DESENVOLVIMENTO DE UMA FERRAMENTA DE COLETA E ARMAZENAMENTO DE DADOS DISPONÍVEIS EM REDES SOCIAIS

Bruno Edgar Führ, Evandro Franzen[1]

Resumo: O crescimento da internet na última década fez dela a maior fonte de dados de acesso público em todo mundo. Esses dados formam o que se chama de *Big Data*, termo que se refere ao enorme volume de dados que existe hoje nas mais diversas fontes. A partir desses dados, organizações têm a oportunidade de analisar o comportamento de seus clientes, o que o público pensa de seus produtos ou serviços, além de outras possibilidades. Porém, para um sistema conseguir realizar alguma análise, primeiro é necessário obter e armazenar um grande volume de dados. Considerando essa necessidade, este trabalho apresenta uma arquitetura de uma ferramenta para coleta e aplicação de algoritmos para análises de dados oriundos da *web*. Por meio dela é possível definir uma estrutura de armazenamento e configurar coletores de dados de diversas fontes, como páginas da internet e redes sociais, por diferentes métodos de busca, como palavras-chave e endereços. Além da arquitetura e do funcionamento, são descritos resultados de testes realizados em outubro e novembro de 2014.

Palavras-chave: *Big Data*. Coleta de dados. Redes sociais.

1 INTRODUÇÃO

O rápido crescimento da internet na última década fez dela a maior fonte de dados de acesso público em todo o mundo (LIU, 2011). Nos dias de hoje, há imensa quantidade de dados disponíveis nos mais diversos locais. Sistemas de informação, páginas de notícias, *blogs* e redes sociais são alguns exemplos de fontes de dados que podem ser utilizadas para extrair informações relevantes.

O volume com que essas informações são criadas e atualizadas diariamente excedem, no entanto, a capacidade de armazenamento e processamento das tecnologias de informação tradicionais. Como exemplo, citam-se as redes sociais *Twitter* e *Facebook*, que geram, por dia, aproximadamente 17 terabytes (TB) de dados (EATON et al., 2012).

A esse grande volume de dados, que não pode ser processado e analisado utilizando-se processos e ferramentas tradicionais, atribui-se o termo *Big Data* (EATON et al., 2012). O conceito refere-se ao enorme crescimento no volume, variedade e velocidade com que os dados estão sendo produzidos e ao conjunto de aplicações que geram, armazenam, processam e analisam esses dados.

Segundo Hurwitz (2013), as pessoas estão se tornando emissoras de informações. Ao navegar pela rede, deixam rastros digitais com inúmeras informações de valor, como por exemplo: como tomam decisões de compra, como influenciam seus amigos, que assuntos procuram nos *sites* de busca, por que amam ou odeiam as marcas; e estão transformando a internet em uma imensa plataforma de pesquisa, que pode ser utilizada para reduzir custos e até como principal fonte de coleta de opiniões.

1 Mestrado em Computação pela Universidade Federal do Rio Grande do Sul, Brasil(2002). Professor assistente do Centro Universitário UNIVATES, Brasil.

Big Data é o conjunto de dados estruturados e não estruturados que excedem a capacidade dos sistemas de banco de dados tradicionais (DUMBILL, 2012). Esses dados podem ser provenientes das mais variadas fontes, como dados de transações financeiras até mensagens em redes sociais.

Um "Sistema de *Big Data*" é um conjunto de recursos computacionais baseados em programação, algoritmos, ferramentas de busca e banco de dados que permitem agregar todos os dados estruturados e não estruturados. Quando unidas, essas fontes de dados geram um enorme volume, do qual uma aplicação pode extrair conhecimentos, buscar padrões e fazer análise que não seriam possíveis sem essa agregação. Não é uma tecnologia única, mas uma combinação de velhas e novas técnicas que ajudam as organizações a obterem conhecimentos práticos. Esses sistemas devem ter a capacidade de gerenciar um vasto volume de dados discrepantes, para permitir análises e decisões em tempo real.

A grande novidade dessas soluções é lidar não apenas com dados estruturados de bancos de dados relacionais, mas também com os chamados dados não estruturados, que até então só podiam ser compreendidos por pessoas. São mensagens em redes sociais, comentários em *blogs* e comportamento de clientes que dependem de contexto para terem sentido.

Um uso comum para sistemas de *Big Data* é extrair de dados desestruturados alguma informação que faça sentido para uso das pessoas ou para servir de entrada para outra aplicação. Como exemplo citar-se o caso de uma empresa que tira fotos de satélite e vende a seus clientes informações em tempo real sobre a disponibilidade de vagas de estacionamento livres em uma cidade, em determinada hora.

Lidar com essa imensa quantidade de dados, não estruturados, traz, porém, novos desafios, principalmente na área de banco de dados. Nesse contexto de dezenas, ou centenas de terabytes de dados não estruturados, os bancos de dados tradicionais não são os mais adequados, principalmente por possuírem uma estrutura rígida, não permitindo armazenar modelos de dados flexíveis (VIEIRA et al., 2012).

Uma das alternativas usadas para armazenar dados não estruturados e para recuperá-los de forma eficiente são os bancos de dados denominados NoSQL, que é uma abreviação de *Not Only* SQL (não apenas SQL), utilizados principalmente quando um Sistema de Gerência de Banco de Dados relacional não apresenta a performance adequada. O propósito das soluções NoSQL não é substituir o modelo relacional, e sim ser utilizado em casos em que se necessite de maior flexibilidade da estruturação da base de dados (CHANG et al., 2010).

Os sistemas de bancos de dados NoSQL apresentam algumas características que os diferenciam dos sistemas baseados no modelo relacional, tais como:

- escalabilidade: característica de um sistema que pode continuar a servir um grande número de requisições com pouca perda de performance;

- disponibilidade: característica de um sistema resistente a falhas de qualquer natureza (*hardware, software*, energia). Tem como objetivo manter os serviços disponíveis o máximo de tempo possível.

Além das características citadas, bancos de dados NoSQL podem ser classificados de acordo com o modelo de armazenamento de dados, que pode ser adequado para diferentes aplicações. Segundo Chodorow (2013), esses sistemas atualmente podem ser agrupados em quatro modelos principais: chave-valor, orientado a colunas, orientado a documentos, orientado a grafos.

Uma das primeiras implementações de um sistema não-relacional foi o *BigTable* da *Google*, lançado em 2004. Outras implementações surgiram, como o *Dynamo*, utilizado pela *Amazon* e o Cassandra, desenvolvido pelo *Facebook* em 2008. Além desses, diversos outros projetos NoSQL de

código livre estão disponíveis, como, por exemplo, o *Apache CouchDB*, o *MongoDB, HyperTable* e *Hbase*.

Além de requerer uma estrutura de armazenamento mais flexível e robusta, um volume de dados também necessita de um ambiente em que possa ser aplicados algoritmos de processamento e de análise, para que dele se extraiam informações úteis e de valor.

Em vista desses novos desafios, este trabalho apresenta o desenvolvimento de uma ferramenta de apoio à criação, obtenção e armazenamento de um grande volume de dados, de modo que possa servir como base para o futuro desenvolvimento de aplicações que utilizem técnicas de mineração de dados.

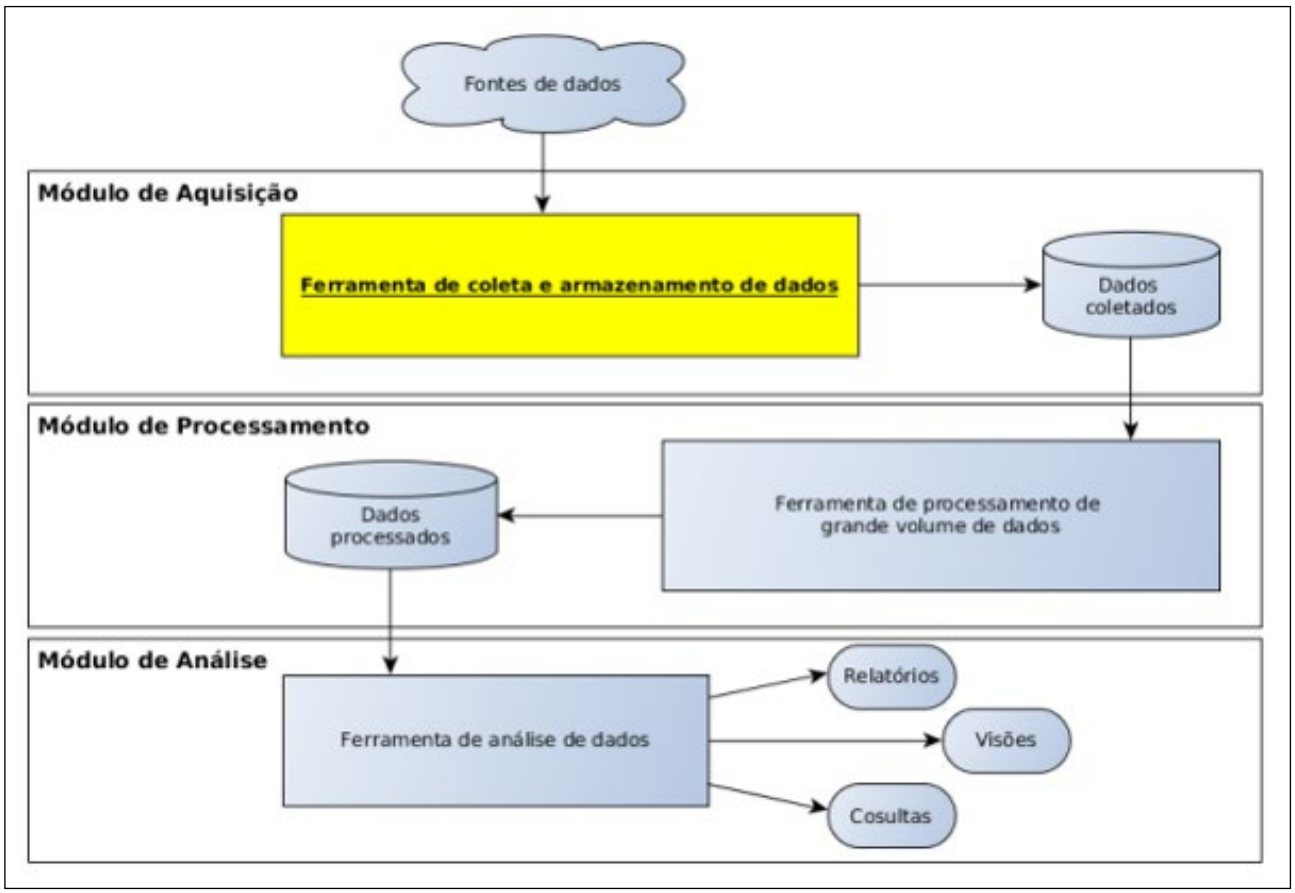
2 DESENVOLVIMENTO DA FERRAMENTA DE COLETA DE DADOS

A ferramenta proposta faz parte de um sistema completo para mineração de dados coletados de fontes disponíveis na *web*. Esse sistema será formado por três módulos principais: aquisição, processamento e análise.

O módulo de aquisição é o responsável pela obtenção e pelo armazenamento dos dados. Esse foi o módulo desenvolvido e é detalhado no decorrer deste artigo. O módulo de processamento deverá conter uma ferramenta que prepare os dados armazenados pelo módulo de aquisição para serem utilizados no módulo de análise. Essa preparação de dados pode abranger atividades como remover tipos de palavras dos textos, como preposições e artigos, encontrar e agrupar termos semelhantes e classificar os registros de acordo com seu contexto.

O módulo de análise deverá aplicar algoritmos e técnicas de mineração nos dados processados, possibilitando a geração de conhecimento a partir desse grande volume de dados obtidos. A Figura 1 mostra a arquitetura e a relação entre os módulos citados.

Figura 1 – Arquitetura da ferramenta para mineração de dados

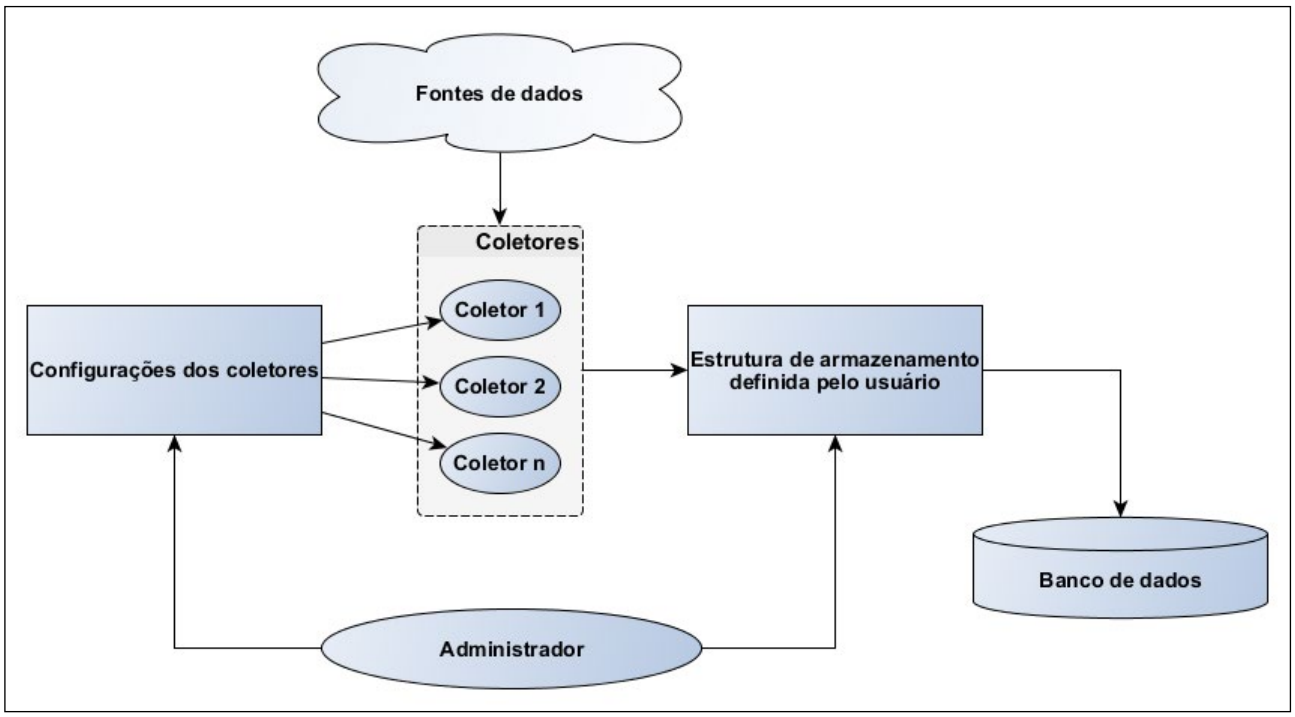


Fonte: Dos autores (2014).

Os resultados apresentados neste artigo compreendem apenas o desenvolvimento do módulo de coleta de dados. Considerando que qualquer análise só pode ser realizada a partir de um conjunto de dados disponível, este módulo é essencial para a continuidade do desenvolvimento da ferramenta proposta.

O módulo implementado é composto por dois módulos: coleta e configuração. O módulo de coleta é o responsável pela execução dos diversos coletores. Cada coletor é um processo isolado, ou seja, pode haver mais de um programa do mesmo tipo buscando dados ao mesmo tempo. Informações obtidas são armazenadas conforme as informações definidas pelo usuário por meio do módulo de configuração, no qual é possível definir diferentes coletores de dados e especificar os parâmetros para a execução deles, além de definir a estrutura de armazenamento dos dados coletados pelos coletores. Neste módulo também é possível cadastrar usuários para a utilização do programa e acessar uma tela para efetuar consultas nos dados. A Figura 2 mostra a estrutura da ferramenta descrita.

Figura 2 – Estrutura do módulo de captura de dados



Fonte: Dos autores (2014).

O módulo é composto por três tipos de programas coletores, responsáveis, respectivamente, por obter dados de três fontes: *Twitter, Facebook* e *Google Plus*. Cada um deles é um programa independente, utilizando recursos específicos conforme sua finalidade. Existe também uma rotina que é executada em segundo plano, cuja responsabilidade é executar os programas paralelamente.

O sistema foi implementados na linguagem de programação PHP, que é uma linguagem de programação voltada para aplicações *web*, e armazena os dados coletados em um servidor de banco de dados orientado a documentos. O Sistema Gerenciador de Banco de Dados (SGBD) escolhido para isso foi o MongoDB, pelo fato de ser livre para uso e não requerer um *hardware* específico para seu funcionamento, além dos três motivos citados por Islam (2011):

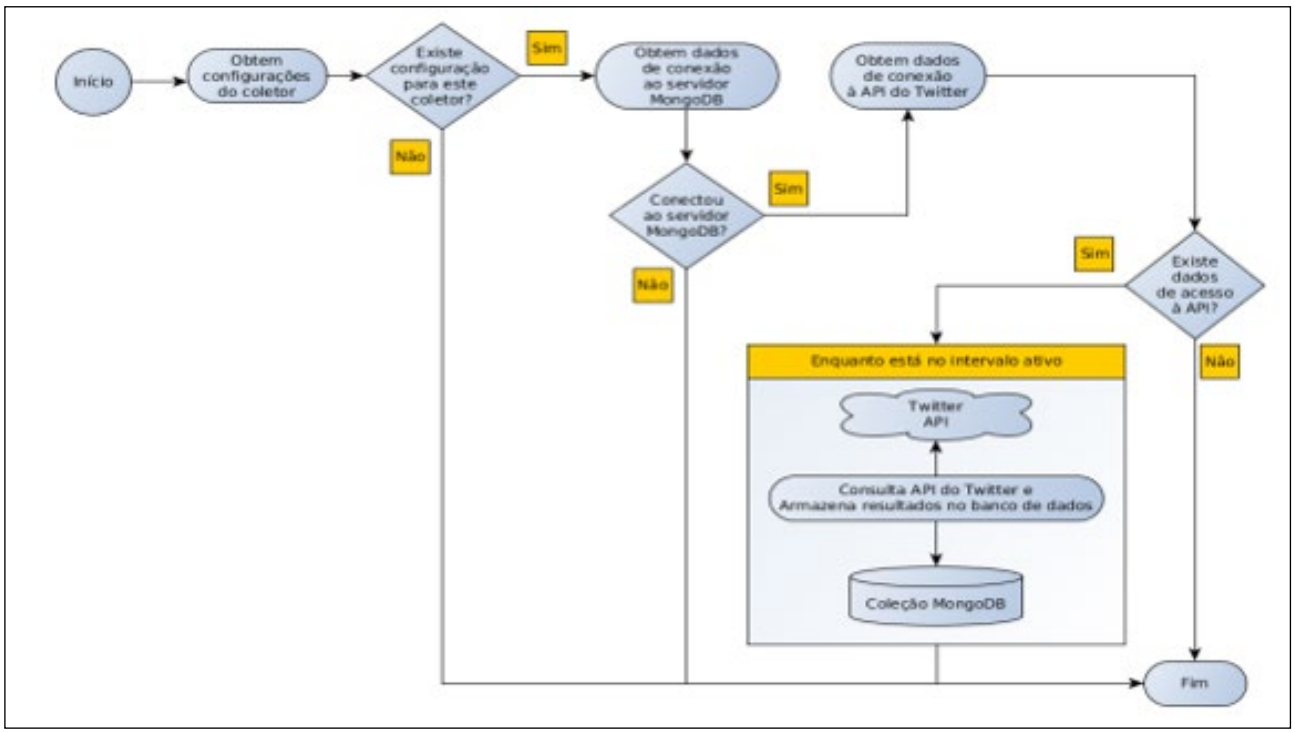
- implementa a ideia de esquema flexível. Não é preciso definir uma estrutura de dados antes de se começar a armazenar informações, o que o torna adequado para armazenar dados não estruturados, como os dados provenientes de redes sociais;

- é altamente escalável. O MongoDB vem com muitos recursos para ajudar a manter um bom desempenho, enquanto o tráfego e o volume de dados crescem com pouca ou nenhuma alteração na aplicação;

- é de fácil aprendizagem por ter muitos conceitos parecidos com bancos de dados relacionais.

O programa coletor para a rede social *Twitter* utiliza-se de uma *Application Programming Interface* (em português Interface de Programação de Aplicativos) (API) para interação com a fonte de dados. A API, que está na versão 1.1, é disponibilizada pelo próprio *Twitter* e a documentação de todos os métodos e funções existentes está disponível no endereço *https://dev.twitter.com/*. A Figura 3 ilustra o funcionamento da alternativa de captura para *Twitter*.

Figura 3 – Fluxo do programa coletor para *Twitter*



Fonte: Dos autores (2014).

O funcionamento do programa coletor ocorre da seguinte forma: inicialmente é recebido o código do coletor que foi cadastrado pelo usuário. Se for passado como parâmetro o código de um coletor que não estiver cadastrado, o programa é finalizado. Caso contrário o código é utilizado para carregar as configurações de conexão ao servidor MongoDB. Com esses dados, é feita uma tentativa de conexão ao servidor. Em caso de não haver sucesso nessa conexão, o programa é finalizado, caso contrário são carregadas as informações de acesso à API. Se não houvere dados para esse acesso, o programa é finalizado. Em caso contrário, verifica-se o horário está entre o intervalo cadastrado pelo usuário para o funcionamento desse coletor. Se o horário em que o coletor estiver sendo executado estiver no intervalo ativo definido pelo usuário, o programa utilizará os dados de consulta, como as palavras-chave, para efetuar consultas à API do *Twitter*.

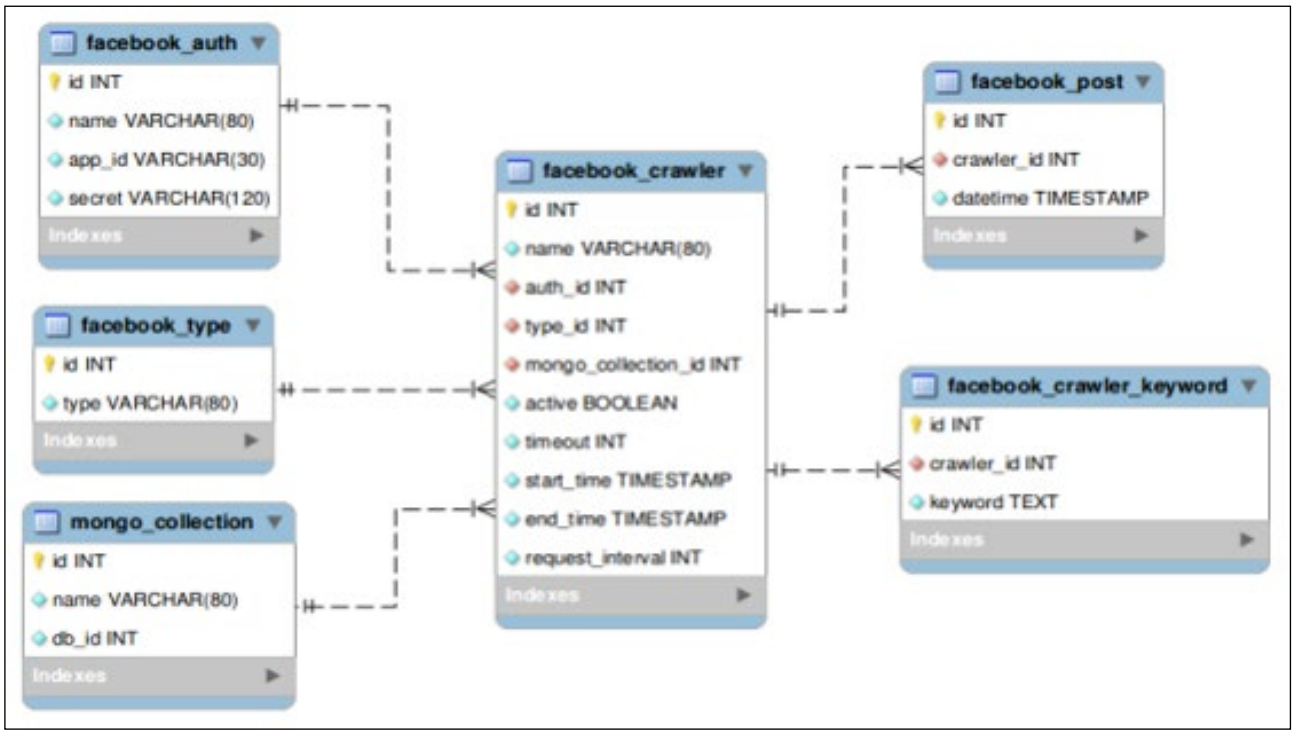
O resultado dessas consultas é armazenado na coleção definida pelo usuário para esse coletor. Essas consultas são realizadas em um intervalo de tempo definido pelo usuário (não menor do que um minuto), enquanto o programa estiver no intervalo ativo do coletor. Assim que o horário não estiver mais entre o intervalo ativo, o programa é finalizado.

O coletor de dados para o *Facebook* também utiliza uma API para acesso aos dados. Essa API é disponibilizada pelo *Facebook* e a documentação de todos os métodos e funções existentes está disponível no endereço *https://developers.facebook.com/*. O funcionamento do programa é idêntico ao coletor de dados do *Twitter*. A API oferecida pelo *Facebook* exige um código de aplicação (App ID) e uma chave de acesso (*App Secret*) para efetuar as consultas nas postagens públicas da rede social. Para obter essas informações, é necessário criar uma aplicação no *Facebook*, através do endereço *https://developers.facebook.com/*.

Para possibilitar a configuração dinâmica das ferramentas de coleta e o armazenamento dos dados obtidos, foram criadas tabelas em um Sistema de Gerência de Banco de Dados Relacional,

o PostGreSQL. Nessas tabelas existe uma referência para a coleção do MongoDB, no qual são armazenados os dados obtidos na captura.

Figura 4 – Modelo ER do coletor de dados do *Facebook*



Fonte: Dos autores (2014).

A Figura 4 mostra o modelo Entidade Relacionamento (ER) correspondente à estrutura para coletores do *Facebook*. A estrutura para os demais é semelhante.

Para o acesso aos dados da rede social *Google Plus*, o programa coletor desta fonte utiliza o projeto de código aberto "*google-api-php-client*", pois não há uma API oficial da *Google* na liguagem de programação PHP. Essa API está disponível no endereço *https://github.com/google/google-api-php-client*. O funcionamento desse programa é igual ao dos demais coletores de dados.

Para possibilitar a execução automática e paralela, foi desenvolvida uma rotina que é executada em segundo plano no servidor em que a ferramenta foi instalada. Esse programa tem a finalidade de, por meio das tabelas de configuração, obter a lista de todos os coletores cadastrados no sistema e, se não estiverem ativos, executá-los em um processo separado. Desse modo, garante-se que os coletores sempre estarão em execução se estiverem no intervalo de funcionamento especificado.

O módulo de configuração dos coletores é composto de uma interface gráfica em que o usuário pode gerenciar toda a ferramenta de coleta de dados. Nesse módulo é possível manter o cadastro de servidores, banco de dados e coleções MongoDB, assim como definir características de coletores para todas as fontes disponíveis no sistema desenvolvido. Também é possível acessar uma tela de consulta dos dados coletados pela ferramenta.

Essa interface é do tipo *web*, ou seja, é acessada por meio de um programa navegador, tal como o *Mozilla Firefox* ou o *Google Chrome*, sendo necessária a sua hospedagem em um servidor de

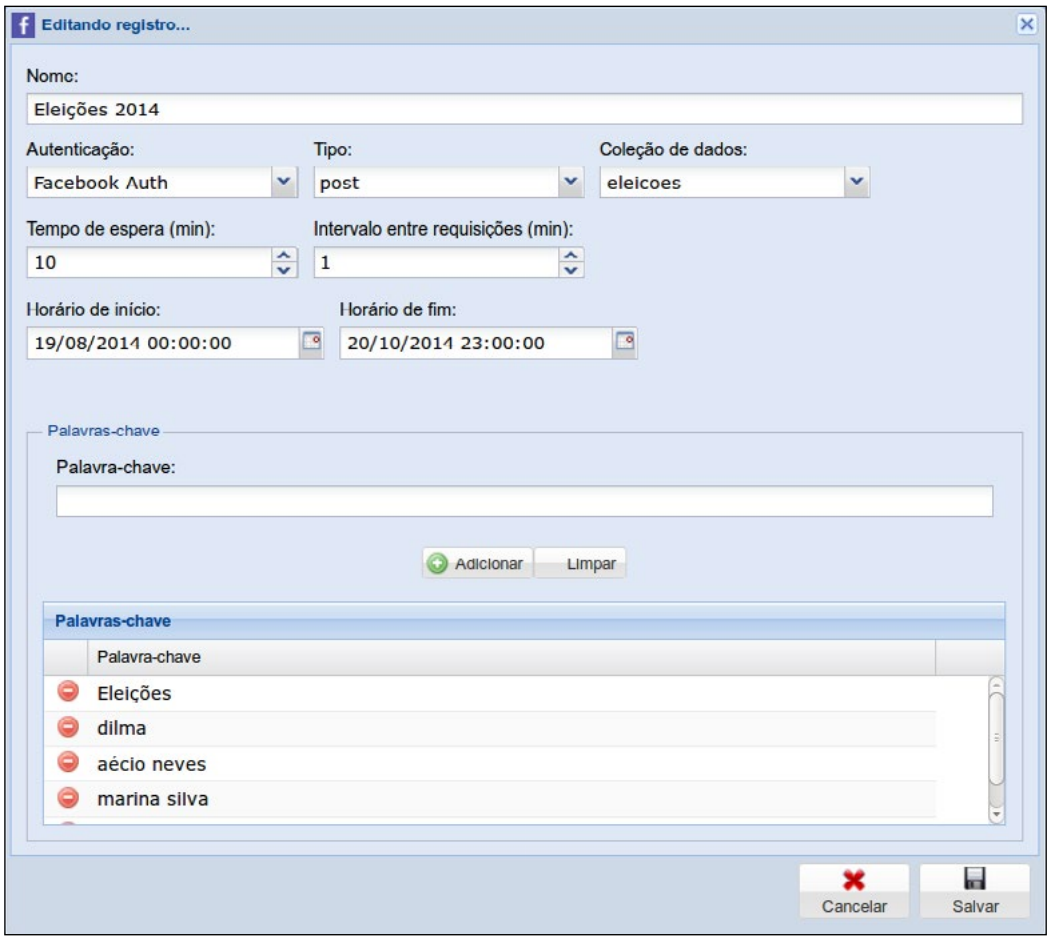
páginas. A escolha por esse tipo de interface deu-se por ser possível acessá-la de qualquer lugar e com qualquer dispositivo que possua um navegador e acesso à internet.

Para o desenvolvimento da interface *web*, foi utilizada a biblioteca Sencha ExtJS, escrita em javascript[2], para as interfaces de interação com o usuário. A escolha dessa biblioteca para a parte gráfica da aplicação deve-se ao fato de ela prover uma interface bastante amigável ao usuário, além de haver muita documentação disponível. Com mais de 100 exemplos, centenas de componentes, uma suíte de documentação completa e construído em temas, o ExtJS fornece as ferramentas necessárias para construir aplicações robustas (SENCHA, 2013).

Para permitir alteração dinâmica das funcionalidades, foi criada uma interface de configuração, mostrada na Figura 5. É possível observar informações relacionadas à autenticação na rede social que é alvo da captura, além de informações relativas ao início e ao fim do processamento.

Cada processo de captura é identificado por um nome, como neste exemplo, "Eleições 2014". Na área inferior da tela são listadas as palavras-chave, que podem ser inseridas na configuração. Elas serão utilizadas no processamento para identificar quais postagens serão obtidas.

Figura 5 – Interface de configuração de captura de dados para *Facebook*



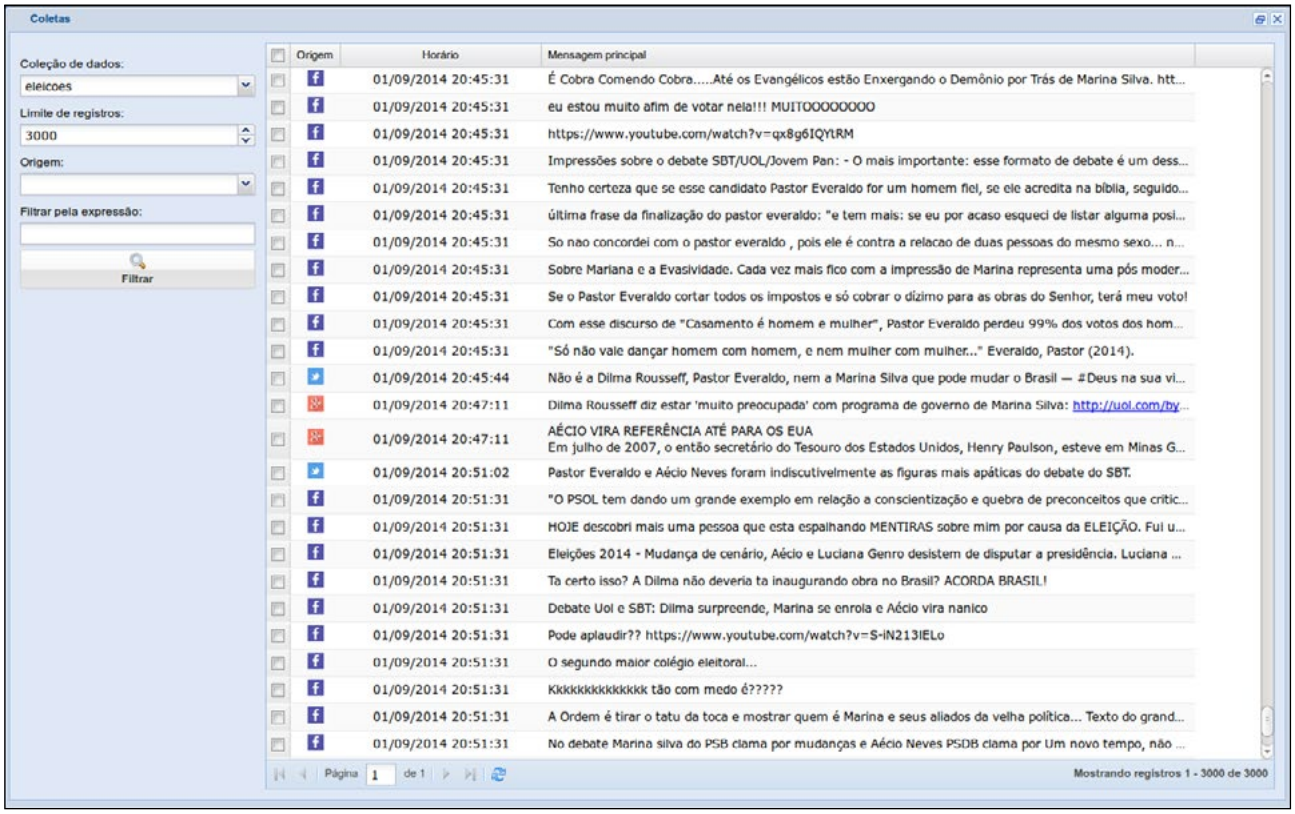
Fonte: Dos autores (2014).

2 Linguagem de programação executada pelo lado do cliente, no próprio navegador. Ou seja, por meio do javascript podem ser executados códigos sem a necessidade de comunicação com o servidor.

Após uma coleta ter sido realizada, é possível consultar as postagens e informações obtidas. Para esse fim foi criada uma tela de consulta, que mostra a lista de itens, com alternativa para ver detalhes de cada um deles (Fig.6).

Na imagem são exibidas informações relativas à data e hora em que as postagens foram buscadas e o texto principal de cada mensagem. O recurso de consulta permite também filtrar por termos, expressões, além de buscar registros em uma coleção específica. Cada coleção corresponde a uma determinada busca realizada, conforme a configuração indicada na Figura 5.

Figura 6 – Interface de consulta dos dados coletados



Fonte: Dos autores (2014).

3 RESULTADOS E DISCUSSÃO

Após o desenvolvimento da ferramenta, foram executadas algumas capturas de dados para verificar seu funcionamento e validar seus objetivos. Nesta seção são descritas algumas dessas coletas, além dos resultados obtidos.

A primeira coleta foi realizada em 17 dias, durante os meses de setembro e outubro de 2014, e teve como tema as eleições do mesmo ano. Essa coleta utilizou os coletores de dados para *Twitter*, *Facebook* e *Google+*, com as mesmas palavras-chave para todos: 'dilma', 'aécio' e 'eleições'. Na tabela a seguir, são apresentados dados sobre registros desta coleta.

Tabela 1 – Número de registros da primeira coleta

Coletor	nº Registros
Twitter	285.879
Facebook	63.645
Google+	2.185
Total	351.709

Fonte: Dos autores (2014).

Com o total de 351.709 registros coletados, essa coleta teve a média de 20.689 registros coletados por dia. O coletor de dados do *Twitter* foi o que mais coletou atividades dos usuários, principalmente pelo fato de todos os *tweets* serem de acesso público, um pouco diferente dos *posts* do *Facebook*, que dependem das configurações dos usuários para serem públicos ou não. Também deve-se ao fato de a API do *Facebook* restringir o número de registros por requisição em 100. A API do *Twitter* retorna 150 registros por requisição. E a baixa coleta do coletor do *Google Plus* deve-se ao fato de esta rede social ainda não ser muito utilizada e não possuir uma grande atividade por parte de seus usuários. Essa primeira coleta evidenciou que a ferramenta é capaz de adquirir um grande volume de dados.

A segunda coleta de validação do sistema desenvolvido foi efetuada nos dias 25, 26 e 27 de outubro de 2014 e buscou registros de duas marcas de refrigerantes, utilizando as palavras-chave 'pepsi' e 'coca-cola' para os três tipos de coletores. A Tabela 2 apresenta os resultados dessa coleta.

Tabela 2 – Número de registros da segunda coleta

Coletor	nº Registros
Twitter	5.349
Facebook	1.891
Google+	16
Total	7.256

Fonte: Dos autores (2014).

Essa coleta obteve um total de 7.256 registros utilizando as duas palavras-chave. Deles, 5.900 referiam-se ao termo 'coca-cola' e 1.356, ao termo 'pepsi'. A partir dos dados dessa coleta, uma ferramenta de análise pode descobrir informações sobre os consumidores dessas marcas, tais como se estão satisfeitos ou não com esses produtos, o quanto consomem ou o quanto gostariam de consumir, entre muitas outras análises possíveis.

A última coleta de validação da ferramenta buscou registros da palavra 'univates' nas fontes de dados e foi efetuada em cinco dias, durante os meses de outubro e novembro de 2014. Essa coleta evidenciou a facilidade da obtenção de assuntos que podem ser relevantes a uma instituição em uma determinada região, podendo-se obter um *feedback* de seu público quase que instantaneamente. A Tabela 3 apresenta os totais dessa coleta.

Tabela 3 – Número de registros da terceira coleta

Coletor	nº Registros
Twitter	801
Facebook	264
Google+	0
Total	1.065

Fonte: Dos autores (2014).

Com um total de 1.065 registros obtidos em apenas cinco períodos de coleta, a ferramenta desenvolvida, após essa última validação, mostrou-se capaz de atingir os objetivos propostos.

A Figura 7 mostra exemplos de registros coletados e armazenados no MongoDB. Nela observam-se a estrutura complexa e o conjunto de dados que compõem uma postagem em uma rede social como o *Twitter*. No exemplo dessa imagem, observam-se duas pessoas de diferentes lugares do país, que, ao expressarem sua opinião em uma rede social, fizeram parte da análise de dados utilizada como teste da ferramenta.

Pode-se constatar também que em cada *twitter* capturado existem diversas informações, além do texto principal da mensagem. Informações sobre o instante da criação, usuário, contagem de *retweets*, entre outros dados, que podem ser úteis em análises posteriores.

Figura 7 - Registros armazenados durante captura


```
{
    "_id" : ObjectId("5432b8f30b16375b0a8b6e5e"),
    "metadata" : { "iso_language_code" : "pt", "result_type" : "recent" },
    "created_at" : "Mon Oct 06 15:44:26 +0000 2014",
    "id" : NumberLong("519151200537944065"),
    "id_str" : "519151200537944065",
    "text" : "Vou votar na Dilma sim e não escondo. Sou do estado que tem Beto no governo e Álvaro Dias no senado.",
    "source" : "twitter",
    "truncated" : false,
    "user" : { "id" : NumberLong(53793436), "id_str" : "53793436", "name" : "cavanhinho", "screen_name" : "alcavanha", "location" : "Curitiba", "description"
    "geo" : null,
    "coordinates" : null,
    "place" : null,
    "contributors" : null,
    "retweet_count" : NumberLong(0),
    "favorite_count" : NumberLong(0),
    "entities" : { "hashtags" : [ ], "symbols" : [ ], "urls" : [ ], "user_mentions" : [ ] },
    "favorited" : false,
    "retweeted" : false,
    "lang" : "pt"
}

{
    "_id" : ObjectId("544c0ad40b163702078b706a"),
    "metadata" : { "iso_language_code" : "pt", "result_type" : "recent" },
    "created_at" : "Sat Oct 25 20:40:26 +0000 2014",
    "id" : NumberLong("526111060244586496"),
    "id_str" : "526111060244586496",
    "text" : "Eu não sou influenciado por pesquisas eleitorais. Eu vou votar no Aécio para presidente. #VotoAecioPeloBR45IL",
    "source" : "twitter",
    "truncated" : false,
    "user" : { "id" : NumberLong(294261353), "id_str" : "294261353", "name" : "marcos ", "screen_name" : "marcos_bezerra_", "location" : "Rio de Janeiro", "p
    "geo" : null,
    "coordinates" : null,
    "place" : { "id" : "e433fbca595f29e5", "url" : "https://api.twitter.com/1.1/geo/id/e433fbca595f29e5.json", "place_type" : "admin", "name" : "Rio de Janei
    "contributors" : null,
    "retweet_count" : NumberLong(0),
    "favorite_count" : NumberLong(0),
    "entities" : { "hashtags" : [ { "text" : "VotoAecioPeloBR45IL", "indices" : [ NumberLong(89), NumberLong(109) ] } ], "symbols" : [ ], "urls" : [ ], "user
    "favorited" : false,
    "retweeted" : false,
    "lang" : "pt"
}
```


Fonte: Dos autores (2014).

4 CONSIDERAÇÕES FINAIS

O estudo realizado com base em diversas fontes bibliográficas evidenciou a oportunidade de organizações adquirirem conhecimentos importantes para a melhoria contínua de seus processos de negócio, por meio da análise de grandes volumes de dados por sistemas de *Big Data*.

Ao propor o desenvolvimento de uma ferramenta de coleta de dados de diversas fontes, o autor busca atender a um requisito comum a todos os sistemas de *Big Data*: a coleta e o armazenamento de grande volume de dados, para posterior análise e mineração de informações relevantes.

A ferramenta de coleta de dados apresentada no presente artigo atinge os objetivos propostos, sendo uma contribuição para a criação de um sistema de análise de grande volume de dados em trabalhos futuros, podendo ser utilizada e/ou ampliada conforme a necessidade.

REFERÊNCIAS

CARVALHO, L. A. V. Data Mining: **A Mineração de dados no Marketing**. Rio de Janeiro: Editora Ciência Moderna, 2005.

CHODOROW, K. **MongoDB: The Definitive Guide**. 2.ed. Sebastopol: O'Reilly Media, 2013.

DUMBILL, E. **Big Data Now**: 2012 Edition. Sebastopol: O'Reilly Media, 2012.

EATON, C.; DEROOS, D.; DEUTSCH, T.; LAPIS, G.; ZIKOPOULOS, P. **Understanding Big Data**: **Analytics for Enterprise Class Hadoop and Streaming Data**. The McGraw-Hill Companies, 2012.

HURWITZ, J.; NUGENT, A.; HALPER, F.; KAUFMAN, M. **Big Data for Dummies**: A Wiley Brand. Hoboken: John Wiley & Sons, 2013.

ISLAM, R. **PHP and MongoDB Web Development**: Begginer's Guide. Birmingham: Packt Publisinhg, 2011.

LIU, B. **Web Data Mining**: Exploring Hyperlinks, Contents and Usage Data. 2.ed. Chicago: Springer, 2011.

SENCHA. **Sencha ExtJS**: Javascript Framework for Rich Desktop Apps. Disponível em: http://www.sencha.com/products/extjs/ Acesso em: 02 nov. 2013.

VIEIRA, M. R.; FIGUEIREDO, J. M.; LIBERATTI, G.; VIEBRANTZ, A. F. M. **Bancos de dados NoSQL: Conceitos, ferramentas, linguagens e estudos de caso no contexto de Big Data**. Simpósio Brasileiro de Bancos de Dados. 2012.