

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Business Intelligence

Marcelo de Barros Gaspar

**RECUPERAÇÃO DE DADOS EM MÍDIAS SOCIAIS SEM VIOLAÇÃO DE
PRIVACIDADE**

Belo Horizonte

2018

Marcelo de Barros Gaspar

**RECUPERAÇÃO DE DADOS EM MÍDIAS SOCIAIS SEM VIOLAÇÃO DE
PRIVACIDADE**

TCC apresentado ao Programa de Pós-Graduação em Business Intelligence da Pontifícia Universidade Católica de Minas Gerais, como requisito para obtenção do título de Especialista em Business Intelligence.

Orientador: Prof. Cristiano Carvalho

Belo Horizonte

2018

Marcelo de Barros Gaspar

TRABALHO DE CONCLUSÃO DE CURSO:

Recuperação de Dados em Mídias Sociais sem Violação de Privacidade

TCC apresentado ao Programa de Pós-Graduação em Business Intelligence da Pontifícia Universidade Católica de Minas Gerais, como requisito para obtenção do título de Especialista em Business Intelligence.

Orientador: Prof. Cristiano Carvalho

Prof. Cristiano Carvalho - PUC Minas (Orientador)

Belo Horizonte, 11 de abril de 2018

RESUMO

Recuperar e minerar dados obtidos de mídias sociais se tornou uma prática extremamente comum entre empresários que almejam alcançar seus objetivos, porém coletar dados de terceiros é uma tarefa que demanda responsabilidade e coerência por parte de profissionais de Tecnologia da Informação. Este trabalho possui por finalidade detalhar a instalação e configuração de ferramentas Open Source como o Apache NiFi e Solr, para que um ambiente Linux seja capaz de coletar dados de forma legal, sem infringir o Marco Civil da Internet ou a privacidade de daqueles que produziram o conteúdo coletado. O processo de coleta de dados a partir do programa Apache Nifi se baseia na elaboração de um fluxo capaz de identificar termos e idiomas, para que uma base de dados possa ser gerada para atender demandas específicas. Uma vez coletados, os dados estarão disponíveis para consulta através do Solr.

Palavras-chave: Mineração de Dados. Coleta de Dados, Mídias Sociais.

SUMÁRIO

1 Introdução	1
2 Mídias Sociais	2
2.1 Políticas de Privacidade em Mídias Sociais	2
2.2 Tipos de Usuário do Facebook	4
2.3 API	5
3 Marco Civil da Internet	5
3.1 Invasão de Privacidade Segundo o Marco Civil da Internet	5
4 Coletando Dados de Mídias Sociais	6
4.1 Coletando dados através de questionários virtuais	6
4.2 Coletando Dados Através de Testes de Facebook	7
4.3 Coletando Dados Através de Aplicativos do Facebook	8
4.3.1 Escândalo Envolvendo o Vazamento de Dados Pessoais de Usuários do Facebook	10
4.4 Coletando dados através de aplicativos Open Source	11
4.4.1 Preparando o ambiente para a instalação dos serviços Apache Nifi e Solr	11
4.4.2 Instalando Apache Nifi	13
4.4.3 Instalando Solr	14
4.4.4 Configurando workflow do Apache Nifi para coletar dados do Twitter	15
4.4.6 Verificando dados coletados através do Apache Nifi com o Solr	22
5 Conclusão	23
6 Referências Bibliográficas	24
Emerson Alecrim. A controvérsia dos 50 milhões de perfis do Facebook manipulados pela Cambridge Analytica. Disponível em: < https://tecnoblog.net/236612/facebook-cambridge-analytica-dados/ > Acesso em 14 abr. 2018.	25

1 Introdução

Com mais de 2 bilhões de usuários ao redor do mundo, sendo 150 milhões somente no Brasil, mídias sociais como o Facebook e Twitter atraem a atenção de empreendedores que almejam um maior retorno financeiro para seus investimentos.

Diante da recessão econômica que impactou, a partir de 2015, de forma extremamente negativa, no faturamento do comércio brasileiro, empreendedores começaram a buscar nas mídias sociais publicações que envolvessem o nível de satisfação de seus clientes, atualmente considerada uma prática extremamente comum entre investidores que possuam visão de futuro e a intenção de expandir seus negócios.

Partindo do princípio que a propaganda boca a boca é a melhor forma de se divulgar o nome de qualquer produto ou estabelecimento comercial, a implantação de técnicas que possibilitam a mineração e recuperação de dados deixou de ser classificada por empreendedores como mais um gasto operacional, passando a ser considerada como um investimento estratégico, realizado no sentido de se fortalecer no mercado.

Cientes de que objetivos empresariais possam ser alcançados com maior rapidez, sem a necessidade da aplicação da tradicional, e onerosa, fórmula de tentativa e erro, visionários apostam na mineração de dados para alavancarem seus empreendimentos, pois desta forma saberão não somente como alcançar com mais precisão seu público alvo, mas também como identificar possíveis críticas envolvendo suas marcas, para que uma solução pós-venda possa ser oferecida caso seja necessário, evitando futuras quedas de faturamento.

Embora coletar dados publicados em redes sociais seja uma demanda crescente entre profissionais de Tecnologia da Informação, a preocupação com a invasão de privacidade dos usuários de mídias sociais, cujos dados estão sendo coletados para fins comerciais, não acompanha o mesmo ritmo de crescimento.

Nesse contexto, a proposta deste trabalho de conclusão de curso é apresentar conceitos, definições e metodologias capazes de recuperar dados de mídias sociais com responsabilidade, tarefas que deverão sempre respeitar as configurações de privacidades previamente configuradas por proprietários de perfis em mídias sociais.

2 Mídias Sociais

Programadas para serem plataformas sociais, ou seja, um ambiente virtual que conecte pessoas com interesses em comum, as mídias sociais com o decorrer dos anos se tornaram muito mais do que uma simples ferramenta de lazer e interação pessoal, tendo em vista que suas aplicações abrangem também soluções de marketing online para empresas que buscam por maior visibilidade.

Atualmente são raros os exemplos de pessoas físicas e jurídicas que não possuam sequer um perfil em mídias sociais, membros de uma nova sociedade nominada 2.0, onde a cede por informações é tão grande que a conectividade se tornou praticamente um sinal vital.

2.1 Políticas de Privacidade em Mídias Sociais

Para se cadastrar em plataformas sociais como o Facebook, ou Twitter por exemplo, os novos usuários deverão preencher um simples cadastro que solicita informações básicas, como seu nome completo, e-mail e telefone, formulário que ao seu final conterá uma Política de Dados, automaticamente aceita toda vez que uma nova conta é criada.

Ao finalizar o cadastro, os novos usuários de mídias sociais renunciam ao direito autoral de todo conteúdo por ele publicado, que passa a ser de propriedade da plataforma social escolhida para se relacionar, sendo assim qualquer postagem poderá ser coletada para análise e enviada a parceiros.

Imagem 1: Termo para cadastro no Facebook

O que você faz e as informações que fornece.

Coletamos o conteúdo e outras informações fornecidas por você quando usa nossos Serviços, como quando se cadastra em uma conta, cria ou compartilha conteúdos, envia mensagens ou se comunica com os outros. Isso pode incluir informações presentes no conteúdo ou a respeito dele, como a localização de uma foto ou a data em que um arquivo foi criado. Também coletamos informações sobre como você usa nossos Serviços, por exemplo, os tipos de conteúdo que você vê ou com que se envolve e a frequência ou duração de suas atividades.

Fonte: <https://www.facebook.com/about/privacy>

Imagem 2: Termo para cadastro no Twitter

Informações públicas: Podemos compartilhar ou divulgar suas informações públicas, como suas informações de perfil de usuário público, Tweets públicos, ou as pessoas que você segue ou que seguem você. Lembre-se: suas configurações de privacidade e visibilidade controlam se seus Tweets e determinadas informações de perfil são tornados públicos. Outras informações, como seu nome e nome de usuário, sempre são públicas no Twitter, a menos que você exclua sua conta, conforme descrito abaixo.

Fonte: <https://twitter.com/pt/privacy>

Para atender as exigências de usuários preocupados com a violação de privacidade, plataformas sociais oferecem a opção de acesso a publicações antes mesmo que elas sejam realizadas, podendo o autor escolher entre: uma postagem pública, que estará disponíveis a todos; ou entre amigos.

Imagem 3: Política de privacidade para publicações no Facebook



Fonte: <https://www.facebook.com>

Imagem 4: Política de privacidade para publicações no Twitter

Privacidade dos Tweets ☒ **Proteger seus Tweets**

Se a opção for selecionada, somente as pessoas que você autorizar receberão seus Tweets. Seus futuros Tweets não estarão disponíveis publicamente. Os Tweets publicados anteriormente ainda poderão ser vistos publicamente em alguns lugares. [Saiba mais.](#)

Fonte: <https://twitter.com/settings/safety>

2.2 Tipos de Usuário do Facebook

Como qualquer plataforma virtual, as mídias sociais possuem diferentes classificações para seus usuários, que poderão ser cadastrados como:

- Pessoa Física: perfil pessoal, onde por via de regra será apenas divulgado publicações de interesse pessoal, sem fins comerciais;
- Negócio ou Estabelecimento Comercial: perfil voltado para a divulgação de uma empresa de pequeno porte, que possui apenas um ponto comercial, atuando somente em apenas uma cidade;
- Empresa, Organização ou Instituição: perfil voltado para atender as demandas de uma empresa de grande porte, de forma que possa alcançar seu público nacionalmente.
- Marca ou Produto: perfil voltado para divulgação de um produto específico.
- Artista, Banda ou Figura Pública: perfil voltado para pessoas públicas e artistas, de forma que projetos e shows possam ser divulgados para usuários interessados em cultura ou uma específica área;
- Entretenimento: perfil voltado para a divulgação de programas de TV, filmes em cartaz nos cinemas, bibliotecas, festivais em geral, premiações e séries de TV.
- Causa ou Comunidade: perfil voltado para a divulgação de ONGS ou ações comunitárias.

Para atender a demanda de cada usuário cadastrado, plataformas sociais como o Facebook por exemplo, possuem diferentes funcionalidades para cada tipo de usuário, como demonstrado na figura abaixo:

Imagem 5: Funcionalidades por tipo de usuário do Facebook

	Livros e revistas, marcas e produtos	Empresas e organizações	Serviços locais	Filmes, música, televisão	Pessoas, desportos	Sites e blogs
Breve descrição	✓	✓	✓	✓	✓	✓
Site	✓	✓	✓	✓	✓	✓
Serviços	✓	✓	✓	✓	✓	✓
Classificações e críticas	✓	✓	✓	✓	✓	✓
E-mail	✓	✓	✓	✓	✓	✓
Telefone	✓	✓	✓	✓	✓	✓
Morada		✓	✓		✓	
Mapa		✓	✓		✓	
Horário de funcionamento		✓	✓		✓	
Visitas		✓	✓			

Fonte: <https://www.facebook.com>

2.3 API

Para se coletar, analisar ou recuperar dados de qualquer plataforma social, a utilização de uma chave chamada de API (Interface de Programação de Aplicações) é obrigatória, credencial que somente poderá ser adquirida por usuários previamente cadastrados na mídia social onde pretende se conectar.

Muito utilizada por desenvolvedores de aplicações e cientistas de dados, a utilização do API abriu as portas para a coleta de dados publicados em mídias sociais.

3 Marco Civil da Internet

Elaborada para dar maior segurança jurídica, e normatizar a utilização da internet em todo território nacional, a Lei N° 12.965/14, aprovada pelo Congresso Nacional, foi sancionada pelo Palácio do Planalto com a finalidade de garantir não somente a liberdade de expressão de seus usuários, mas também o direito da privacidade de dados pessoais enviados através da web.

3.1 Invasão de Privacidade Segundo o Marco Civil da Internet

Questões que abrangem a invasão de privacidade fazem parte de um dos pilares mais importantes do Marco Civil da Internet, conforme consta em seu artigo terceiro: “Em qualquer operação de coleta, armazenamento, guarda e tratamento de registros, de dados pessoais ou de comunicações por provedores de conexão e de aplicações de internet em que pelo menos um desses atos ocorra em território nacional, deverão ser obrigatoriamente

respeitados a legislação brasileira e os direitos à privacidade, à proteção dos dados pessoais e ao sigilo das comunicações privadas e dos registros”.

De acordo com a Legislação vigente, a publicação da “Política de Privacidade” e “Termos de Uso” por parte de empresas que prestam qualquer serviço online é obrigatório, informações que deverão conter a política de acesso aos dados enviados ou publicados pela internet.

No caso do Twitter por exemplo, ao se cadastrar os novos membros são informados de que toda e qualquer postagem poderá ser coletada para análise, e enviada a parceiros, tornando a coleta de dados para as postagens públicas, com a utilização do API disponibilizado pela plataforma, uma ação legal.

Imagem 6: Termo para cadastro no Twitter

Consentimento ou Orientação do Usuário: O Twitter pode compartilhar ou divulgar suas informações de acordo com as suas orientações, por exemplo, quando você autoriza que um cliente Web ou aplicativo de terceiros acesse a sua conta, ou quando você nos orienta a compartilhar seu feedback com uma empresa. Quando você utiliza o Digits by Twitter para inscrição ou acesso a um aplicativo de terceiros, você está orientando o Twitter a compartilhar suas informações de contato, tais como seu número de telefone, com esse aplicativo. Se você compartilhou informações, como Mensagens Diretas ou Tweets protegidos, com outro usuário que acessa o Twitter utilizando um serviço de terceiros, lembre-se de que as informações poderão ser compartilhadas com o serviço de terceiros.

Fonte: <https://twitter.com/pt/privacy>

4 Coletando Dados de Mídias Sociais

Atualmente existem diferentes formas de se coletar e analisar dados provindos de mídias sociais como o Facebook por exemplo.

4.1 Coletando dados através de questionários virtuais

Coletar dados através de questionários virtuais tornou-se uma prática extremamente comum entre empresas que buscam alcançar seu público alvo com o auxílio de dispositivos móveis. Esta solução pode ser perfeita para uma pesquisa de mercado, por exemplo.

Existem serviços online, como o Survey Monkey, que oferecem a facilidade de se criar questionários ou enquetes que poderão ser publicadas em mídias sociais, e os resultados exportados em arquivos Excel.

Utilizar dados coletados através de questionários virtuais, para enriquecer ainda mais a análise de um determinado tópico, não invade a privacidade de internautas que decidiram preenche-los, tendo em vista que sua participação é opcional.

Uma vez publicado, o questionário virtual estará disponível para que qualquer internauta o preencha, portando o impulsionamento de tais publicações é extremamente recomendável, para que possa alcançar o número máximo de participantes.

Outra forma muito comum de se realizar tais questionários, seria através do Google Forms.

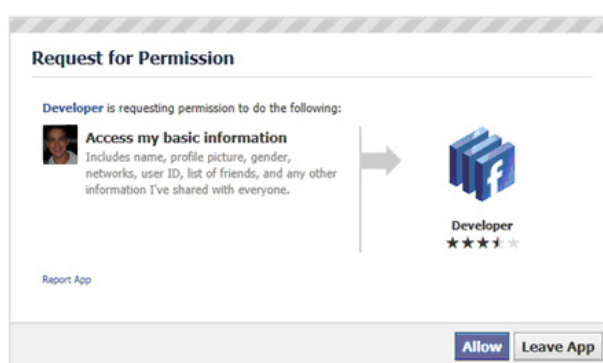
4.2 Coletando Dados Através de Testes de Facebook

Com uma maior adesão de jovens e adolescentes no Facebook, os testes se tornaram uma grande ferramenta de coleta de dados, muitas vezes utilizados para coletar dados sem que os usuários percebam o perigo por trás de tais ferramentas.

Desenvolvidos, na maioria das vezes, com a finalidade de coletar dados do usuário, os testes escondem por trás da diversão e interação com o internauta uma ferramenta capaz de acessar publicações e coletar dados pessoais como a data de nascimento e endereço de e-mail.

Para a realização de tais testes, o usuário interessado deverá concordar em dar acesso a seu perfil, permitindo com que o desenvolvedor seja considerado pelo Facebook como de sua confiança.

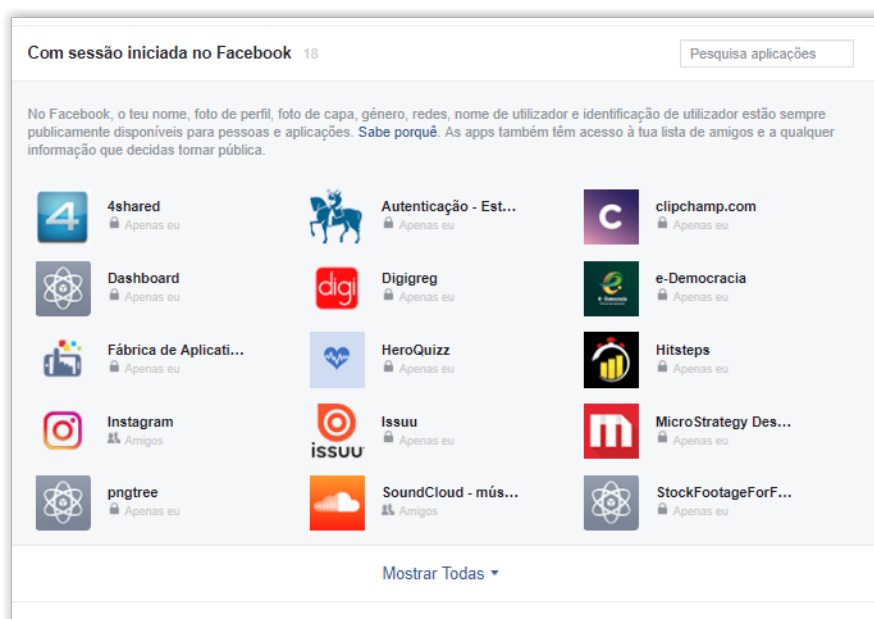
Imagem 7: Exemplo de permissão para acessar informações pessoais no Facebook



Fonte: <https://www.facebook.com>

Uma vez aceito, a empresa responsável pelo teste poderá acessar o perfil do participante para coletar dados, e em alguns casos também de seus amigos. O desenvolvimento de tais testes apenas se torna possível a partir da utilização da API disponibilizada para desenvolvedores pelo Facebook.

Imagem 8: Exemplo do painel que informa quais desenvolvedores possuem autorização para acessar informações pessoais no Facebook



Fonte: <https://www.facebook.com>

Mesmo assim, informando aos participantes que dados serão coletados, e seus perfis analisados, tal feito é considerado como invasão de privacidade, deixando de ser ético para ser criminoso, pois mesmo que o usuário cancele sua conta no Facebook a empresa responsável pelo teste continuará utilizando suas informações para fins comerciais.

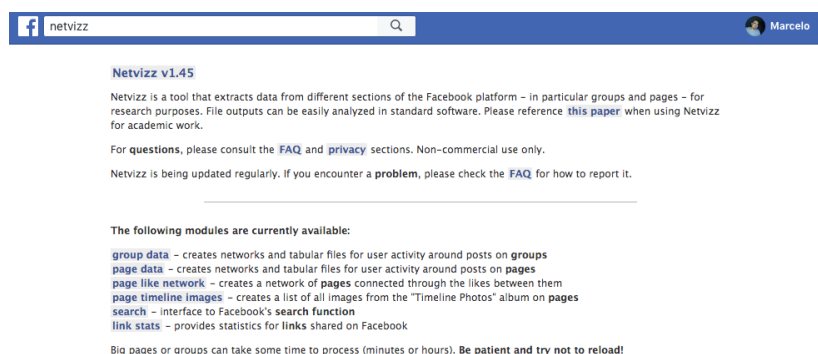
O conhecimento gerado a partir da participação de internautas em testes do Facebook é capaz de influenciar na opinião pública, mudando o rumo de campanhas eleitorais por exemplo.

Uma vez coletados, os dados servirão de apoio para que empresas possam atingir seus objetivos com maior precisão, pois saberão exatamente como alcançar seu público alvo e manipular sua opinião.

4.3 Coletando Dados Através de Aplicativos do Facebook

Embora existam softwares como o Power BI, Tableau, e Qlikview, capazes de se conectarem com o Facebook no intuito de coletar dados de sua página pessoal, um dos aplicativos mais utilizados para a coleta de dados de grupos, perfis e comunidades sempre foi a NetVizz.

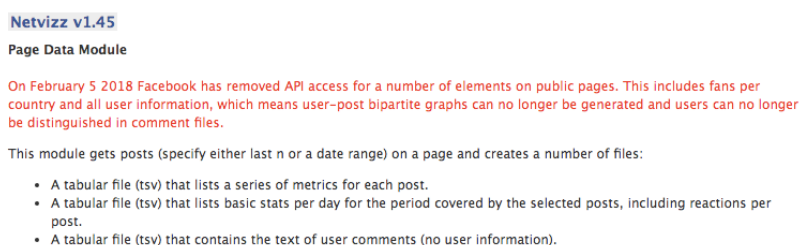
Imagem 9: Aplicativo NetVizz



Fonte: <https://apps.facebook.com/netvizz/>

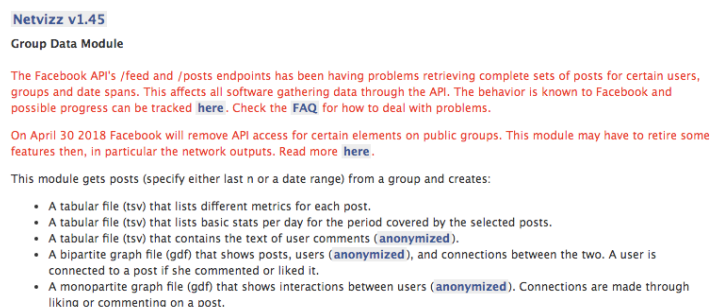
Desenvolvido através do Facebook, com o objetivo de servir como ferramenta de apoio para cientistas de dados, os dados coletados através do NetVizz poderiam facilmente ser utilizados também para manipular a opinião pública, durante campanhas eleitorais por exemplo, fazendo com que o API utilizado em seu desenvolvimento fosse rapidamente limitado, tendo algumas de suas principais funções, como a coleta de dados de páginas e grupos, desabilitadas.

Imagem 10: Anúncio da remoção da funcionalidade Page Data Module



Fonte: <https://apps.facebook.com/netvizz/>

Imagem 11: Anúncio da remoção da funcionalidade Group Data Module



Fonte: <https://apps.facebook.com/netvizz/>

Diante do último escândalo envolvendo o vazamento de dados pessoais de usuários do Facebook, a plataforma social foi obrigada a se reestruturar, pois a sua desvalorização fez com que revissem a política de privacidade, tornando a coleta de dados uma tarefa extremamente limitada, proibindo mecanismo de terceiros que acessem informações pessoais como religião, orientação política, status de relacionamento, formação educacional e profissional, exercícios físicos, músicas, livros, notícias, vídeos e jogos.

4.3.1 Escândalo Envolvendo o Vazamento de Dados Pessoais de Usuários do Facebook

O escândalo que chocou usuários do Facebook ao redor do mundo, e investidores da plataforma social, fez com que a eficácia da Política de Privacidade de uma das maiores mídias sociais disponíveis na internet fosse questionada.

A polêmica gerou ao redor da coleta de informações em perfis de usuários através da antiga Permissão para Amigos, desabilitada pelo Facebook em 2015, que permitia com que um determinado usuário autorizasse a coleta de dados no perfil de todos aqueles em sua corrente de amigos, fazendo então com que estas pessoas fossem vítimas de aplicativos que por traz de jogos e testes coletavam dados pessoais com intenções comerciais, sem seus consentimentos, tornando a prática questionável.

Mesmo desabilitada em 2015, a permissão para amigos já havia sido utilizada por um aplicativo chamado “This is Your Digital Life” (Esta é a sua Vida Digital), desenvolvido pelo psicólogo Aleksandr Kogan, da Universidade de Cambridge, que coletou informações pessoais de aproximadamente 50 milhões de usuários, base de dados que foi então vendida para Cambridge Analytica, empresa que utilizou estes dados em campanhas eleitorais como a do atual Presidente dos Estados Unidos Donald Trump.

O caso chamou a atenção do parlamento americano, fazendo com que funcionários, ex-funcionários e até mesmo o Presidente do Facebook, o programador Mark Zuckerberg, fosse questionado quanto as Políticas de Privacidade e segurança dos dados armazenados pelo Facebook.

“Nos meus 16 meses no Facebook, não me lembro de terem realizado uma só auditoria nos desenvolvedores de aplicativos Facebook” (Sandy Parakilas, Ex-Gerente de Operações do Facebook, 2018)

Embora coletar dados providos do Facebook não seja uma prática ilegal, vendê-los a torna ilícita, pois as condições de uso da plataforma proíbem o compartilhamento de dados com terceiros, codenando imediatamente o aplicativo “This is Your Digital Life”.

4.4 Coletando dados através de aplicativos Open Source

Tendo em vista que a coleta de dados providos do Facebook se tornou limitada, será demonstrado neste documento como se coletar dados de outras plataformas sociais como o Twitter por exemplo, utilizando softwares livres como o Apache NiFi.

Desenvolvido com a finalidade de automatizar o fluxo de dados entre sistemas de software, possibilitando a recuperação de dados providos de websites e plataformas sociais, o Apache Nifi se tornou uma ferramenta de grande valia para cientistas de dados.

Uma das principais vantagens de se utilizar o Apache Nifi, além da possibilidade de se montar uma infraestrutura capaz de coletar dados sem interrupções, é o fato de não exigir o conhecimento aprofundado em nenhuma linguagem de programação.

Outro programa que será utilizado é o Apache Solr, um projeto audacioso desenvolvido com a finalidade de armazenar e procesar dados textuais, permitindo a indexação e consultas dinâmicas.

4.4.1 Preparando o ambiente para a instalação dos serviços Apache Nifi e Solr

Antes de instalar os programas Apache NiFi e Solr, o ambiente onde os serviços serão hospedados deve ser devidamente configurado, neste caso um servidor rodando o sistema operacional Linux, Debian OS 9.4 64 bits.

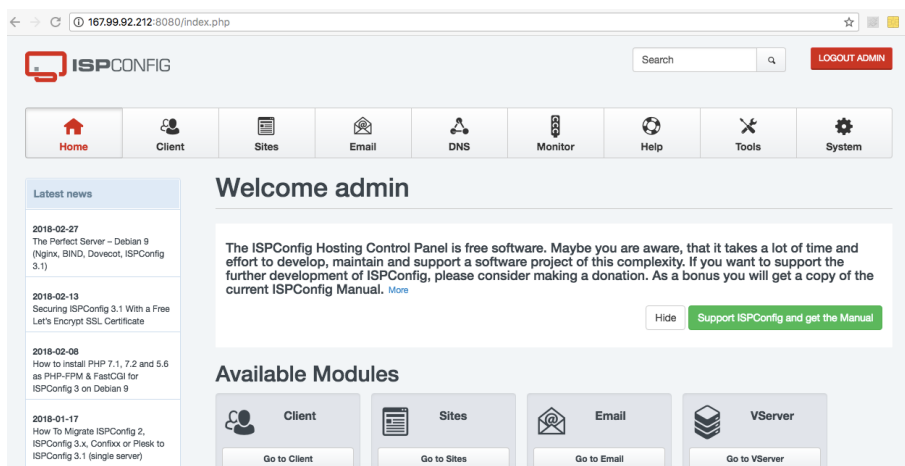
Tendo em vista que os serviços a serem instalados são acessados via web, o primeiro passo será instalar um painel de controle de domínios, neste caso o ISPCONFIG 3.

Para instalar o ISPCONFIG basta acessar o terminal de controle Linux e realizar o download gratuito, através do comando `wget --no-check-certificate https://github.com/servisys/ispconfig_setup/archive/master.zip`.

Após a finalização do download será possível descompactar o arquivo de instalação através digitando o comando `unzip máster.zip`, e iniciar a instalação através do comando `cd ispconfig_setup-master/./install.sh`.

Após o término da instalação o painel de controle do ISPCONFIG poderá ser acessado através do link <https://localhost:8080>.

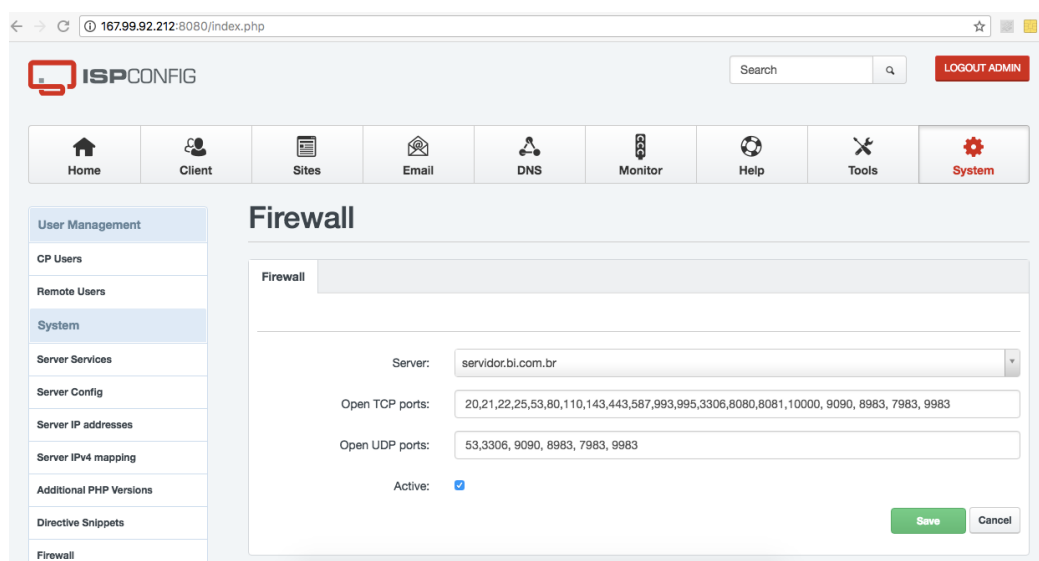
Imagem 12: Paine de controle ISPCONFIG



Fonte: <https://localhost:8080>

Antes de se instalar o Apache NiFi e Solr, deve-se abrir as portas necessárias para que os serviços a serem instalados funcionem corretamente, neste caso em específico abriremos as portas 9090 para o NiFi e 8983, 7983, 9983 para o Solr.

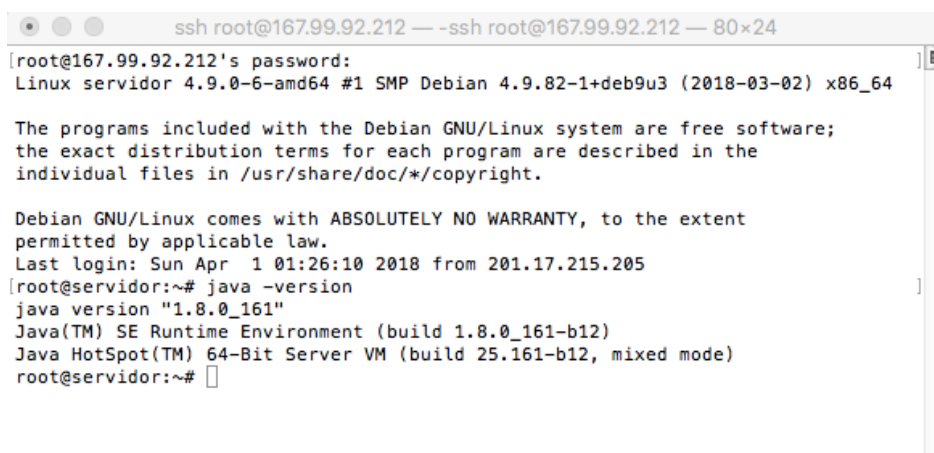
Imagem 13: Configurando firewall no ISPCONFIG



Fonte: <https://localhost:8080>

Uma vez configurado o firewall, deve-se verificar se o servidor possui o Java instalado, para verificar acesse o terminal de controle Linux e digite o comando `java -version`, caso não o possua instalado digite o comando `sudo apt-get install oracle-java8-installer` para instalar.

Imagem 14: Verificando instalação do Java



```
ssh root@167.99.92.212 — -ssh root@167.99.92.212 — 80x24
[root@167.99.92.212's password:
Linux servidor 4.9.0-6-amd64 #1 SMP Debian 4.9.82-1+deb9u3 (2018-03-02) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sun Apr  1 01:26:10 2018 from 201.17.215.205
[root@servidor:~# java -version
java version "1.8.0_161"
Java(TM) SE Runtime Environment (build 1.8.0_161-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.161-b12, mixed mode)
root@servidor:~#
```

Fonte: <https://localhost:8080>

Por último, certifique-se de que o servidor está atualizado, digitando os comandos `apt-get update` e `apt-get upgrade`.

4.4.2 Instalando Apache Nifi

Com o ISPCONIG devidamente configurado, o Apache Nifi já poderá ser instalado, neste documento será detalhado a instalação da versão 1.3.0..

Com o servidor pronto para receber as instalações, no terminal de controle Linux acesse a pasta de arquivos temporários (`cd /usr/tmp`) para baixar o arquivo de instalação do Apache NiFi, digitando o comando: `wget --no-check-certificate http://www-us.apache.org/dist/nifi/1.3.0/nifi-1.3.0-bin.tar.gz`.

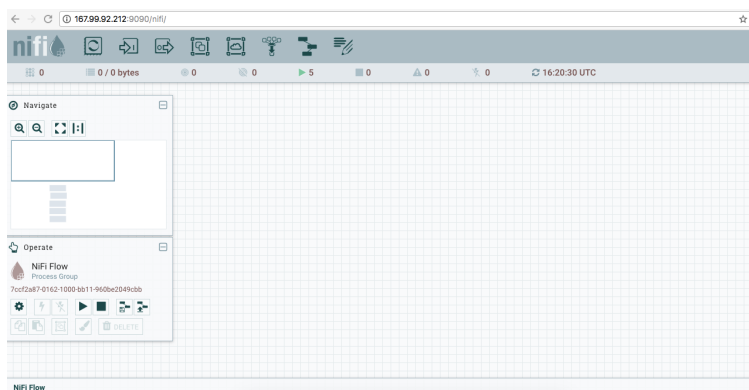
Assim que o download terminar, descompacte o arquivo `nifi-1.5.0-bin.tar.gz` digitando o comando `tar -xzf nifi-1.5.0-bin.tar.gz`, e mova os arquivos para seu devido local de destino, através do comando `mv nifi-1.3.0 /usr/local/nifi`.

Originalmente o Apache NiFi utiliza a porta 8080, porém neste caso esta porta já está sendo utilizada pelo o IPSCONFIG, sendo assim a porta 8080 deve ser alterada para 9090. Para que a alteração seja realizada o arquivo de configuração do NiFi deve ser editado em seu campo `nifi.web.http.port`. Para realizar tal configuração, basta digitar comando `nano /usr/local/nifi/conf/nifi.properties`.

Tendo em vista que a porta 9090 já foi aberta no Firewall, utilizando o ISPCONFIG, o serviço Nifi está pronto para ser iniciado, para isso acesse a pasta `usr/local/nifi/bin` e insira o seguinte comando no terminal de controle Linux: `./nifi.sh start`.

Após o término da instalação o Apache Nifi poderá ser acessado através do link <https://localhost:9090>.

Imagem 15: Apache NiFi instalado



Fonte: <https://localhost:9090>

4.4.3 Instalando Solr

Com o ISPCONIG e Apache Nifi devidamente configurados e instalados, chegou a hora de instalar o Solr, neste documento será detalhado a instalação da versão 6.6.0.

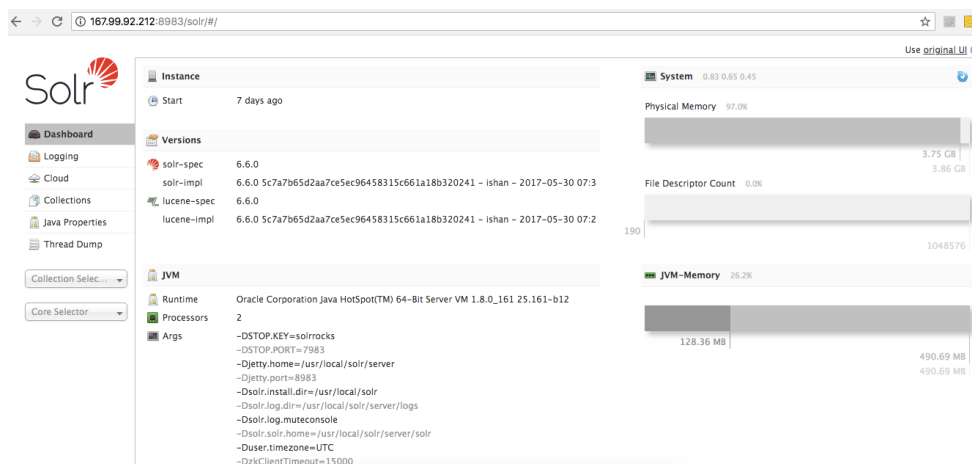
Como qualquer instalação em Linux, primeiramente deve-se abrir o terminal de controle e inserir o comando `wget http://apache.mirror1.spango.com/lucene/solr/6.6.0/solr-6.6.0.tgz`, para se baixar o arquivo de instalação, neste caso o `solr-6.6.0.tgz`.

Uma vez de posse do arquivo `solr-6.6.0.tgz`, basta descompacta-lo e move-lo para seu local de destino, inserindo os comandos `tar -xzf solr-6.6.0.tgz` e `mv solr-6.6.0 /usr/local/solr`.

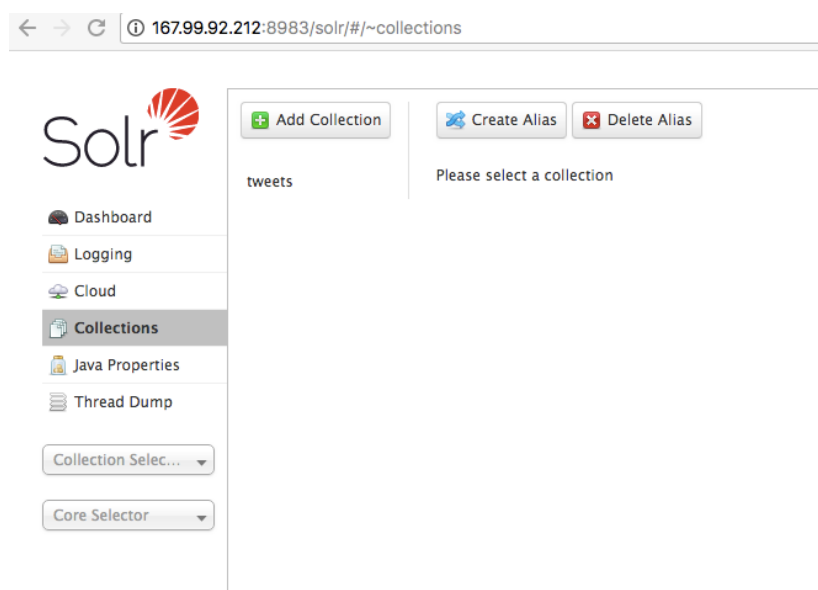
Para que o Solr receba os dados coletados do Apache Nifi, deve-se iniciar o serviço Solr em modo cloud e criar uma biblioteca que armazenará os dados coletados através do Apache Nifi.

Iniciar o serviço Solr em modo cloud é uma tarefa simples, basta acessar a pasta `/bin` do diretório Solr e inserir o comando `solr start -c`. Uma vez iniciado o Solr poderá ser acessado através da URL `http://localhost:8983`.

Com o serviço Solr no ar a biblioteca poderá ser criada, neste caso a chamamos de Tweets, através do comando `./bin/solr create_collection -c tweets -d data_driven_schema_configs -shards 1 -replicationFactor 1`.

Imagem 16: Solr instalado

Fonte: <https://localhost:8983>

Imagem 17: Verificando bibliotecas Solr

Fonte: <https://localhost:8983>

4.4.4 Configurando workflow do Apache Nifi para coletar dados do Twitter

Com o servidor atualizado e devidamente configurado, deve ser criado o workflow responsável pela coleta de dados públicos do Twitter através do Apache Nifi.

Para melhor atender todo tipo de demanda por parte de seus usuários, o Apache Nifi disponibiliza inúmeros tipos de processos que juntos determinarão as funcionalidades do seu workflow, neste caso em específico utilizaremos os seguintes processos: GetTwitter, EvaluateJsonPath, RouteOnAttribute, Merge Content, e PutSolrContentStream.

O primeiro passo para a criação do workflow é inserir os processos a serem configurados, para isso arraste o ícone de processos para o centro da tela.

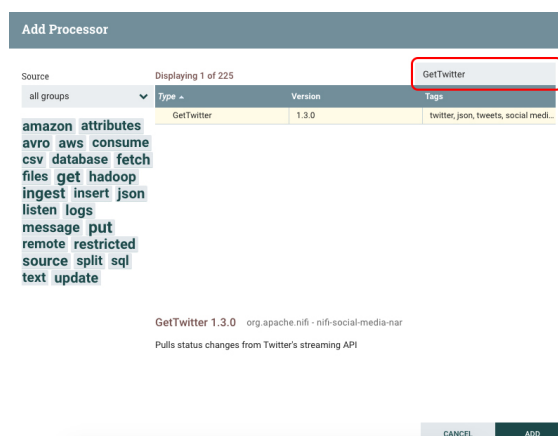
Imagem 18: Inserindo processo em fluxo gerado através do Apache NiFi



Fonte: <https://localhost:9090>

Após arrastar o ícone para o centro da tela, filtre pelo nome dos processos e clique em adicionar.

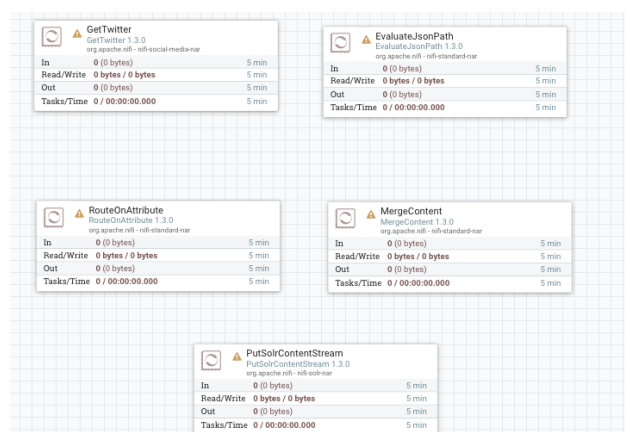
Imagem 19: Painel para busca de processos disponíveis no Apache NiFi



Fonte: <https://localhost:9090>

Após inserir os cinco processos no centro da tela, chegou a hora de configura-los.

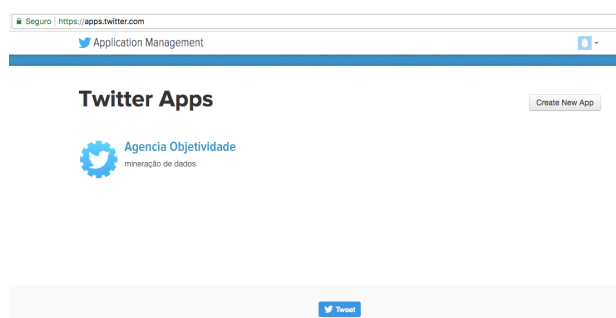
Imagem 20: Processos que compõem o fluxo para coleta de dados



Fonte: <https://localhost:9090>

O primeiro processo a ser configurado é o GetTwitter, responsável pela conexão com a plataforma social do Twitter. Para que este processo seja configurado com sucesso, primeiramente acesse o website <https://apps.twitter.com/> e crie uma aplicação que disponibilizará as seguintes chaves: Consumer Key, Consumer Secret, Access Token e Access Token Secret.

Imagem 21: Adquirindo as chaves Consumer Key, Consumer Secret, Access Token e Access Token Secret



Fonte: <https://apps.twitter.com/>

Em posse das chaves, com o botão direito do mouse clique no processo GetTwitter e selecione a opção configurar. Na aba propriedades insira os dados de acesso fornecidos pelo Twitter.

Imagem 22: Configurando o processo GetTwitter

Property	Value
Twitter Endpoint	Sample Endpoint
Consumer Key	No value set
Consumer Secret	No value set
Access Token	No value set
Access Token Secret	No value set
Languages	No value set
Terms to Filter On	No value set
IDs to Follow	No value set
Locations to Filter On	No value set

Fonte: <https://localhost:9090>

Uma vez inseridas as chaves disponibilizadas pelo Twitter, o Apache Nifi poderá se conectar com a plataforma social, portando para terminar as configurações do processo GetTwitter basta preencher os campos idiomas, optando pelo Português do Brasil (pt-BR), e

termos, ou seja, as publicações públicas que serão armazenadas deverão conter a palavra estipulada no campo termos.

Imagem 23: Configurando idioma e termos coletados pelo processo GetTwitter

Property	Value
Twitter Endpoint	Sample Endpoint
Consumer Key	No value set
Consumer Secret	No value set
Access Token	No value set
Access Token Secret	No value set
Languages	No value set
Terms to Filter On	No value set
IDs to Follow	No value set
Locations to Filter On	No value set

Fonte: <https://localhost:9090>

Após configurar o processo GetTwitter, chegou a hora de configurar o processo EvaluateJsonPath, responsável por conferir os termos e idioma previamente configurados no processo anterior, antes de armazenar as publicações que preenchem tais requisitos.

Nas propriedades do processo EvaluateJsonPath, deve ser inseridos dois novos atributos, que chamamos de twitter.lang, contendo a expressão \$.lang, e twitter.text, com o valor igual a \$.text.

Imagem 24: Inserindo atributos no processo EvaluateJsonPath

Property	Value
Destination	flowfile-content
Return Type	auto-detect
Path Not Found Behavior	ignore
Null Value Representation	empty string

Fonte: <https://localhost:9090>

Imagem 25: Configurando o processo EvaluateJsonPath

Processor Details

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
Destination	flowfile-attribute
Return Type	auto-detect
Path Not Found Behavior	ignore
Null Value Representation	empty string
twitter.lang	\$.lang
twitter.text	\$.text

OK

Fonte: <https://localhost:9090>

Para garantir que o Apache Nifi não armazene campos nulos, e colete publicações públicas sempre dentro do idioma previamente selecionado, o processo RouteOnAttribute realizará a última verificação antes do armazenamento dos dados. Para configurar, basta acessar as propriedades do processo RouteOnAttribute e adicionar uma nova propriedade, que neste caso chamamos de tweet, com a seguinte expressão: `${twitter.text:isEmpty():not():and (${twitter.lang:equals("en")})}`

Imagem 26: Configurando o Processo RouteOnAttribute

Processor Details

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
Routing Strategy	Route to Property name
tweet	\${twitter.text:isEmpty():not():and (\${twitter.lang:equals("en")})}

OK

Fonte: <https://localhost:9090>

Uma vez configurados os processos Gettwitter, EvaluateJsonPath, RouteOnAttribute, chegou a hora de configurar o processo Merge Content, responsável por ir até o Twitter e coletar dados de acordo com os parâmetros previamente configurados nos demais processos.

Configurar o processo Merge Content é uma tarefa simples, pois neste caso apenas alteramos o campo Delimiter Strategy para texto, de forma que o Apache Nifi busque pelos termos informados no processo Gettwitter. Outro campo que pode ser alterado caso queira buscar por mais de um termo é o demarcador, que informa o limitador dos termos que serão coletados.

Imagem 27: Configurando o processo Merge Content

Property	Value
Attribute Strategy	Keep only Common Attributes
Correlation Attribute Name	No value set
Minimum Number of Entries	100
Maximum Number of Entries	500
Minimum Group Size	0 B
Maximum Group Size	No value set
Max Bin Age	60 seconds
Maximum number of Bins	100
Delimiter Strategy	Text
Header	
Footer	
Demarcator	
Compression Level	1
Keep Path	false

Fonte: <https://localhost:9090>

Por último deve-se configurar o processo PutSolrContentStream, que enviará os dados coletados através do Apache Nifi para a biblioteca Solr, previamente criada, chamada Tweets.

Para realizar esta tarefa basta acessar as configurações do processo PutSolrContentStream e alterar o campo Solr Location, informando o endereço de acesso do Solr, e collection, informando o nome dde sua biblioteca Solr.

Imagem 28: Configurando o processo PutSolrContentStream

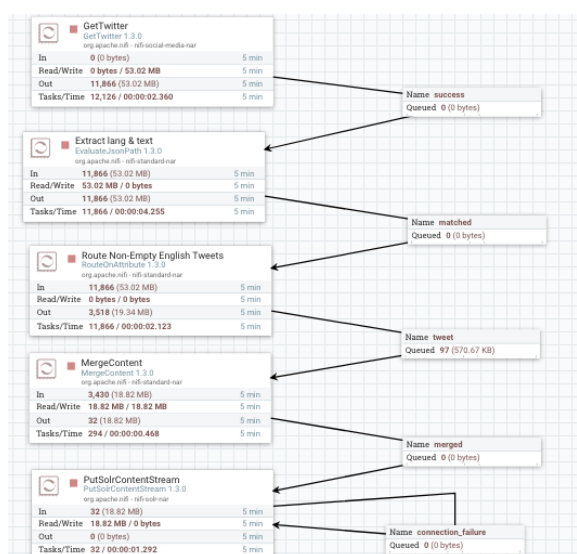
Property	Value
Solr Type	Cloud
Solr Location	localhost:9983
Collection	tweets
Content Stream Path	/update/json/docs
Content-Type	application/json
Commit Within	1000
JAAS Client App Name	No value set
Username	No value set
Password	No value set
SSL Context Service	No value set
Solr Socket Timeout	10 seconds
Solr Connection Timeout	10 seconds
Solr Maximum Connections	10
Solr Maximum Connections Per Host	5

Fonte: <https://localhost:9090>

Após configurar todos os cinco processos responsáveis pela coleta de dados, realize as seguintes conexões:

- 1- GetTwitter e EvaluateJsonPath;
- 2- EvaluateJsonPath e RouteOnAttribute;
- 3- RouteOnAttribute e Merge Content;
- 4- Merge Content e PutSolrContentStream;

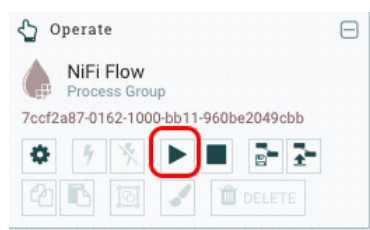
Imagem 29: Conectando processos do fluxo criado para coletar dados do Twitter



Fonte: <https://localhost:9090>

Para que o workflow comece a coletar dados, basta clicar no botão Play.

Imagem 30: Iniciando processos



Fonte: <https://localhost:9090>

Observe o painel superior para verificar se os cinco processos que compõem o workflow estão operando sem erros.

Imagem 31: Verificando erros no Apache NiFi

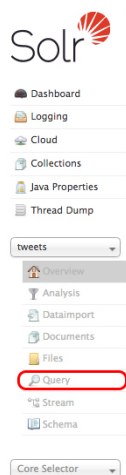


Fonte: <https://localhost:9090>

4.4.6 Verificando dados coletados através do Apache Nifi com o Solr

Após configurar o Apache Nifi para coletar termos de seu interesse, acesse o Solr, para verificar se os dados foram corretamente inseridos na biblioteca Tweets, através do endereço <http://localhost:8983>. Selecione a biblioteca Tweets e clique na opção Query.

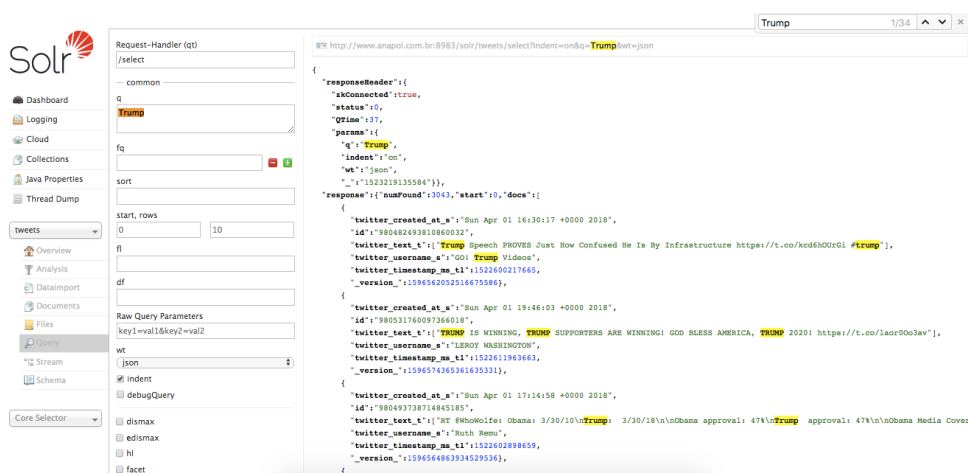
Imagem 32: Acessando o painel de pesquisa da biblioteca Tweets



Fonte: <https://localhost:8983>

Uma vez selecionada a opção Query, utilize a opção q para inserir o termo desejado para verificar os resultados.

Imagem 33: Pesquisando por termos previamente configurados no Apache NiFi



Fonte: <https://localhost:8983>

5 Conclusão

Conforme os objetivos estabelecidos, conclui-se que o trabalho possibilitou o desenvolvimento de um ambiente Linux capaz de recuperar dados de mídias sociais com responsabilidade, tarefas que deverão sempre respeitar as configurações de privacidade previamente configuradas por proprietários de perfis em mídias sociais.

A partir da instalação de programas Open Source, como o Apache NiFi e Solr, é possível se extrair dados de plataformas sociais para a realização de pesquisas que envolvem a análise do comportamento humano, e suas tendências. Minerar os dados utilizando ferramentas como Rapidminer, por exemplo, é uma prática extremamente comum no processo de desenvolvimento de estratégias assertivas.

Embora existam outros softwares disponíveis no mercado que também ofereçam os mesmos serviços apresentados neste documento, a utilização de ferramentas Open Source tornam o projeto de coleta e mineração de dados viável, devido ao seu baixo custo de implantação e manutenção, quando comparados com soluções tais como Power BI, Tableau, SAP e Microstrategy.

De acordo com pesquisas recentes, observou-se a existências de várias propostas para projetos de coleta de dados, planejamento tecnológico que muitas vezes não se preocupam com a integridade dados coletados ou com as normativas previstas pela Lei do Marco Civil da Internet, no que se refere a invasão de privacidade.

Com a implantação da solução proposta neste trabalho, será possível configurar um ambiente capaz de recuperar e armazenar termos de sua preferência em plataformas sociais como o Twitter, dados que poderão ser analisados posteriormente com mais profundidade com o auxílio de ferramentas como Rapidminer.

6 Referências Bibliográficas

Palácio do Planalto. **Neutralidade, liberdade de expressão e privacidade: conheça os pilares do Marco Civil**, Disponível em:

<<http://www2.planalto.gov.br/noticias/2015/04/neutralidade-liberdade-de-expressao-e-privacidade-conheca-os-pilares-do-marco-civil>> Acesso em: 15 mar. 2018.

Patrícia Gnipper. **A evolução das redes sociais e seu impacto na sociedade – parte 3**.

Disponível em: <<https://canaltech.com.br/redes-sociais/a-evolucao-das-redes-sociais-e-seu-impacto-na-sociedade-parte-3-109324/>> Acesso em: 15 mar. 2018.

Marketingdeconteúdo. **Política de Privacidade: O que é e como montar uma**. Disponível em: < <https://marketingdeconteudo.com/politica-de-privacidade/>> Acesso em: 18 mar. 2018.

Luís Lima. **Pior recessão da história complica retomada da economia brasileira**.

Disponível em:: <<https://epoca.globo.com/economia/noticia/2017/03/pior-recessao-da-historia-complica-retomada-da-economia-brasileira.html>> Acesso em: 18 mar. 2018.

Facebook. **Quando public algo, como posso selecionar quem vê esses conteúdos**.

Disponível em: <<https://www.facebook.com/help/120939471321735>> Acesso em: 20 mar. 2018.

Facebook. **Que funcionalidades estão disponíveis com base na categoria da minha página?**. Disponível em:

<https://www.facebook.com/help/918592541485077?helpref=faq_content> Acesso em: 20 mar. 2018.

Renato Santino. **Testes de Facebook são uma invasão de privacidade enorme**. Disponível em:

<<https://olhardigital.com.br/noticia/testes-de-facebook-sao-uma-invasao-de-privacidade-enorme/53252>> Acesso em 23 mar. 2018.

Portal Terra. **Como os testes de Facebook usam seus dados pessoais e como empresas ganham dinheiro com isso.** Disponível em: <https://www.terra.com.br/noticias/tecnologia/como-os-testes-de-facebook-usam-seus-dados-pessoais-e-como-empresas-ganham-dinheiro-com-isso,a55234888a73d9e0984db908dde5d3371xsr6c9s.html> Acesso em 24 mar. 2018.

Senado da República. **O Facebook e o direito à privacidade.** Disponível em: https://www12.senado.leg.br/ril/edicoes/51/201/ril_v51_n201_p17.pdf Acesso em 26 mar.2018.

Apache NiFi. **Apache Nifi Users Guide.** Disponível em: < <https://nifi.apache.org/docs/nifi-docs/html/user-guide.html> > Acesso em 28 mar. 2018.

Apache Solr. **Apache Solr Reference Guide.** Disponível em: https://lucene.apache.org/solr/guide/6_6/ Acesso em 28 mar.2018.

Anderson Castro Soares de Oliveira. **A utilização de redes sociais da internet para obtenção de dados.** Disponível em: <http://www.ufmt.br/dest/arquivos/2426356bc2ad9e7864f5f17f06d71bfa.pdf> Acesso em 14 abr. 2018.

Folha de São Paulo. **Vazamento de dados do Facebook atinge 443.117 usuários brasileiros.** Disponível em: <https://www1.folha.uol.com.br/mundo/2018/04/vazamento-de-dados-do-facebook-atinge-443117-usuarios-brasileiros.shtml> Acesso em 14 abr. 2018.

Emerson Alecrim. **A controvérsia dos 50 milhões de perfis do Facebook manipulados pela Cambridge Analytica.** Disponível em: <https://tecnoblog.net/236612/facebook-cambridge-analytica-dados/> > Acesso em 14 abr. 2018.