a função retorna os valores baseado na tarefa que foi executada.

O conceito central do MapReduce é a possibilidade da utilização de algoritmos capazes de processar grandes quantidade de dados, já que as operações são independentes(HURWITZ et al., 2013).

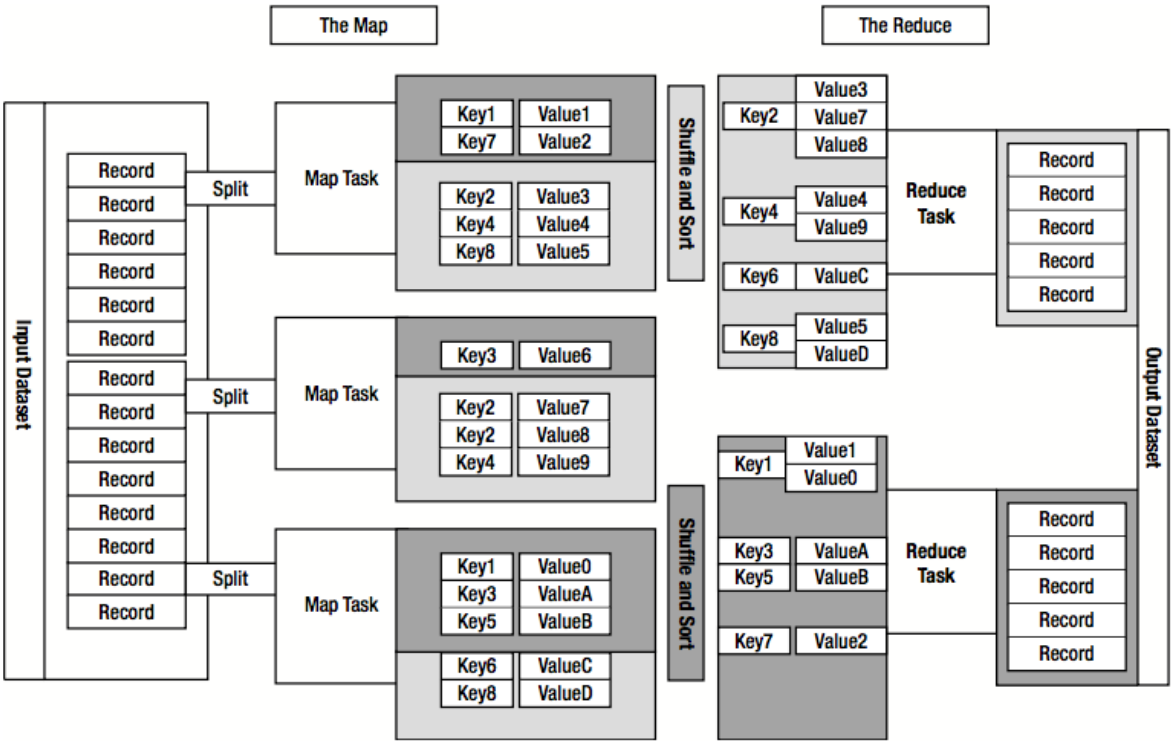


**Figura 8: Tarefas do MapReduce**

Fica claro na figura acima a separação das tarefas de *Map* e as tarefas de *Reduce*. Como já explicado anteriormente, os dados serão divididos em "pedaços" menos, depois processados e recolocados juntos ao final do processo. A função *Map* não poder ser executada sem uma ordem e em qualquer nó de uma máquina do cluster.

É importante ressaltar que alguns comportamentos do MapReduce dever ser tratados e vistos com uma maior atenção. A função de redução somente será iniciada quando a tarefa de mapeamento estiver completa totalmente [HURWITZ et al., 2013]

Portanto, nota-se que essas tarefas são prioridades com relação ao número de nós no cluster. Caso haja mais tarefas do que nós, as *Map Tasks* serão gerenciadas pelo framework até seu término. Em seguida, após o sucesso das *Map Tasks* obterem sucesso, as *Reduce Tasks* tomarão o mesmo caminho até o seu término por completo. Caso exista a necessidade de que ambos os processos sejam executados simultaneamente será necessário um mecanismo de sincronização. Nisso o framework irá entender em que função estará e guardará as informações do momento em que esta ocorrendo a execução e do que também esta a ser executado. Quando toda a função do mapeamento terminar inicia-se a função de redução enquanto todos os dados serão copiados e ordenados através da rede.

Para que um processamento eficiente ocorra, o código de mapeamento é relocado para os diversos nós em que os dados serão processados. O MapReduce oferece 2 processos que podem gerenciar os trabalhos:

• JobTracker (master): faz a gerência das tarefas, ou seja, provê controle e monitoramento do trabalho e a coordenação na distribuição das tarefas para os nós que executam o Task- Tracker;

• TaskTracker (slaves): gerencia a execução individual das funções map e reduce em um nó de um cluster.

Existe, geralmente, um processo JobTracker em um cluster, normalmente no nó master, e vários TaskTracker nos diversos nós slaves. No processo de DataNode [Venner seção 2.2] observa-se que o processo JobTracker é um ponto de falha crítico, diferentemente do Task-Tracker. A maioria das implementações do MapReduce possuem uma manipulação de erros robusta e tolerância a falhas. O framework é capaz de reconhecer alguma falha e criar a correção necessária [HURWITZ et al., 2013]

Uma vantagem de possuir esta arquitetura de processos é evidenciada por [Venner], onde cita que pode-se adicionar novos nós TaskTracker no cluster enquanto uma tarefa está sendo executada e o trabalho será dividido para essas novas máquinas.

Uma arquitetura baseada em racks é ideal para a construção de um cluster Hadoop, com cada Daemon essencial em um diferente rack.

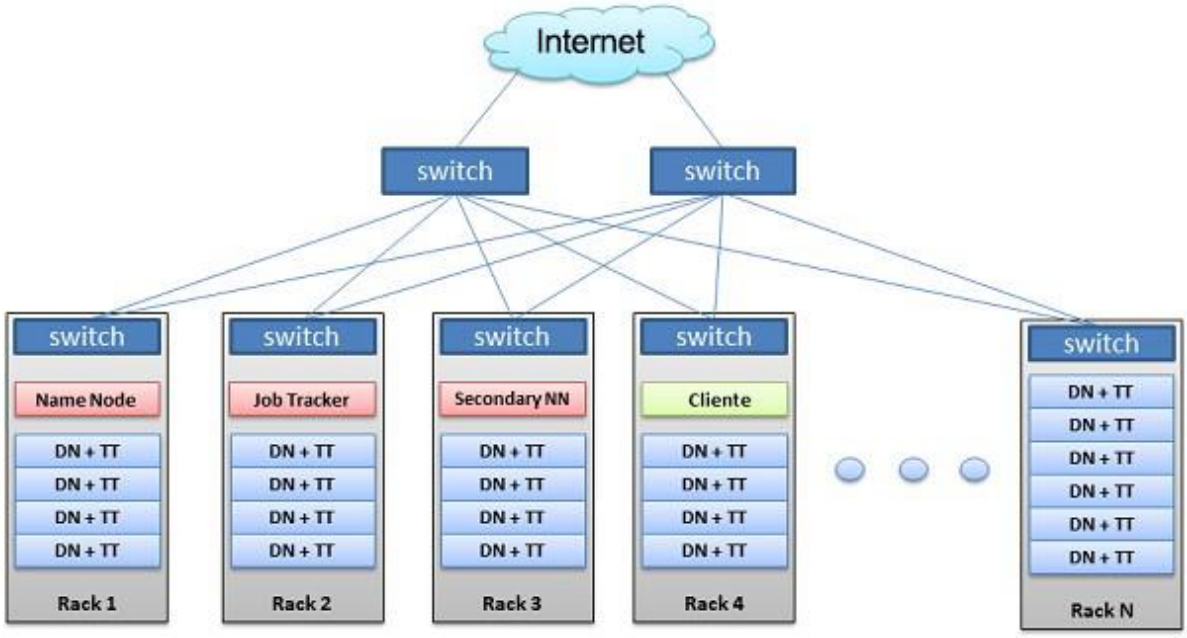


**Figura 9: Arquitetura de cluster Hadoop**

A figura Fig. 2.9 demonstra esta arquitetura. Cada rack possui um daemon vital (NameNode, JobTracker ou NameNode Secundário) ou o cliente independente, e seus DataNodes/Task-Trackers para armazenamento e processamento. Além disso, cada rack é interligado a dois switches, criando assim redundância caso um switch venha a falhar. A maioria dos servidores são nós escravos, com muito espaço em disco e com poder de processamento e de memória moderados. Algumas máquinas serão nós principais (master) que podem ter uma configuração diferente favorecendo memória e CPU e com menos espaço em disco.

Assim como acontece na comunicação entre cliente com o NameNode do HDFS, já mencionado, o mesmo é implementado no MapReduce. Uma porta TCP é definida para o JobTracker, que irá gerenciar os trabalhos.

Podemos avaliar então que o através do MapReduce podemos obter sucesso em muitos propósitos diferentes, devido a alguns fatores, como, por exemplo, a sua facilidade de utilização, mesmo para programadores iniciantes sem muita experiência em sistemas distribuídos e paralelos, já que o framework oculta os detalhes da paralelização, da distribuição dos dados, do balanceamentos das cargas e da tolerância a falhas. Outro fator também de que uma vasta gama de problemas podem ser facilmente expressados através do MapReduce, como os encontrados nos serviços de busca do Google, a ordenação de enormes quantias de dados para a mineração de dados, para aprendizado de máquinas e outros sistemas. Vê-se também que o MapReduce é prático para a escalabilidade em clusters computacionais abrangendo uma quantia de máquinas considerável, tornando o uso dos recursos muito eficientes, montando assim um ambiente eficiente para o processamento de muitos problemas computacionais que envolve grandes quantidades de dados.

## 2.5 GEPHI

A apresentação do grande volume de dados precisa ser clara para que possamos interpretá-la de maneira usual. Os recursos gráficos ajudam nesse ponto pois tornam mais didáticas as informações obtidas a partir dos processos descritos anteriormente. No âmbito de criar formas mais lúdicas de se apresentar a informação extraída dos dados coletados e tratados, o Gephi se mostra uma excelente ferramenta. Através de grafos, o software é capaz de apresentar uma enorme quantidade de informação útil ao usuário, o qual passa a ter uma visão ampla e sistêmica do processo globalizado.

Um grafo G é definido como sendo um par ordenado (V,E), onde V é um conjunto e E uma relação binária sobre V. Os elementos de V são denominados de vértices ou pontos ou nós, e os pares ordenados de E são denominados de arestas ou linhas ou arcos do grafo. [RABUSKE, 1992]

Um grafo pode ser dirigido ou não dirigido. Um grafo é dito dirigido se suas arestas possuem orientação. [RABUSKE, 1992]

Gephi é um software de código aberto para visualização e exploração de grafos dos tipos dinâmicos ou hierárquicos e de redes e sistemas complexos. Essas redes podem ser de computadores, de relacionamentos ou ainda de informação. Escrito em Java, é uma plataforma suportada pelos principais sistemas operacionais – Windows, Linux e Mac OS X – e que é distribuída gratuitamente.

O Gephi permite a visualisação e exploração de grandes redes em tempo real através de uma ferramenta de renderização tridimensional. Sua arquitetura é leve, flexível e adaptada a multi-tarefas possibilitando trabalhar com complexos conjntos de dados e gerando resultados visualmente interessantes e repletos de informação útil para o usuário. Permite ainda, exportar dados de redes e trabalhá-los espacialmente, com filtragens, navegação, manipulação e agrupamento dos dados. [Bastian, 2009]. Seguem nas figuras 10 e 11 abaixo, alguns exemplos: na figura 10, agrupamento; na figura 11, filtragem.

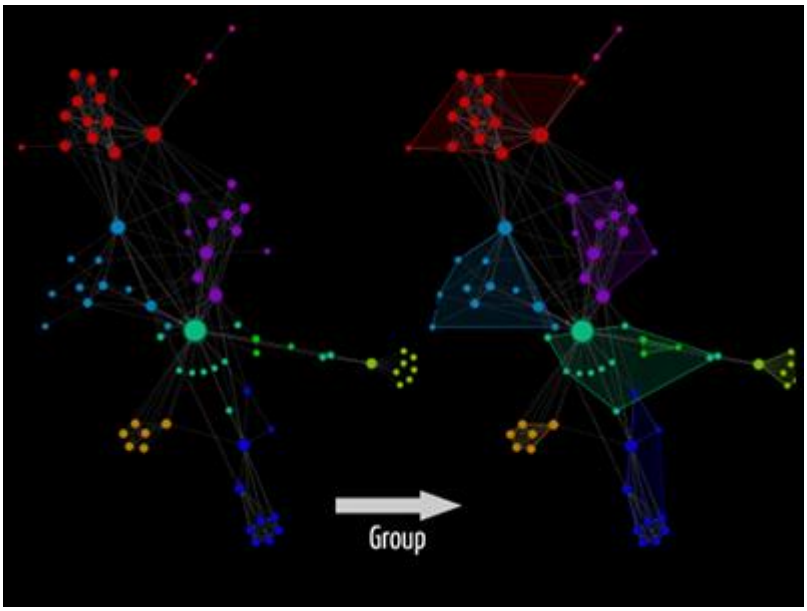


**Figura 10: Agrupamento de nós no Gephi**

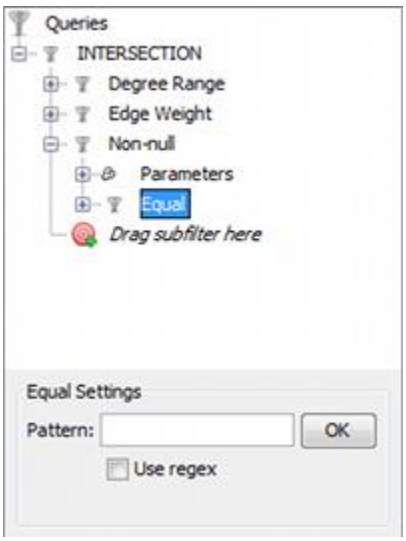


**Figura 11: Filtragem no Gephi**

Os maiores diferenciais do Gephi estão presentes na tratabilidade proporcionada aos dados, podendo serem visualizados e explorados em tempo real, terem seu layout de apresentação alterado para melhorar a visibilidade da informação almejada, seus parâmetros modificados de métrica, análise de redes dinâmicas, criação de cartografias, criação de grafos hierárquicos e com clusterização, ou agrupamento, filtragem dinâmica, centrados no usuário, modular e com uma vasta gama de *plugins* disponíveis.

A figura 12, apresenta uma aplicação em tempo real, usada para descobrir padrões de comportamento em grande grafos iteragindo visualmente através de filtros dinâmicos e ferramentas avançadas que possibilitam uma manipulação significativa, podendo suportar redes de até 50.000 nós e 1.000.000 de laços [Bastian, 2009].

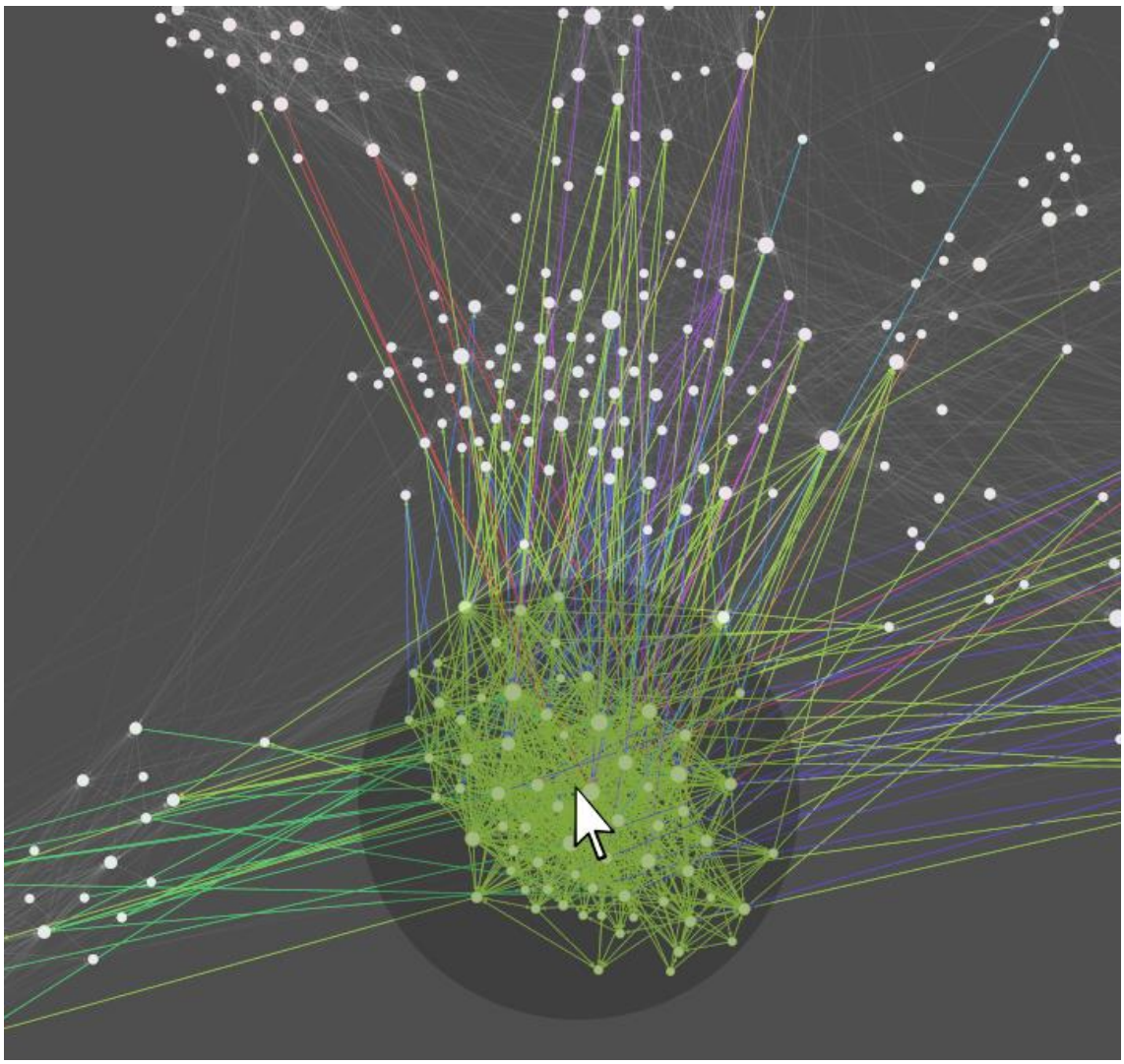

**Figura 12: Visualização em tempo real**

O Gephi apresenta diversas possibilidades de customização do layout do grafo a partir de suas ferramentas. A figura 13 mostra uma comparação de alguns layouts disponíveis dentre os diversos algoritmos criados a partir do arquivo modelo disponível junto com a instalação que descreve a rede de relacionamentos da obra romântica do autor Victor Hugo, Os Miseráveis.

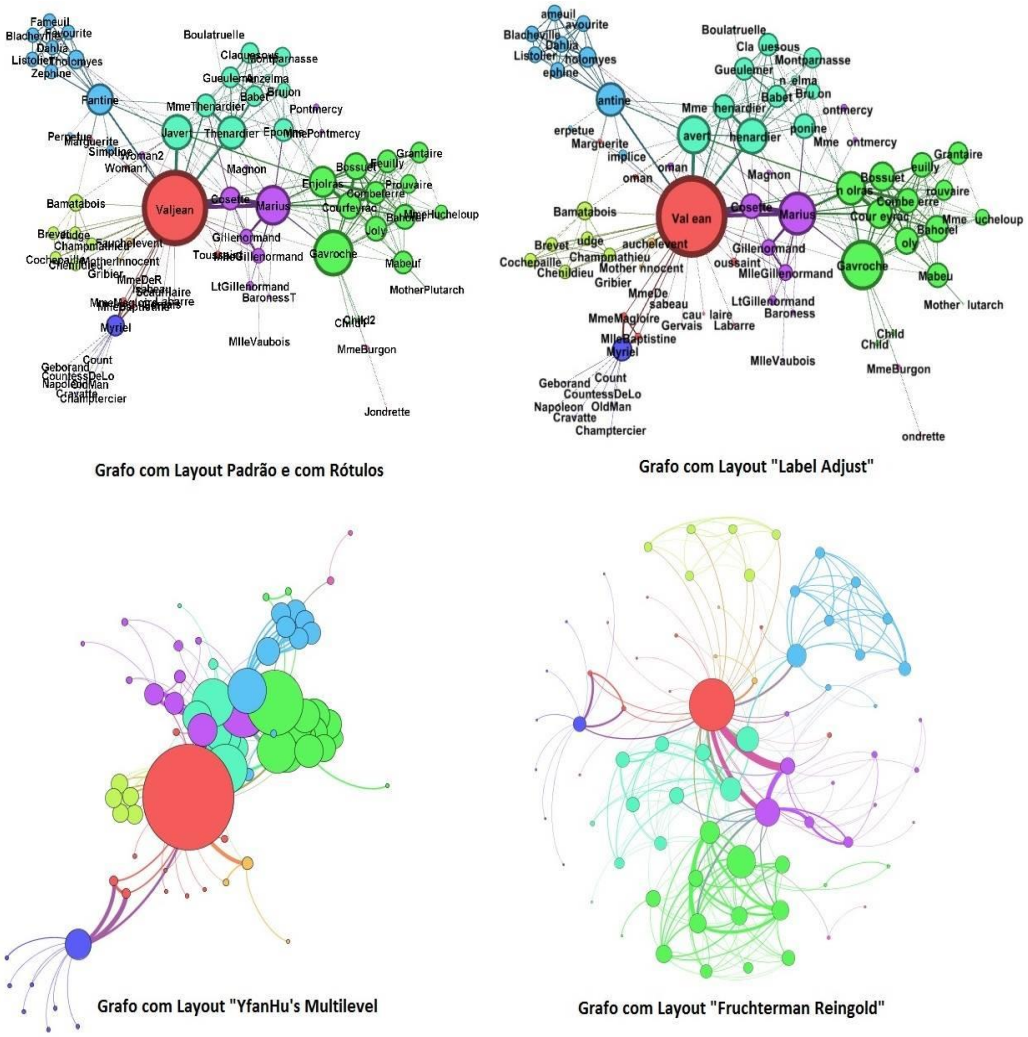


**Figura 13: Rede de relacionamentos de *Les Miserables*, layouts variados.**

Outra importante forma de extrair informações de valor é a por meio de gráficos mais usuais, o Gephi também oferece um bom suporte para a análise estatística e métrica para a análise de redes sociais e redes de escala livre. Ou seja, o software é capaz de produzir gráficos a partir de medidas realizadas sobre os dados, estatísticas ou analíticas. A visualização de gráficos mais comuns ajuda a encontrarmos correlações e proximidade entre redes, o diâmetro da rede e coeficientes de agrupamento [Bastian, 2009].

A figura 14 exemplifica como tais gráficos podem ser apresentados na plataforma Gephi.

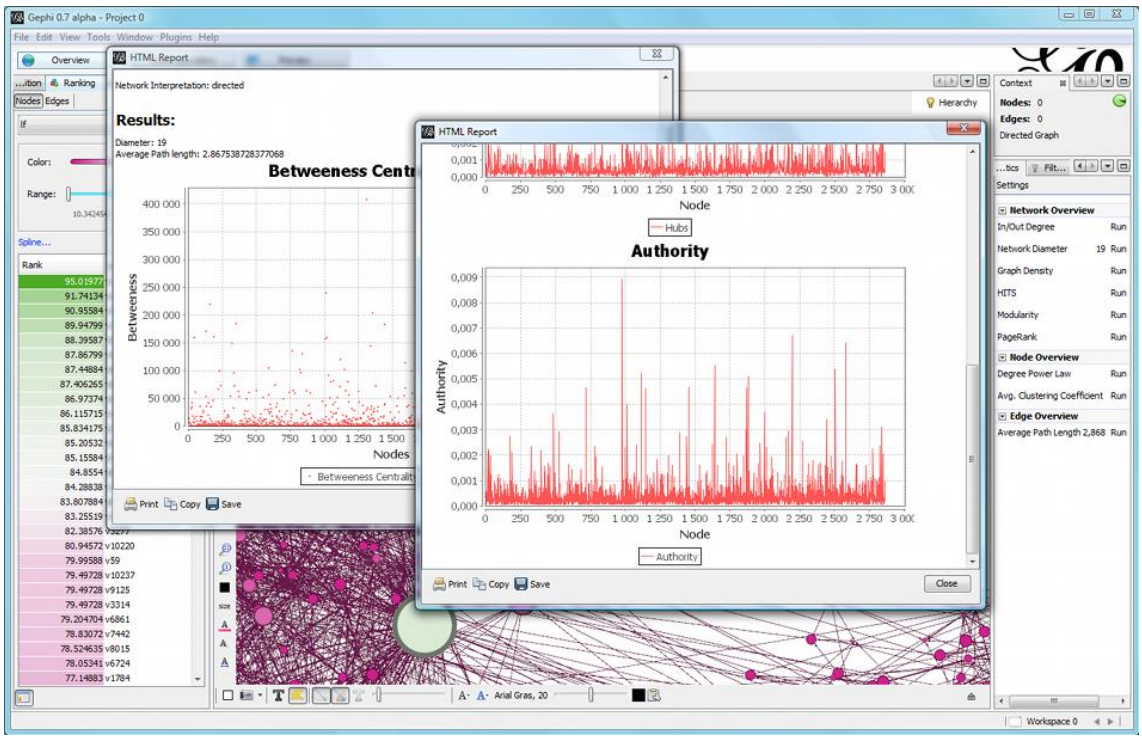


**Figura 14: Métrica na plataforma Gephi.**

Outro diferencial do Gephi é o uso de linhas de tempo como filtro para estruturas dinâmicas como as redes socias, Twitter, Facebook, Linkedin, etc. O software permite que as redes dinâmicas sejam apresentadas em gráficos cuja estrutura ou conteúdo varia ao longo do tempo proporcionando uma recuperação de parte da rede a partir de um intervalo determinado. Essa apresentação pode ser visualizada como uma sequência de um filme, facilitando a compreensão das mudanças e atualizações da rede como um todo. A figura 15 apresenta um streaming de um grafo que apresentou mudanças durante o tempo. no caso da figura, a escala de tempo apresentada é de seis meses.


**Figura 15: Streaming de vídeo do Gephi.**

Reunindo todas as ferramentas mencionadas acima, o Gephi nos permite criar uma cartografia completa da rede estudada, configurar as cores, rótulos, tamanhos e estampas, adicionar sentido a rede configurando a rede espacial mostrada para que a interpretação fique o mais clara possível. Os recursos visuais podem ser explorados completamente de acordo com a vontade do usuário, sendo o processo de criação de informação fácil e dedutível a partir das funções do Gephi. A plataforma é bem simples de ser entendida e a capacidade de criação infindável como foi mostrado brevemente neste capítulo.

Outro benefício do uso do Gephi é a possibilidade de fazer reajustes ainda na fase vetorial, ou seja, antes de renderizar a imagem. Isso possibilita economia na capacidade de processamento e permite a gravação de configurações pré-estabelecidas que podem ser utilizadas em outros projetos. A figura 16 representa uma breve amostra da capacidade de customização do programa Gephi.

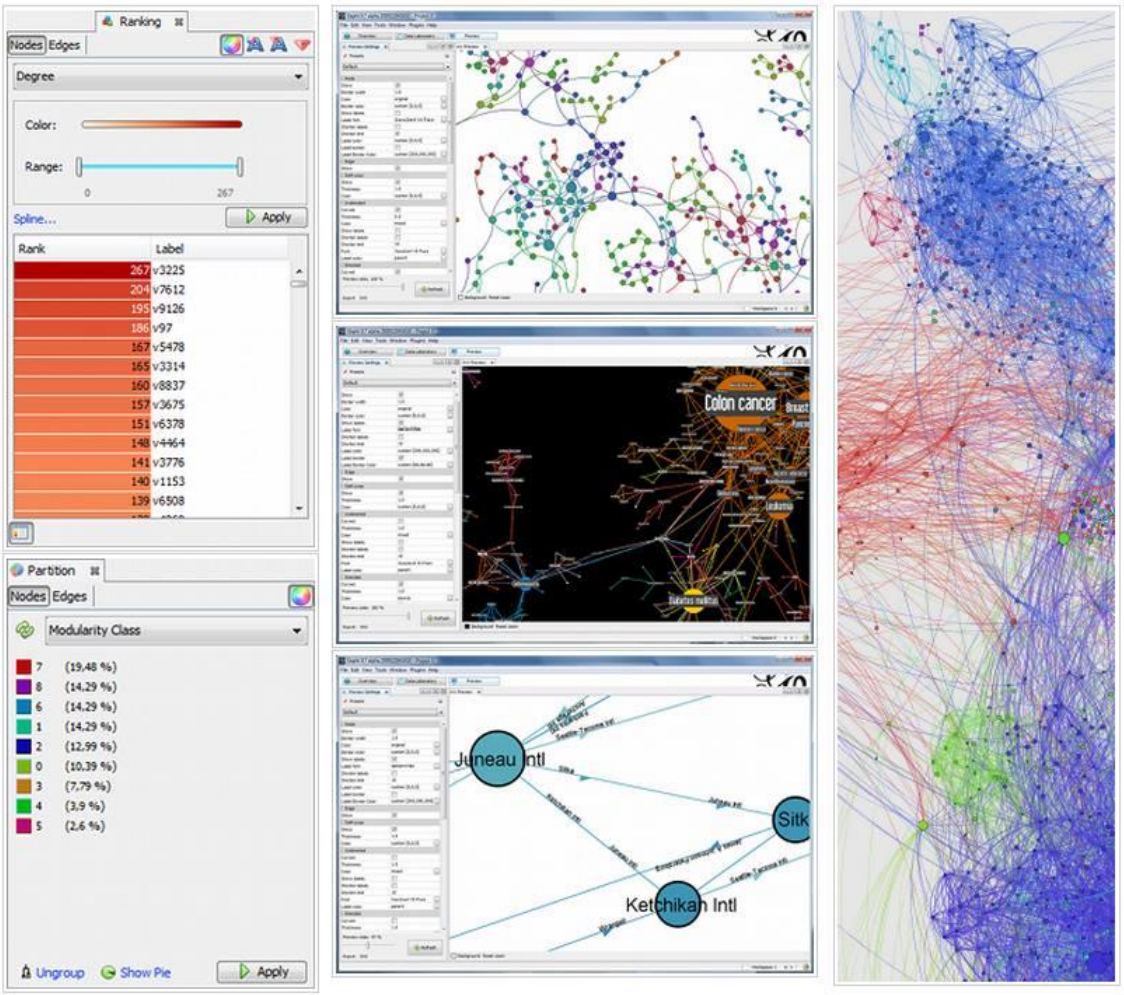


**Figura 16: Criação de Cartografia no Gephi.**

De acordo com o exposto, podemos perceber que o Gephi é uma excelente ferramenta, com muitos recursos gráficos diferentes, de visualização e exploração de redes, sejam elas de comunicação, relacionamento, computadores ou de informação. Os seus recursos permitem a criação de subterfúgios visuais que propiciam entendimento amplo das redes representadas por grafos. A utilização do software não requer conhecimento específico visto que a plataforma é didática e auto explicativa.

Os grafos e gráficos ganham tantos recursos visuais quanto forem necessários para tornar a informação disponível de forma clara e concisa, baseada nos requisitos fornecidos pelo usuário. Seus recursos são modulares, pois podem ser adicionados novos plugins e o programa é feito a partir da plataforma Netbeans que é facilmente

extensível, podendo ser adicionadas novas funcionalidades conforme mude a necessidade.

# 3 AMBIENTE DE VISUALIZAÇÃO

*Nesse capítulo abordaremos as ferramentas utilizadas para a visualização gráfica que interligam os processos de tratamentos dos dados com os sistemas de visualização.*

## 3.1 ESTUDOS DAS PESQUISAS E SEGMENTAÇÕES

O texto dos tweets foi previamente tratado. Todas as palavras encontradas no campo texto foram normalizadas para ASCII minúsculas e sem acentos ou caracteres especiais

Para a segmentação, iremos filtrar os tweets já capturados para que se encaixem em uma classificação segundo algum destino. Caso contrário o tweet é desprezado, porém continuando no sistema de arquivo para futuras análises.

O tweet primeiramente será segmentado por operadora, e depois pelo tipo de serviço.

O estudo de como será feita as segmentações é baseado em amostragens de tweets em diversos momentos e retirado as palavras chaves que mais se relacionam ao determinado filtro.

Os tipos de serviços oferecidos pelas operadoras serão filtrados da seguinte forma, como consta na tabela abaixo.

| **Serviços** | **Palavras Chaves** |
|---|---|
| Telefonia | telefonia, fixo, celular, tel, cel, sinal movel, sinal de celular |
| Internet | internet, inet, conexao, net, 3g, sinal de internet, virtua, 4g |
| Atendimento | atendimento, call center, cancelamento, contato |
| TV | tv, televisao |
| Marketing | loja, oferta |

**Tabela 2: Serviços segmentados com as palavras chaves procuradas**

Neste caso, quanto mais palavras chaves inserirmos nessa tabela, mais tweets iremos obter. Como os tweets em forma original estarão no sistema de arquivos, podemos alterar a tabela, adicionando ou removendo palavras chaves para aumentar a precisão.

Para a descoberta do teor negativo do tweet, analisamos uma grande quantidade de tweets e filtramos a ocorrência das palavras negativas com maior frequência: “ruim”, “cancelar”, “cancelem”, “horrivel”:

| Satisfação | Palavras Chaves |
|---|---|
| Negativa | ruim, cancelar, cancelem, horrivel, manutencao, resolvam, lenta, instavel, lerdo, lerda, nao funciona, injusto, ligacao cai, sem sistema, nao consigo, parou, complicado, tentando, persiste, decepciona, corta, dificil, sinal cai, TV trava, pior, caiu, absurdo, falha, lixo, sem internet, sem telefone, sem net, sem servico, espera, limite, vergonha, problema, anatel, chateacao, desprezo, cortar, indisponivel, nunca atende, descaso, sem atendimento, fora do ar & gírias + palavras de baixo calão. |

**Tabela 3: Palavras Chaves que serão procuradas para segmentar a insatisfação do cliente.**

A partir disso conseguimos extrair de que operadora o tweet está relacionado, de qual serviço daquela operadora e se possui informações negativas. Tweets que não entrarem nessas segmentações não serão processados, porém ficarão no sistema de arquivos para futuras modificações nas palavras chaves, e aprimoramento do algoritmo desenvolvido.

As palavras na Tabela 3 foram escolhidas após observação e análise feita em grupo pelos autores, que contabilizaram as palavras de maior frequência baseados em conjunto de tweets restrito obtido da mesma forma que os utilizados em todo o restante do trabalho.

## 3.2 EXPORTAÇÃO DE DADOS (BANCO DE DADOS)

Os dados foram exportados de duas formas distintas de acordo com o ambiente criado, para os resultados obtidos e apresentados neste trabalho, os dados foram exportados para o programa Gephi, utilizando queryes dentro da própria programação do sistema escolhido, diretamente do banco de dados presente no sistema web. A porta utilizada, para acessar o banco de dados foi a 3306, ou seja, a mesma utilizada para o serviço do Haddop. Neste ponto, o software Gephi foi um grande avanço pois facilitou muito o trabalho de exportação dos dados. O software oferece a possibilidade de se fazer consultas na mesma linguagem em que o banco de dados foi criado, no nosso caso, em MySQL.

Para a parte de visualização do sistema web, foram utilizados templates obtidos através de um plugin gerador de gráficos estáticos chamado HighCharts. Para que esse plugin funcione em tempo real, adaptamos um código PHP para atualizar os dados a cada rotina de 10 em 10 minutos. Esse código PHP, pesquisa nos resultados coletados a partir da função MapReduce as variáveis preestabelecidas e as atualiza na própria estrutura do código onde o HighCharts coleta os dados. Dessa forma foi possível obter atualização em “tempo real” dos parâmetro apresentados nos gráficos.

## 3.3 RESULTADOS ESPERADOS

Capacidade de prover informações reais que possam ser determinantes para a tomada de decisões para as empresas e para os usuários.

Comprovação de uma previsão de cerca de 10% a 15% de tweets com referências negativas das empresas de telecomunicações do Brasil.

# 4 IMPLEMENTAÇÃO

*A decidir ainda o objetivo 3, caso for um contexto, apresentar nesse capitulo todas as informações, e onde entra o profissional de redes.*

## 4.1 HARDWARE UTILIZADO

Um dos objetivos deste projeto é a construção de um ambiente de baixo custo, mas ao mesmo tempo sendo eficiente em suas tarefas.

A escolha de duas máquinas virtuais se dá pelo fato da quantidade de dados que serão coletados e processados. Foi feito um estudo inicial para determinar e prever o volume de dados e o poder de processamento necessário para um modelo ideal do trabalho. Foi concluído que uma máquina virtual é ideal, porém no escopo do projeto é utilizado duas com intuito de demonstrar o uso real do cluster, do paralelismo do processamento e da redundância do sistema de arquivos HDFS.

Novas máquinas podem ser adicionadas ao cluster sem nenhum problema, deixando-o mais robusto e rápido. No projeto atual, ambas as máquinas virtuais são configuradas de forma igual, descrito na tabela 2.

| | |
|---|---|
| **Processador** | 2 Core Virtual |
| **Memória RAM** | 4 GB |
| **Capacidade de armazenamento** | 30GB SSD Disk |
| **Rede** | 1Gbps = 1000 Mbps |

**Tabela 2: Configuração da máquina virtual que é utilizada**

Conforme mencionado, a escolha da configuração se dá pelos mesmos motivos que os da escolha de número de nós no cluster. A partir do estudo da quantidade de informação que é capturada e processada, foi concluído que, para o projeto em ambiente de produção, a configuração ideal pode ser representada pela tabela 3. Com essa configuração, pode-se aumentar a portabilidade, visto que a máquina não necessariamente seja utilizada somente para as atividades do projeto.

| | |
|---|---|
| **Processador** | 1 Core Virtual |
| **Memória RAM** | 1 GB |
| **Capacidade de armazenamento** | 30GB SSD Disk |
| **Rede** | 1Gbps = 1000 Mbps |

**Tabela 3: Configuração ideal da máquina virtual no projeto**

## 4.2 SISTEMA WEB

Nesse tópico falaremos um pouco sobre a implementação do Sistema Web. Para a construção da ferramenta de visualização da análise sentimental dos usuários da rede social Twitter quanto às suas operadoras de telecomunicações fora utilizado algumas ferramentas de suma importância.

O servidor Apache HTTP (Apache HTTP Server) é um servidor de fácil acesso e utilização já que é um software livre. Criado em 1995 por Rob McCool, um ex-funcionário da NCSA, este servidor web será o responsável por processar as solicitações HTTP do nosso sistema. Este servidor foi escolhido pois é o que apresenta uma maior segurança, tem uma excelente performance e uma grande compatibilidade entre várias plataformas e seus recursos, fato muito importante já que na óptica de Big Data, o uso de várias ferramentas se faz necessário para que o melhor resultado seja apresentado. Ele está disponível para os sistemas operacionais Linux, bem como outros baseados em Unix, para o Windows, para o Novell Netware e para o OS/2, que torna hábil que ele rode em computadores obsoletos, desde que atenda aos requisitos mínimos do sistema.

Com sua capacidade de execução de códigos PHP, aproveitamos dessa ferramenta para que o sistema fosse mais facilmente modelado, já que havia um conhecimento prévio nessa linguagem pelos autores deste trabalho. Dessa forma aproveitamos uma das mais usadas combinações existentes entre o Apache e o PHP, que é a com o banco de dados MySQL.

# 5 RESULTADOS OBTIDOS

*Este capítulo apresenta os resultados obtidos graficamente para a visualização dos dados explorados neste trabalho e a interpretação dos autores em relação à obra produzida.*

## 5.1 ANÁLISE DOS DADOS

Este tópica se divide em duas partes abordadas em separado: a primeira revela a análise dos gráficos e dados construídos para o sistema web desenvolvido no trabalho; a segunda parte remete à análise dos gráficos obtidos a partir do software Gephi, que como foi dito anteriormente, é um software específico de customização de gráficos.

### 5.1.1 Sistema Web

O sistema web desenvolvido para este trabalho apresenta seis gráficos obtidos a partir de queries de consulta ao banco de dados obtido após o processamento do MapReduce nos dados coletados. Isto proporcionou obter o comportamento dos gráficos em tempo real, com atualização online no sistema web. Abordaremos agora os gráficos gerados para tal sistema explicando o que foi possível retirar de informação valiosa da pesquisa relacionada a este trabalho.

A figura 17 corresponde ao primeiro gráfico apresentado na interface web revela o total de tweets capturados com filtro sentimental negativo para cada operadora. Como o mesmo gráfico mostra uma curva para cada operadora, é normal que nosso primeiro impulso seja de comparar uma com as outras, no entanto, essa comparação por si só não revela uma boa avaliação dos serviços prestados pelas operadoras e deve ser evitada. A análise a ser realizada é completamente quantitativa e revela tão somente o número de tweets onde foi percebido um sentimento negativo de acordo com o algoritmo de inteligência artificial estabelecido para o sistema.

Por não levar em conta o tamanho das operadoras, comparar uma com a outra baseado somente no número de reclamações apresentado no gráfico nos levaria a cometer um erro, de certa forma, injusto, visto que, uma operadora com um maior número de usuários poderia ter um maior número de reclamações representando, no entanto, uma porcentagem menor de usuários insatisfeitos do que em outras operadoras. Para a comparação entre as operadoras seria necessária uma análise mais completa do que a que é possível a partir dos dados apresentados nos gráficos.

A comparação dos dois gráficos a seguir, os quais foram obtidos em dois meses consecutivos de análise do sistema web revela uma periodicidade em relação a quantidade de tweets coletados e filtrados com sentimento negativo dos usuários. Nos finais de semana é possível perceber que de forma geral a quantidade de reclamações diminui quando comparada aos dias úteis da semana.

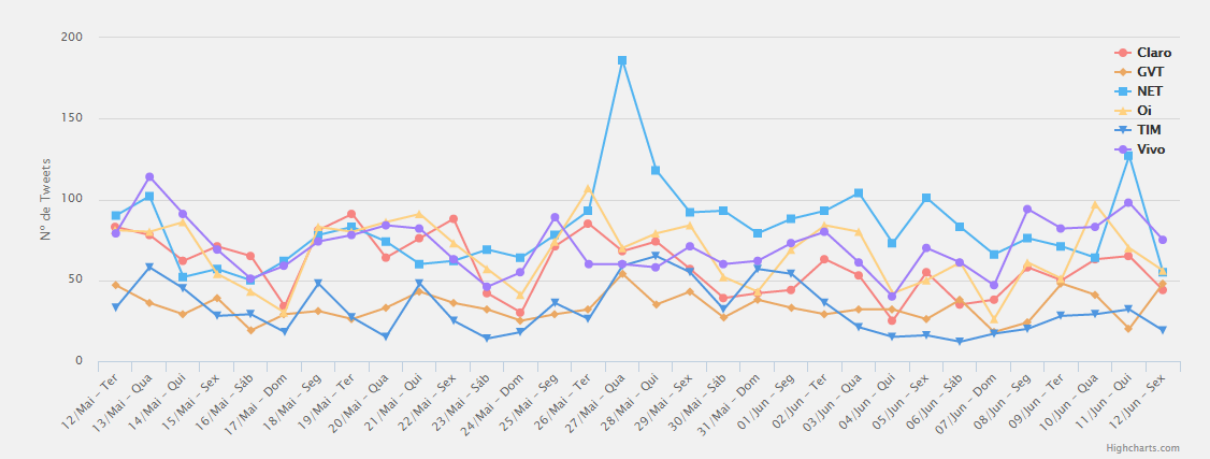


**Figura 17: Simulacao de um mês atrás 12/07/2015 23:05**

A figura a seguir mostra o gráfico obtido a partir do sistema web no dia 13/07/2015, aproximadamente as zero hora e quinze minutos.

.

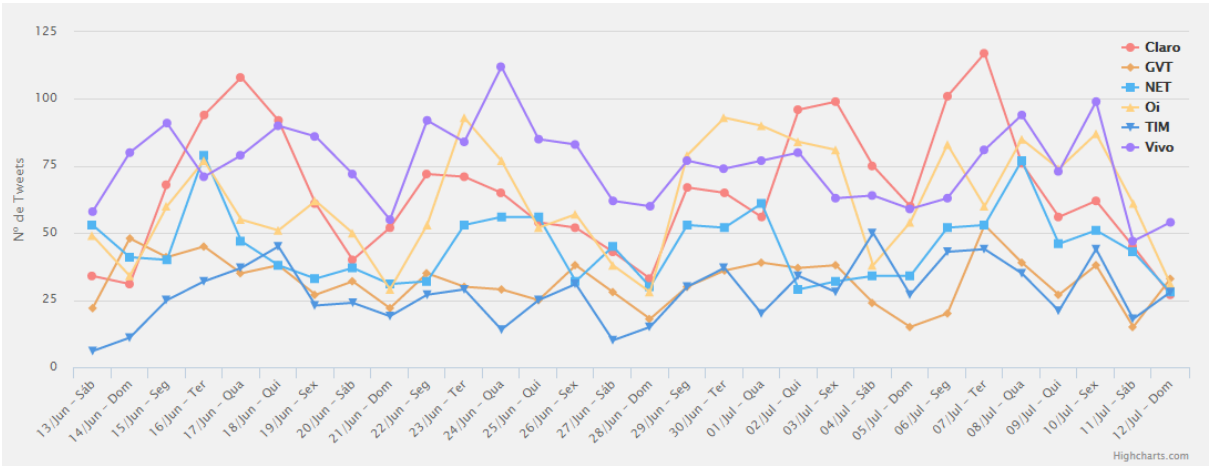


**Figura 18: telecomnasredes.com.br Gráfico 1 13/07/2015 00:15**

O gráfico apresentado revela um padrão que foi percebido em diversas observações feitas em dias e horários diferentes. Primeiramente, é possível perceber uma disparidade entre o número de reclamações obtido durante os dias úteis da semana (de segunda feira a sexta feira) e os finais de semana (sábados e domingos). Em geral, o número de reclamações percebido pelo sistema é ligeiramente maior nos dias úteis em relação aos finais de semana, isso se deve, segundo nossa interpretação, por uma menor demanda de serviços de telecomunicações, visto que os usuários diversificam suas atividades e ficam menos dependentes destes serviços.

Outro motivo é a grande utilização dos serviços de telefonia e internet por parte das empresas e do governo, organizações públicas e privadas, concentradas nos dias da semana. Dessa forma, é possível constatar que grande parte do problema das empresas provedoras de serviços de telecomunicações no Brasil se dá pela ineficiência em administrar os recursos de acordo com o número de usuários em dias cuja demanda é mais acentuada, ou seja, garantir uma escalabilidade no serviço prestado. Em segundo lugar, a observação do gráfico revela alguns dias em que ocorrem picos na curva de alguma das operadoras. Esses picos podem estar relacionados com problemas pontuais que as operadoras estejam enfrentando em determinada ocasião. Por exemplo, uma

operadora que esteja realizando manutenção programada de sua rede pode causar um aumento esporádico no número de reclamações percebido pelo sistema, revelando a insatisfação de alguns usuários em relação ao serviço prestado na ocasião citada.

Na sequência, o sistema web apresenta três gráficos de setores comparando em porcentagem o número de tweets com sentimento negativo em sua composição, segmentados por operadoras em períodos pré-estabelecidos.

O primeiro gráfico da figura 19 apresenta a segmentação realizada em relação à última hora. Como no processo de redução dos dados, estes foram normalizados em relação ao horário tendo sido descartadas as informações de minutos e segundos do *timestamp* o gráfico é construído, na verdade, com os dados obtidos nas duas horas anteriores ao tempo presente, por exemplo, às dezenove horas e vinte e três minutos serão apresentados os dados coletados e filtrados com o sentimento negativo desde às dezoito horas zero minutos até às dezenove horas e vinte minutos, pois o gráfico é atualizado em *Jobs* que ocorrem a cada dez minutos e a janela de tempo do gráfico vai de sessenta a cento e vinte minutos, conforme os minutos avançam dentro da mesma hora.

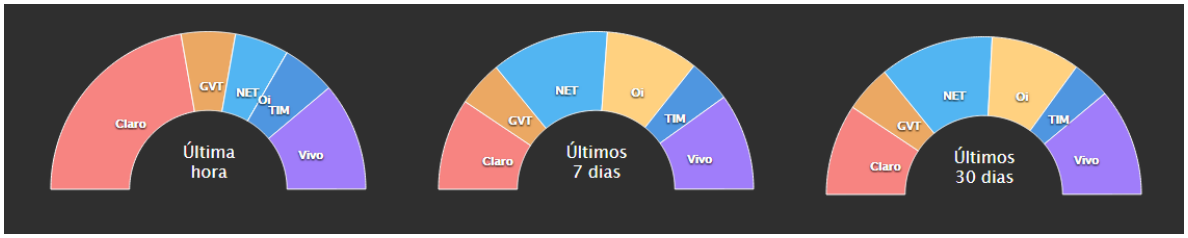


**Figura 19: Operadoras X Tempo, 12/07/2015 19:00**

A figura acima foi tirada do sistema web no dia 12/07/2015, aproximadamente às dezenove horas. Nela podemos observar que os gráficos dos últimos sete dias e o dos últimos trinta dias tendem a ser muito próximos, pois revelam uma tendência temporal de normalização no número de reclamações de cada operadora.

Já o gráfico da última hora tem uma variação bastante expressiva, pois revela um período de análise curto em relação à normalização dos dados que o viés temporal causa. As figuras a seguir mostram algumas variações deste gráfico obtidas em dias e horários variados.

Mais uma vez vale ressaltar que as informações obtidas pelo algoritmo de inteligência artificial não respaldam a comparação entre as operadoras e suas respectivas qualidades nos serviços prestados, pois não apresentam nenhuma relação com o tamanho das redes de usuários de cada operadora. Dessa forma, o objetivo dos gráficos mostrados é somente apresentar de forma quantitativa o número de tweets coletados com sentimento negativo para cada operadora.

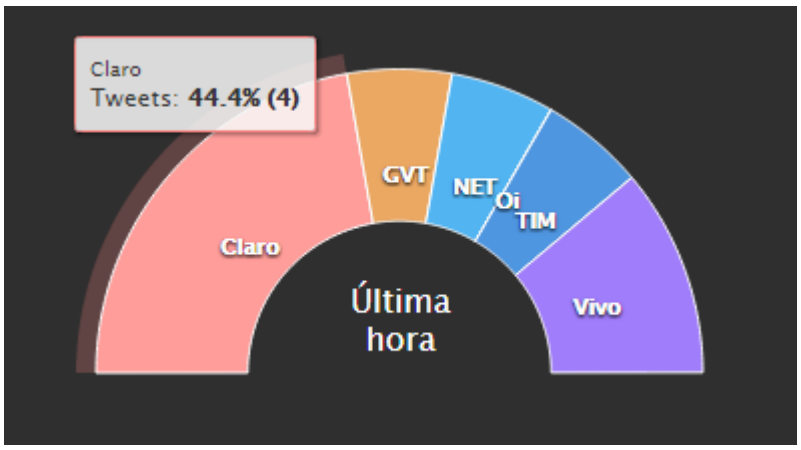


**Figura 20: Operadoras X Tempo, 12/07/2015 19:42**


**Figura 21: Operadoras X Tempo, 11/07/2015 08:05**

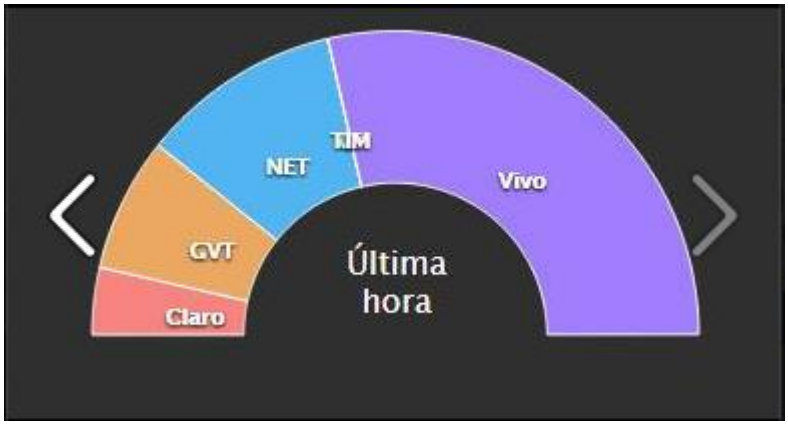


**Figura 22: Operadoras X Tempo, 12/072015 16:20**

A figura 20 poderia levar à uma percepção errônea de que uma das operadoras, a Claro, estaria prestando um serviço pior do que as outras, caso não fosse apresentada a informação de que somente quatro pessoas twittaram na última hora em desfavor da Claro. Dessa forma, em alguns momentos, a observação do gráfico pode ser perigosa, principalmente em horários de menor quantidade de manifestações, pois a porcentagem em relação ao total de tweets apresentados no gráfico pode parecer elevada, representando, no entanto, poucos tweets. Isso pode também ser observado na figura 21, pois no horário relacionado, havia somente dois tweets coletados com sentimento negativo, representando assim, uma porcentagem alta para cada operadora, 50%, representando somente um tweet por operadora.

Comparando as figuras 20, 21 e 22 é possível perceber que o gráfico da última hora apresenta uma variabilidade muito grande conforme exposto neste trabalho.

Outra informação a ser retirada do sistema é a soma de reclamações por operadora segmentada em relação aos serviços prestados por cada operadora. Essa análise é de crucial importância, pois o acúmulo de reclamações pode indicar uma insatisfação por parte dos clientes das operadoras e porque revela em quais serviços cada empresa tem um maior número de reclamações, indicando suas fragilidades.

Foi observado que o serviço de internet é disparado o campeão de reclamações por parte dos usuários seguido do serviço de telefonia fixa e móvel, por impossibilidade do sistema em separar esses serviços, pois depende de especificações que seriam provenientes dos usuários e não são fornecidas.

Dentro deste contexto, foram criados dois gráficos distintos, pois caso deixássemos todos os serviços representados em uma mesma escala, ficaria mais complicado de se comparar os serviços segmentados. O primeiro gráfico possui uma escala maior que a do segundo para facilitar a dimensão do número de reclamações percebidas pelo sistema de inteligência artificial.

As figuras 23 e 24 mostram a segmentação por serviços e por operadoras separadas em dois gráficos distintos que ajudam a compreendermos os serviços de telecomunicações no Brasil durante o período apurado.

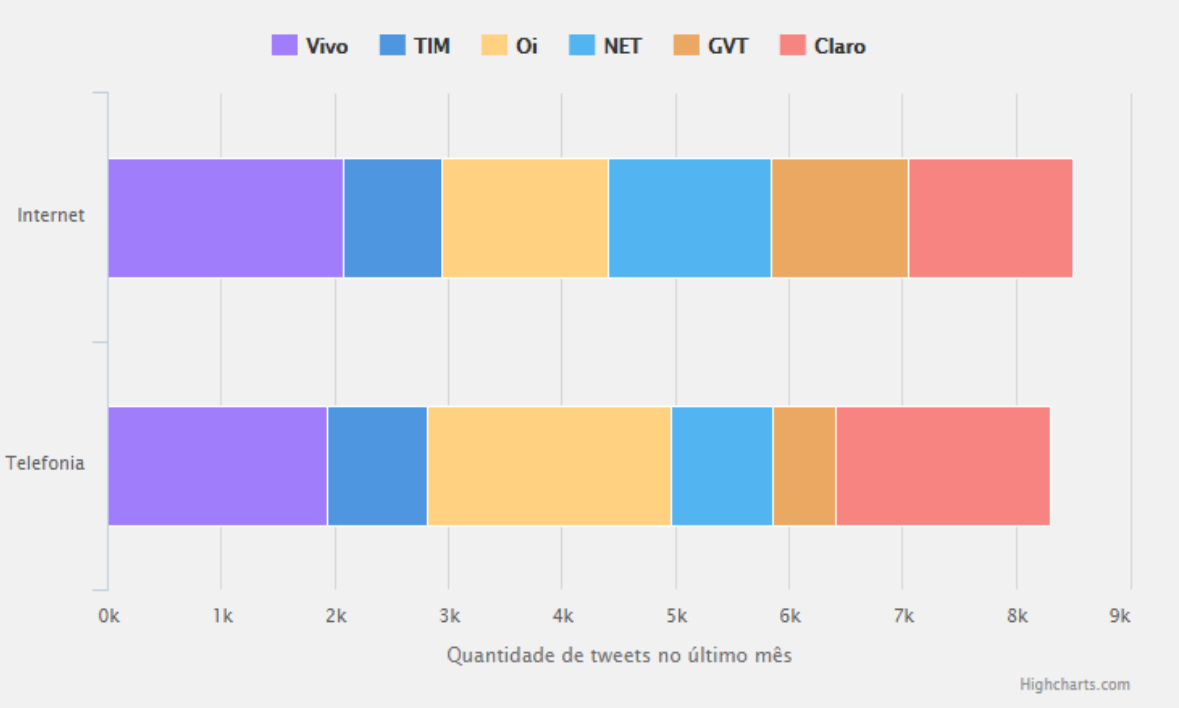


**Figura 23: Serviço X Operadora, 12/07/2015 22:20**

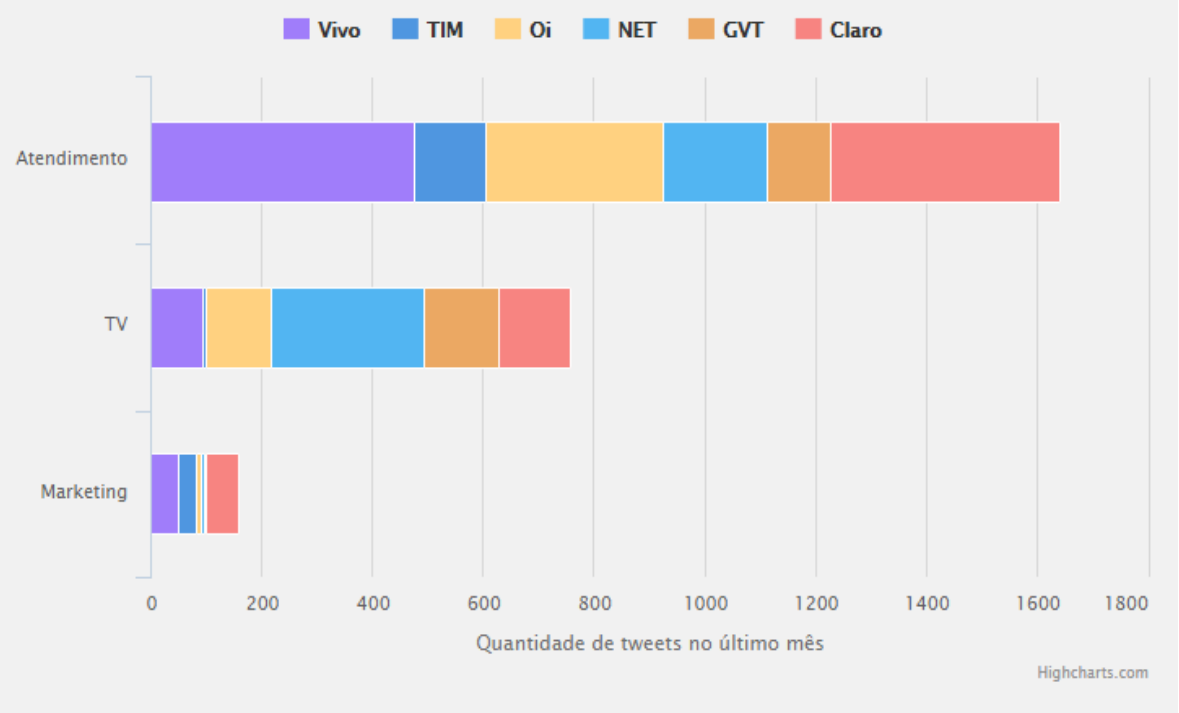


**Figura 24: Serviço X Operadora, 12/07/2015 22:20**

Ao se observar os gráficos pode-se facilmente perceber que os serviços de internet e telefonia são os mais frágeis em relação à percepção do usuário final, enquanto o serviço de televisão por assinatura é o mais estável dos serviços do setor de telecomunicações. Isso se deve à estrutura da transmissão da televisão ser mais bem fomentada do que a estrutura dedicada aos demais serviços. Além disso, o serviço de televisão por assinatura não sofre com o aumento esporádico de usuários, pois não importa para a qualidade do serviço o número de usuários que estão assistindo televisão no momento, e sim o número de assinantes no total.

Outra análise importante a ser observada é traçar uma linha temporal em relação à quantidade de reclamações postadas pelos usuários, gerando expectativas e previsões acerca do comportamento esperado dos usuários ao longo do dia ou da semana para identificar os horários ou dias em que a utilização de determinado serviço se torna mais caótica. Para observar este fenômeno temporal, foram criados dois gráficos baseados no total de tweets coletados com percepção de sentimento negativo pelo sistema em uma análise temporal que pode ser apresentada nas figuras a seguir em relação às horas do dia ou aos dias da semana.

A figura 26 revela o padrão de comportamento do usuário durante o dia. É possível notar que no horário comercial, ou seja, entre oito e dezoito horas, o número de reclamações aumenta, pois a utilização dos serviços críticos, como telefonia e internet, também aumenta. A partir das dezoito horas, percebe-se que o número de reclamações relacionadas a telefonia começa a decair, pois encerra-se o expediente comercial e os usuários, geralmente, deslocam-se para suas residências ou outros lugares que não demandam tanto telefone quanto em seus trabalhos. Já quanto ao serviço de internet, no mesmo horário, começa a subir o número de reclamações, pois o uso da internet não está relacionado somente com o ambiente laboral, mas principalmente com o ambiente domiciliar, visto que as pessoas utilizam muito a internet em seus momentos de descanso e lazer.

O crescente uso de dispositivos móveis se relaciona com este fato, pois ocasiona o aumento da demanda pelo serviço de internet nas horas em que o usuário se encontra em casa.

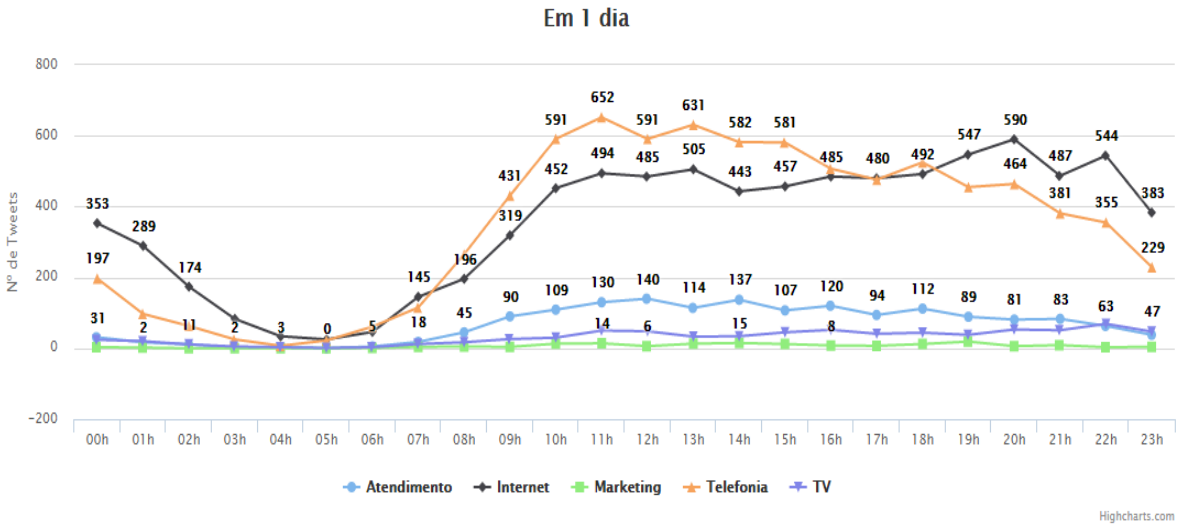


**Figura 25: Serviço X Tempo, 13/07/2015 00:40**

Finalmente, analisaremos a linha temporal semanal. Esta análise, bem como esperado, revela que o uso das tecnologias de telecomunicações tem alavancado o número de reclamações durante os dias úteis da semana, principalmente em relação aos serviços de telefonia, internet e atendimento. É possível perceber ainda, que o número de reclamações de telefonia é descendente nos finais de semana, enquanto o de internet também apresenta uma redução, porém menos acentuada que a observada na telefonia.

Dessa forma, podemos novamente relacionar o uso do telefone à atividade laboral e o uso da internet não somente na atividade laboral, mas também nas atividades de descanso e lazer das pessoas. A figura 19, a seguir, ilustra a linha temporal supracitada.

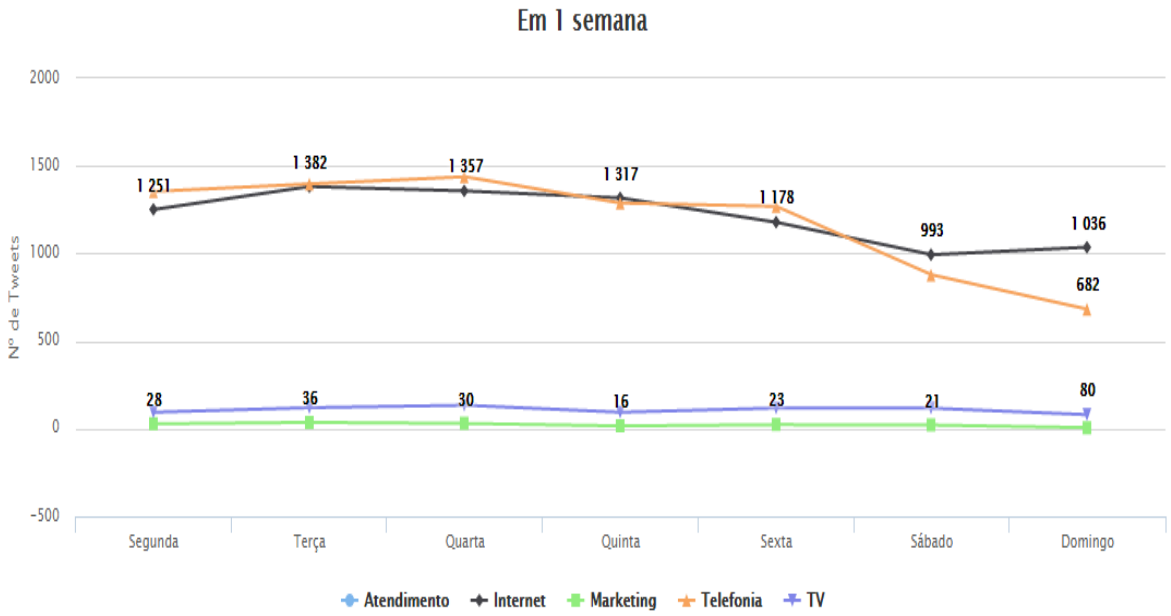


**Figura 26: Serviço X Tempo, 13/07/2015 01:41**

### 5.1.2 Gephi

Diante do desafio de apresentar o grande volume de dados obtidos e analisados uma ferramenta se destacou pela facilidade de utilização e pelo espetacular resultado visual alcançado na confecção de gráficos, o software Gephi.

O uso do Gephi propiciou uma melhora na qualidade visual dos gráficos formulados facilitando a compreensão das relações entre os agentes encontrados na análise dos dados. A aplicação dessa ferramenta nos permitiu apresentar grafos mais dinâmicos, onde a informação pode estar presente nos nós e também nas arestas. Cada aresta pode representar uma série de valores, como sua largura, seu comprimento, sua cor e sua direção. Os nós, da mesma forma, podem receber valores para cada um de seus atributos, como tamanho, largura da borda, cor, rótulo, que pode ou não ser mostrado, tamanho do rótulo, cor do rótulo e formato do nó.

A seguir, algumas figuras ilustrarão parte do potencial artístico do software Gephi. Lembramos que o tempo de exploração dos recursos do Gephi foi muito curto, dessa forma, mesmo apresentando um resultado além do esperado, ficamos ainda, muito aquém das possibilidades infinitas que o software proporciona.

A figura a seguir relaciona as operadoras selecionadas para a nossa prospecção de dados com os serviços escolhidos para a nossa análise. Nessa figura, o tamanho dos nós foi escolhido manualmente somente com o intuito de destacar os nós que representam as operadoras daqueles que representam os serviços. As cores dos nós e arestas remetem à identidade visual das operadoras.

A espessura das arestas foi obtida a partir da contagem de tweets que relacionam cada operadora negativamente com determinado serviço, dessa forma, quanto maior a espessura da aresta, maior o número de reclamações encontrado pelo sistema até o dia doze de julho de dois mil e quinze. O arranjo espacial pôde ser alterado de forma a organizar melhor as informações obtidas.

Vale ressaltar novamente, que análise obtida foi puramente quantitativa, não representando de forma alguma, a qualidade dos serviços de cada operadora. Cabe ao consumidor, interpretar as informações apresentadas de acordo com suas necessidades, eximindo os autores dessa obra de qualquer responsabilidade sobre sua tomada de decisão.

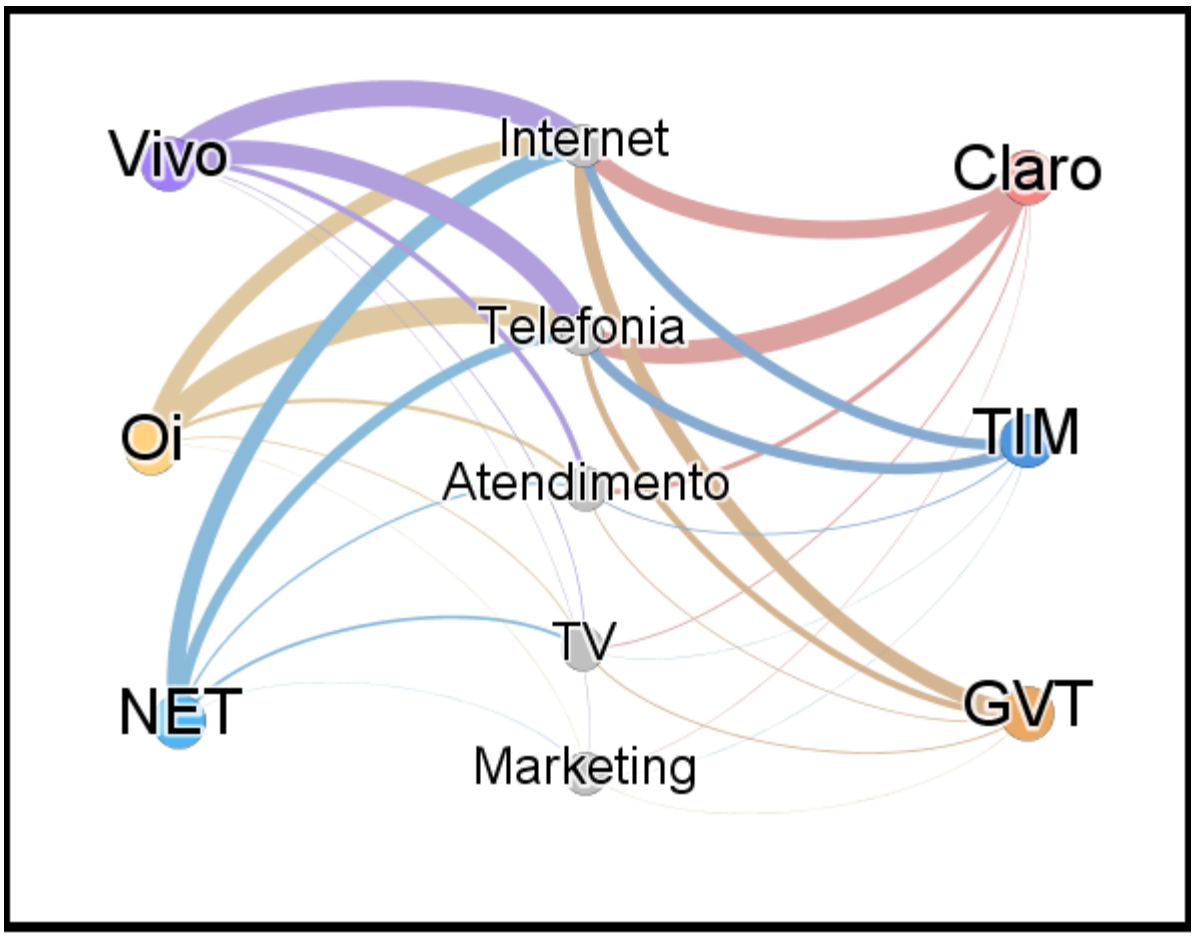


**Figura 27: Operadoras X Serviços, Gephi**

No âmbito de entender melhor este trabalho, nos predispusemos ainda a criar um grafo que expressasse a dimensão da rede de usuários de cada operadora. Para melhor se relacionar com o tema do trabalho, escolhemos apresentar somente a rede de usuários que tiveram tweets coletados com percepção de sentimento negativo, ou seja, extraídos de tabelas do nosso banco de dados do sistema web depois da segmentação que caracteriza a aparição de reclamações nos tweets. Dessa forma, criamos um gráfico onde cada aresta liga um usuário a uma operadora, não impedindo que o mesmo usuário esteja relacionado com mais de uma operadora, mas nunca pela mesma aresta.

Foi possível, portanto, visualizar a quantidade de usuários com percepção negativa de cada operadora e como o software Gephi nos permitiu agrupar estes nós em pontos mais próximos a cada operadora, foi possível adquirir uma visão ampla do que está sendo exposto no trabalho através do sistema web.

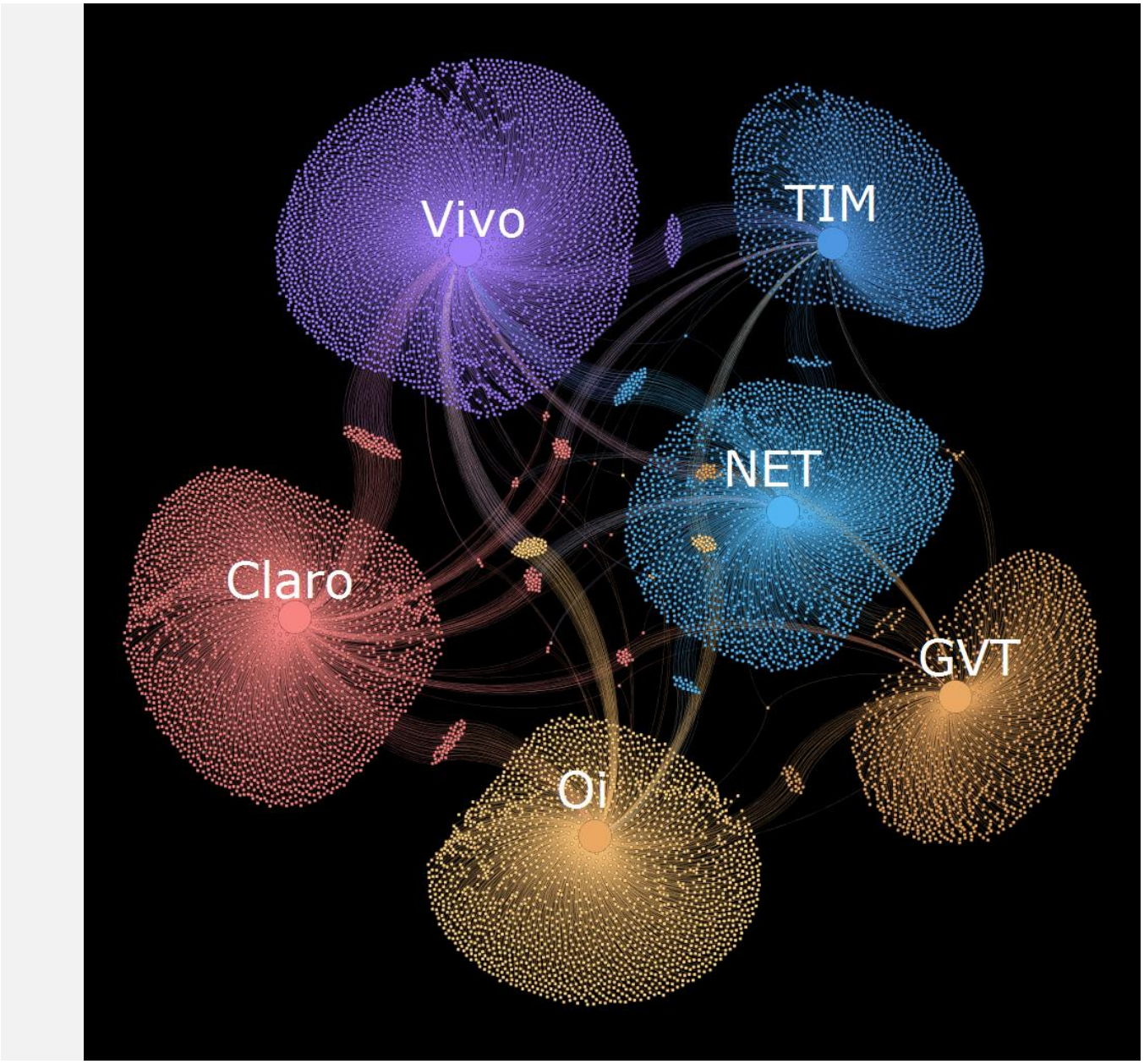


**Figura 28: Operadoras X Usuários, Gephi**

Ao observar a figura 28, podemos perceber que alguns nós parecem um pouco perdidos, agrupados em porções menores localizadas entre as operadoras e suas respectivas redes. Estes pontos representam usuários que por alguma razão, indiferente à motivação deste trabalho, postaram comentários que foram enquadrados pelo sistema como reclamações relacionadas a mais de uma operadora.

A figura 29 ilustra um desses casos somente com o intuito de exemplificar a ocorrência destes grupos de nós. A identidade do autor foi preservada a fim de evitar qualquer exposição desnecessária.

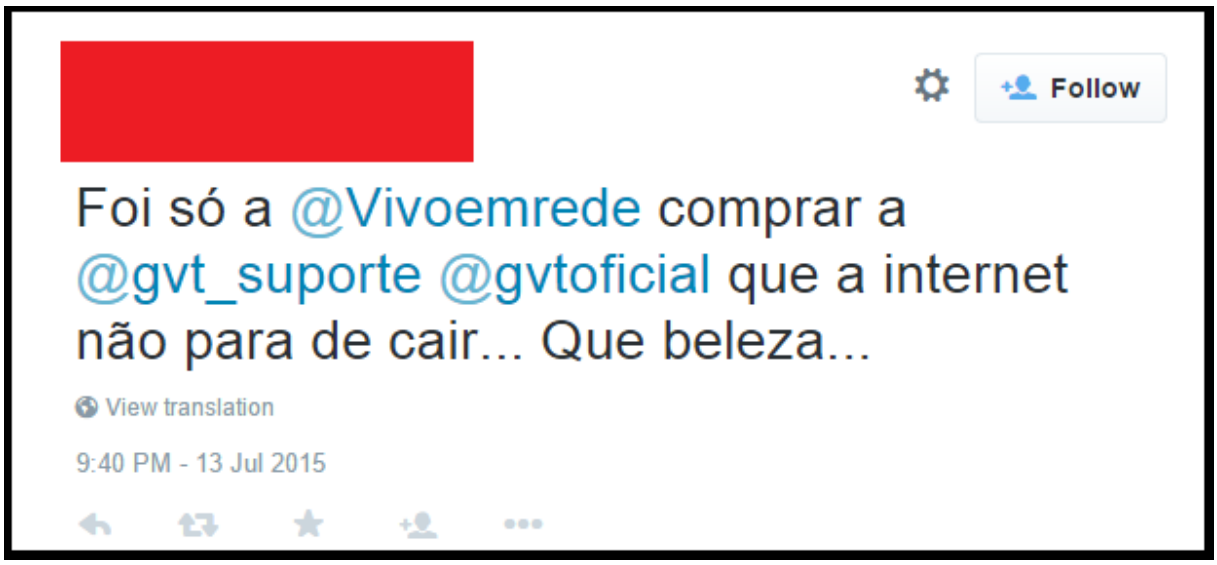


**Figura 29: Tweet exemplificando reclamação associada a mais de uma operadora.**

# 6 CONCLUSÃO

Nesse trabalho, foram apresentadas ferramentas utilizadas na captura, no processamento, na segmentação, no armazenamento e na visualização de um imenso volume de dados semiestruturados, possibilitando a obtenção de informações de valor inestimável para organizações públicas e ou privadas. A partir do sistema desenvolvido, é possível auditar um grande número de reclamações dos usuários não somente em relação às operadoras de telecomunicações, mas de qualquer sociedade empresarial que esteja disposta a reconhecer as redes sociais como fonte de informação relevante para o desenvolvimento de suas atividades.

A melhoria na qualidade dos serviços prestados pelas organizações pode ser continuada se for levada em consideração a reclamação dos usuários, seja ela em meio formal ou informal, advinda de fontes confiáveis ou de estudos estatísticos, como os que podem ser desenvolvidos a partir deste trabalho.

Apesar de não ter havido tempo suficiente para que fossem coletadas informações bastantes para se entender os padrões de consumo através das reclamações dos usuários, bem como para desenvolver uma ferramenta de filtragem mais eficaz, os dados obtidos no período disponível, nos permitiram apoiar as teorias sobre o perfil dos usuários das redes de comunicações apoiadas pelo frequente uso de dispositivos moveis com acesso à internet.

Esse trabalho foi incrementado com um estudo prático sobre o software Gephi, possibilitando a análise gráfica tridimensional dos dados coletados no ambiente de Big Data. Pudemos concluir que o uso dessa poderosa ferramenta, aliado a um bom banco de dados escalável é de suma utilidade para a obtenção de resultados no tratamento de Big Data.

A importância de um recurso gráfico mais bem elaborado na apresentação de resultados, visto que, a percepção humana pode ser distorcida quando se trata de um volume muito grande de dados, caso esses sejam analisados sob uma perspectiva comum. Portanto, um gráfico que apresente uma maior qualidade visual pode alterar a percepção humana tornando mais aceitável ou menos desagradável a informação trazida pelos gráficos.

Fica como legado desse trabalho, o ambiente web que se mostrou uma valiosíssima ferramenta para a captura e tratamento dos dados sociais pesquisados.

Como o sistema foi construído visando a escalabilidade do número de tweets recebidos, ele está pronto para receber dados em nível global e pode ser adaptado para outros segmentos empresariais caso haja interesse.

# REFERÊNCIAS BIBLIOGRÁFICAS


1) Hurwitz, J., Nugent, A., Halper, F. and Kaufman, M. 2013. Big data for dummies. Hoboken, N.J.: Wiley.
2) Mayer-Schönberger, V. and Cukier, K. 2013. Big data. Boston: Houghton Mifflin Harcourt
3) Ohlhorst, F. 2013. Big data analytics. Hoboken, N.J.: Wiley.
4) Franks, B. 2012. Taming the big data tidal wave. Hoboken, New Jersey: John Wiley & Sons, Inc.
5) Berman, J. J. 2013. Principles of big data. Amsterdam: Elsevier, Morgan Kaufmann.
6) Sathi, A. 2012. Big data analytics. Boise: MC Press.
7) Big Data Now. 2012. [e-book] Sebastopol, CA: O'Reilly Media, Inc.
8) O'NEIL, P.; QUASS, D. Improved Query Performance with Variant Indexes. In Proc. Of the ACM SIGMOD Conference, Tuscon, Arizona, May, 1997. (O'NEIL; QUASS, 1997)
9) HAN, J; KAMBER, M. Data Mining, Southeast Asia Edition: Concepts and Techniques, 2a edição. Morgan Kaufmann, São Francisco, 2006. (HAN; KAMBER, 2006)
10) WHITE, T. Hadoop: The Definitive Guide, Third Edition. [S.l.]: O'Reilly Media, Inc., 2012.
11) INTEL IT CENTER. Planing Guide Getting Started With Big Data, http://www.intel.com/content/www/us/en/big-data/getting-started-with-big-data-planning-guide.html, 2014 (INTEL 2014)
12) White, T. (2012). Hadoop: The Definitive Guide. O'Reilly Media, Inc., 3th edition
13) Ranger, C., Raghuraman, R., Penmetsa, A., Bradski, G., and Kozyrakis, C. (2007). Evaluating mapreduce for multi-core and multiprocessor systems. In High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on, pages 13 –24.
14) Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003). The google file system. SIGOPS Oper. Syst. Rev., 37(5):29–43.
15) Dean, J. and Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. Commun. ACM, 51(1):107–113.
16) Matthew A. Russell, Mining the Social Web, Second Edition, October 2013, Second Edition
17) RABUSKE, Márcia. Introdução à teoria dos grafos. Florianópolis: UFSC, 1992.
18) Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
19) Helbing, D. and Balietti, S. 2011. Big data, privacy, and trusted web: What needs to be done.

20) Jerome, J. Buying and Sellling Privacy: Big Data's Different Burdens and Benefits.

21) Hilbert, M. 2013. Big Data for Development: From Information-to Knowledge Societies. University of Southern California-Annenberg School for Communication.

22) Tene, O. and Polonetsky, J. 2013. Big Data for All: Privacy and User Control in the Age of Analytics. HeinOnline.

23) Lee, Y., Kang, W. and Son, H. 2010. An internet traffic analysis method with mapreduce.

24) Einav, L. and Levin, J. D. 2013. The data revolution and economic analysis.

25) Jeon, Y. 2012. Impact of Big Data: Networking Considerations and Case Study. International Journal of Computer Science & Network Security, 12 (12).