Grupo de Estudo de Aspectos Empresariais e de Gestão Corporativa e da Inovação e da Educação e de Regulação do Setor Elétrico-GEC

Uma plataforma de ciência de dados para o setor elétrico brasileiro

ANDRÉ EMILIO TOSCANO (1); MARCOS DE ALMEIDA LEONE FILHO (1); MAKOTO KADOWAKI (1); RAFAEL GIORDANO VIEIRA (1); JOÃO BORSOI SOARES (1); VENIDERA PESQUISA E DESENVOLVIMENTO (1);

RESUMO

O Sistema Interligado Nacional (SIN) é o sistema que atende o mercado consumidor de energia elétrica do Brasil. Um sistema peculiar pela abundante oferta de energia renovável hidrelétrica, com múltiplos empreendimentos de geração com as mais variadas fontes primárias. Sua malha de transmissão, de porte continental, integra os geradores aos centros consumidores, garantindo o suprimento de energia elétrica do mercado consumidor nacional e de países vizinhos.

Um sistema de grande porte e complexo como o SIN é organizado e gerido por diferentes órgãos, como o Ministério de Minas e Energia (MME), a Agência Nacional de Energia Elétrica (ANEEL), a Câmara de Comercialização de Energia Elétrica (CCEE), o Operador Nacional do Sistema (ONS), a Empresa de Pesquisa Energética (EPE), a Agência Nacional de Águas (ANA) e outros órgãos que gerenciam aspectos regionais do sistema. Tais organizações mantêm sua transparência divulgando informações do sistema em suas fases de planejamento, operação e histórico. Diversos conjuntos de dados sobre características, estados e funcionamento do SIN são publicados de forma distribuída, mas a distribuição não é padronizada entre os órgãos, o que dificulta a sistematização do uso desses dados.

Neste contexto, este informe técnico apresenta uma plataforma de ciência de dados desenvolvida como solução para aquisição, associação, armazenamento, indexação, processamento e exportação/visualização dos dados do setor elétrico brasileiro considerando suas fontes de dados diversificadas. A metodologia aplicada favorece a aquisição de dados em massa (streaming) através de coletores, a análise dos dados coletados para submissão de armazenamento por um misto de técnicas de modelagem de dados como a orientada a objetos, a relacional e a em grafos, com indexações espacial (GIS), temporal (séries temporais) e textual. A arquitetura orientada a serviços (SOA) e micro-serviços com comunicação RESTful foram aplicadas e requisitos de *big data* são atendidos. O resultado final foi a criação de um sistema que pode ser visto como uma base de dados do setor elétrico brasileiro. A orquestração de diferentes técnicas fornece o suporte para a coleta em massa de dados, o armazenamento otimizado e a leitura e exportação com alto desempenho, não sofrendo impactos sensíveis com o crescimento elevado do conjunto de dados armazenados.

O caso básico de uso da plataforma de ciência de dados é a coleta de dados em massa de fontes públicas ou privadas, realizada de forma automatizada e periódica. Algoritmos de robotização fazem a aquisição dos dados na web sem qualquer intervenção dos usuários. Como exemplo, os dados publicados pelo MME (relatórios), ANEEL (normativas, revisões tarifárias), CCEE (InfoPLD, contratos, decks de preço, PLD), ONS (EAR, ENA, CMO, cart Anarede), EPE (decks de leilão, estudos de energia firme e garantia física) e ANA (operação dos reservatórios), bem como notícias do setor e dados meteorológicos são automaticamente coletados, tratados e armazenados com um modelo de dados híbrido que acomoda as informações associadas e geolocalizadas. Índices de alto

(*) RUA MONTE LÍBANO, 536, JARDIM CHAPADÃO, CEP 13070086, São Paulo, – Brasil
andre@venidera.com

desempenho são aplicados, permitindo buscas diversas e oferecendo rápida rastreabilidade dos dados armazenados. Os dados lidos da plataforma podem ser utilizados para visualização ou para exportação para sistemas corporativos ou científicos. A exportação para formatos transacionais como o XML e JSON está disponível e provê facilitada integração com ferramentas computacionais diversas.

A plataforma também provê um serviço de computação para o suporte a aplicações diversas de tratamento e uso dos dados, tais como para execução de simulações, análises estatísticas e de anomalias, geração de previsores, dentre outras. Logo, os atuantes nas áreas de geração, transmissão e distribuição podem se valer dos dados de maneira ágil e amigável para os usos que desejarem. A escala computacional provida pela automatização dos processos de aquisição e processamento eliminam os limites que a operação manual imporia, bem como adiciona elevada consistência nas atividades de coleta e uso dos dados.

Diversos subprodutos da plataforma de ciência de dados foram desenvolvidos, como estudos de dados com todos os decks Newave já publicados, análises de contratação de energia, relatórios informativos sobre o setor, e a análise de demanda e oferta do sistema. Os resultados apontam a entrega de valor estratégico aos usuários de órgãos públicos ou privados, provendo agilidade de aquisição e utilização dos dados do setor elétrico brasileiro para as mais diversas finalidades.

PALAVRAS-CHAVE

Ciência-Dados, Data-Science, Big-Data, Setor-Elétrico-Brasileiro

1.0 - INTRODUÇÃO

O Setor Elétrico Brasileiro (SEB) é instituído por quatro níveis, como ilustrado na FIGURA 1. O primeiro nível é definido pelas políticas e planejamento, constituído pelo Congresso Nacional, Presidência da República, Ministério de Minas e Energia (MME) e a Empresa de Pesquisa Energética (EPE). Este nível organiza o setor elétrico em foco nacional no contexto das políticas públicas e planejamento que visem assegurar e viabilizar o planejamento do setor elétrico considerando os interesses do país. O segundo nível é o de regulação e fiscalização, desempenhado pela Agência Nacional de Energia Elétrica (ANEEL). O terceiro nível faz a implementação do SEB do ponto de vista de planejamento, operação e comercialização da energia elétrica do *grid* nacional ou Sistema Interligado Nacional (SIN). Finalmente, o quarto nível é definido pelos consumidores contratando energia elétrica para execução das mais variadas atividades exercidas nos diferentes contextos tais como os domésticos, comerciais e industriais.

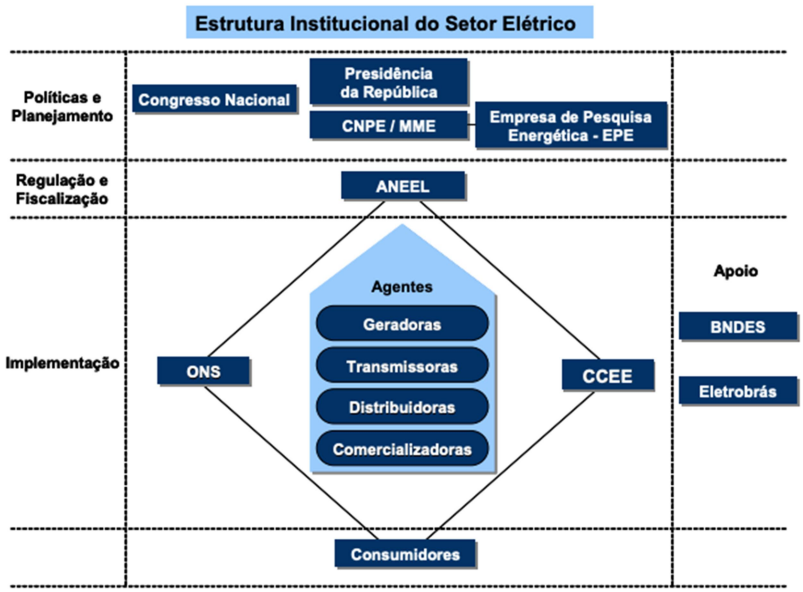

Figura 1 – Estrutura Institucional do Setor Elétrico. Fonte: ANEEL/SPG/2013

Cada um dos níveis que organizam o SEB são fornecedores de dados sobre o SIN. No entanto, a geração desses dados tem se tornado cada vez maior com os avanços tecnológicos, dado que atualmente qualquer dispositivo pode ser conectado na internet e, portanto, gerar dados em sua operação. Por sua vez, os fabricantes utilizam-se, quando autorizados, destas informações para tornar a utilização pelos usuários uma experiência

personalizada, de forma que os conteúdos e recursos que tenham correlação com a utilização do usuário possam ser sugeridos de maneira a tornar as suas experiências de uso mais atraentes e prazerosas.

Os adventos modernos que possibilitam o registro de utilização de recursos por usuários definem métricas que podem ser utilizadas de diversas formas, muitas ainda por serem descobertas. Tais descobertas são realizadas através da análise profunda dos dados gerados em cada atividade registrada. Um exemplo disso é a análise do perfil de consumo de energia elétrica de uma residência, que pode permitir identificar todos os eletrodomésticos em uso basicamente pela análise isolada do perfil de consumo de um largo conjunto de eletrométicos.

Um desafio considerável sobre o montante de dados gerado na atualidade é o de construir sistemas capazes de gerenciar quantidades gigantescas de informações de maneira rápida e consistente. Diferentes tecnologias vêm sendo desenvolvidas para esta atividade, tendo estas sido consideradas no conjunto definido de tecnologias aplicadas ao *Big Data (3)*, um termo que tornou-se extremamente notável nos mais variados contextos como um indicativo de tecnologia inovadora e atual. O termo *Big Data* define o campo de análise, extração sistemática de informações e a manipulação de conjuntos de dados enormes e complexos de maneira a não serem tratáveis por técnicas e aplicações computacionais tradicionais.

Um vasto leque de técnicas para análise de dados está disponível. Tais técnicas que vão desde a organização dos dados em conjuntos estruturados, a visão de dados em dimensões de análise, a mineração de dados, entre outras técnicas. A formalização do processo de experimento com os dados seguindo uma abordagem científica é proposta pela Ciência de dados (*Data Science),* uma extensão moderna do que a estatística propõe para análise de dados. De fato, a ciência de dados tem todas as características que cunham o termo “ciência” quando aplicado a um objeto de estudo. Trata-se de uma prática multi-disciplinar de conhecimentos somados nas áreas de tecnologia da informação, matemática e estatística, análise de dados, que considera um campo de aplicação e estudos e que segue conceitos clássicos de definição de objeto de estudo, estabelecimento de hipóteses e a definição de experimentos para validação ou não das hipóteses definidas para análise.

Como objetivos finais, tanto da Ciência de Dados como das diferentes técnicas de análise de dados, estão a identificação de dinâmicas nos dados, de comportamento que possam emergir e estejam expressos nos dados e finalmente a entrega de valor para uso em atividades de negócio diversas.

Este trabalho apresenta uma plataforma de Ciência de Dados, e a aplicação de técnicas de coleta e análise de dados, e recursos computacionais foram desenvolvidos tendo como caso base o SEB.

A utilização de dados do SEB propõe alguns desafios. A diversificação de fontes é uma delas, pois determinados dados tais como os contratos de energia são publicados pela Câmara de Comercialização de Energia Elétrica (CCEE). Documentos de operação e outros são publicados como procedimentos de rede pelo Operador Nacional do Sistema (ONS). O deck NEWAVE e DECOMP de cálculo para o Preço de Liquidação das Diferenças definido no Planejamento Mensal de Operação é disponibilizado pela CCEE e pelo ONS. O ONS provê ainda o histórico da operação, bem como outros documentos que são de acesso privado aos agentes do setor. A ANEEL publica documentos diversos de interesse dos agentes tais como a revisão tarifária para os empreendimentos de geração. E dadas as características do SIN de ser um sistema hidrotérmico com predominância de geração hidrelétrica, estudos meteorológicos se fazem necessários para suporte à decisões de comercialização e operação, e portanto o uso de fornecedores de dados climáticos, aos quais podemos citar o INPE (Instituto Nacional de Pesquisas Espaciais) e a agência americana NOAA (National Oceanic and Atmospheric Administration). Estes são alguns dos fornecedores de dados de interesse para pesquisas no SEB, outras fontes ainda podem ser incluídas. A utilização de uma plataforma de ciência de dados deverá fornecer a coleta periódica de dados nos mais diversos formatos, a análise destes e preparação para armazenamento, o armazenamento e a indexação para rápida leitura, e posterior acesso rápido as informações para submissão como dados de entradas para os algoritmos de análise que se desejar aplicar. Finalmente, a publicação e visualização facilitada dos dados coletados e produzidos é o artifício final do ciclo de ciência de dados.

2.0 - MIRAN: UMA PLATAFORMA DE CIÊNCIA DE DADOS

A atual e crescente disponibilidade de dados tem gerado cada vez mais demandas para a análise e uso das informações registradas, demandas estas que podem ser definidas desde a disponibilização e aquisição de informações até o armazenamento destas informações em bancos de dados e análise efetiva das informações obtidas. Objetivos diversos podem ser definidos nas análises de grandes volumes de dados e, ao atendimento destes objetivos, diferentes abordagens podem ser aplicadas, tais como *Data Warehouse*, OLAP (*Online Analytical Processing*), *Big Data*, dentre outras. De qualquer forma, a finalidade comum a todas as abordagens consiste na extração de conhecimento a partir da abstração dos dados e das informações (4).

Nesse contexto, essa seção tem como objetivo detalhar os requisitos, especificações e implementações de uma plataforma de ciência de dados voltada ao setor elétrico brasileiro, denominada *plataforma Miran*. O termo "plataforma" é aqui designado como um conjunto de módulos (ou serviços) voltados tanto às análises que se desejam realizar, bem como de artifícios e componentes de *software* que possibilitem o armazenamento e recuperação dos dados necessários de forma eficiente e flexível.

2.1 Infraestrutura computacional baseada em serviços

Nesta seção, será detalhado como a infraestrutura de dados da plataforma Miran foi concebida, especificada e implementada. Inicialmente, como forma de criar um ambiente de armazenamento de dados, tornou-se necessário que os seguintes requisitos fossem atendidos:

1. **Suporte e capacidade de entrada massiva de dados**: um conceito denominado *stream* é o que há de mais moderno em suportar um grande volume de conexões de dados simultaneamente. Este requisito só pode ser atendido considerando componentes eficientes que façam bom uso dos recursos de comunicação de rede e que forneçam suporte ao tratamento dessas massas de dados.
2. **Suporte a computação elástica em nuvem:** o termo *elastic compute cloud* define uma rede distribuída sob demanda que fornece a flexibilidade para aumentar ou diminuir os recursos computacionais usando de esforços mínimos de administração e seguindo políticas definidas de atendimento as requisições. Nesse contexto, tem-se servidores provendo serviços e clientes acessando estes serviços, porém sob o conceito de que um serviço é provido por um conjunto de agentes que buscará atender a requisição da forma mais eficiente possível com os recursos alocados.
3. **Armazenamento híbrido de dados estruturados e não estruturados:** uma característica inerente a maioria dos dados obtidos na internet é a forte despadronização apresentada por eles. Por esse motivo, tornou-se necessário possuir um esquema base de dados que acomodasse níveis de estruturação diferentes, levando à necessidade de um esquema denominado *schema-hybrid*, formado por meta campos de dados definidos para um objeto, sendo que estes campos podem acomodar qualquer esquema de dados desejado.

A infraestrutura computacional utilizada para escalar os serviços da plataforma Miran levou em consideração os requisitos apresentados anteriormente. O modelo de dados definido para a implementação suporta a acomodação de diferentes tipos de dados e a representação de suas relações através da teoria de grafos. Embora os modelos tradicionais ER sejam utilizados de forma eficaz para a comunicação de dados e definições de relacionamentos com o usuário, a sua estrutura no contexto de monitoramento de grandes quantidades de dados está longe do ideal. Isso ocorre porque, quanto mais dados houver na base, mais lentas as consultas se tornarão, inviabilizando, na prática, o correto funcionamento da aplicação (5). Assim, a utilização de modelos que consultem relações em menor tempo, como é o caso do modelo NoSQL, torna-se crítica para atender às necessidades de uso desta infraestrutura.

Como forma de criar um ambiente de armazenamento de dados eficiente para grandes massas de dados, foi necessária a composição de diferentes bases de dados NoSQL, cada uma com uma finalidade, além de componentes que orquestram entrada e saída de dados:

1. **Base de dados de documentos em grafo:** no paradigma NoSQL, as bases de dados de documentos são um tipo específico que buscam armazenar os dados em documentos definidos no formato JSON (JavaScript Object Notation). O papel na infraestrutura de dados para uma base de documentos é o de armazenar objetos de dados, e temos como objeto de dado a abstração de quaisquer características que consigam ser agrupadas em ente individual, portanto, sendo esta a unidade mínima de dados na plataforma. Dentre as bases de dados pesquisadas, o OrientDB foi a que se destacou, pois forneceu o conjunto completo das necessidades da plataforma. Trata-se de uma base de dados de grafos implementada sobre o paradigma de orientação a objetos, em que cada objeto pode ser constituído ou não de um modelo. Um outro aspecto é que o OrientDB fornece indexação Lucene em seu motor de indexação. Os índices Lucene fornecem alto desempenho na busca de dados em formato textual.
2. **Base de dados de séries temporais:** séries temporais constituem conjuntos de dados associados a observações cronológicas. Na atualidade, o modelo NoSQL de bases de dados em colunas (ou linhas) fornecem as melhores opções para a gestão de séries temporais. A gerência de alto desempenho de séries temporais também é necessária, e este requisito foi atendido com o uso do OpenTSDB, uma base de dados NoSQL de alto desempenho para gerência de séries temporais. Essa base fornece poder suficiente de entrada de dados escalável, armazenamento eficiente e leitura extremamente rápida. Ele é construído sobre a base de dados HBase e, portanto, pode se aproveitar de todos os benefícios e funcionalidades que esta provê dentro do ecossistema Hadoop.

Os componentes que formam a infraestrutura de dados são disponibilizados através de computadores virtuais definidos em *containers* do *kernel* Linux. Este recurso de virtualização permite implementar sistemas compatíveis

com a maioria dos fornecedores de IAAS e PAAS do mercado, tais como o Amazon, IBM e Microsoft. Os *containers* podem ter recursos limitados, podem existir com interfaces de redes privadas e de contexto fechado, e podem assumir qualquer configuração definida como de nuvem pública, privada ou híbrida, sem perda de funcionalidade ou abstração dos serviços providos pelas companhias de computação em nuvem.

Por sua vez, a gerência de computação elástica foi implementada com o uso de um servidor *web* e *proxy* HTTP/TCP Nginx. O Nginx consiste de um servidor *web* completo para publicação de aplicações, mas também pode ser utilizado como *proxy* para balanceamento de carga. Assim, uma instância de Nginx foi criada para assumir a publicação de cada serviço, e esta instância faz o controle de quais serviços fornecer para o atendimento de cada requisição. O Nginx é quem realiza a distribuição de requisições nas instâncias de serviços, e ele é quem possibilita o crescimento ou diminuição da capacidade de atendimento a requisições de maneira transparente aos clientes que consomem os recursos da infraestrutura de dados. A organização de todos os componentes da infraestrutura de dados é ilustrada na Figura 2.

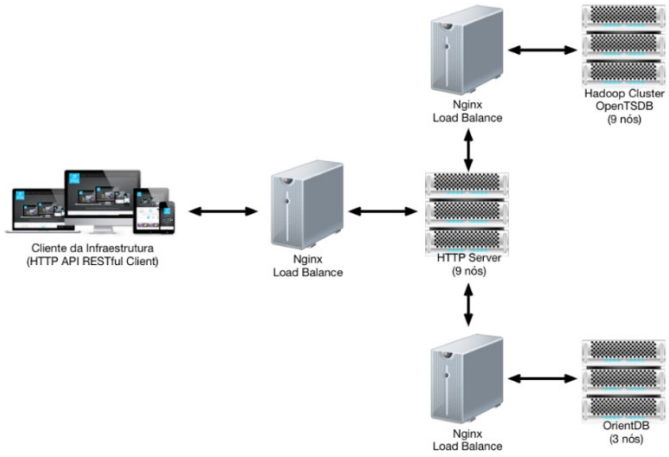

Figura 2 – Organização dos componentes na infraestrutura de dados

2.2 Módulo para aquisição massiva de dados

A coleta dos dados da plataforma Miran é realizada automaticamente e sistematicamente através de *web crawlers* programados individualmente para cada fonte de dado. Um *web crawler* (também chamado de indexador automático, *bot* ou *web spider*) consiste em um programa de computador que navega pela internet de forma metódica e automatizada (6) com o intuito de capturar algum tipo de dado ou informação. A plataforma Miran compreende web crawlers foram concebidos para navegar, visitar e extrair as informações relevantes de uma lista pré-definida de endereços, mantendo assim a base de dados regularmente atualizada.

Algumas das fontes de informações mais relevantes que foram adquiridas, armazenadas e processadas são: decks do NEWAVE, decks do DECOMP, planilhas com dados individuais do Infomercado, cotas do Proinfa, ofertas nos leilões pelos geradores, resultados dos leilões, histórico de PLD, dados de audiências públicas da Aneel, centro de documentação da Aneel (CEDOC), resultado dos processos tarifários de distribuição, processos das usinas, notas técnicas de regulamentação, resoluções regulamentadoras, informações de empreendimentos eletro-energéticos (bases BIG e SIGEL da ANEEL), informações georreferenciadas das áreas de concessão das distribuidoras de energia elétrica (base SIGEL da ANEEL, previsões de término das usinas (Dados dos PMOs), séries históricas de vazões, séries históricas de ENAS, dados dos procedimentos de rede, dados da rede elétrica (Anarede), notícias do setor eletro-energético e faturas de energia elétrica.

O processo de captura dos dados segue uma metodologia de funcionamento, a qual está apresentada de modo geral na Figura 3. Ela é composta por cinco etapas, que são enumeradas a seguir:

1. Inicialmente, o agendador de tarefas executa os web crawlers requeridos, sendo o horário de execução de cada web crawler definido de acordo com critérios como a frequência de atualização e o volume (em termos de tamanho) de cada tipo de dado.
2. Os web crawlers realizam o processo de busca das informações requeridas nas respectivas fontes de dados. A forma como essas informações são buscadas por cada web crawler pode variar de acordo com a complexidade da fonte de dados. Por exemplo, quando há a necessidade de autenticação ou medidas antiscrapping, torna-se necessário emular a navegação convencional utilizando ferramentas específicas.

3. As informações descobertas pelos web crawlers são agora capturadas e pré-processadas, como forma de extrair URLs com o endereço dos arquivos a serem baixados e também informações relevantes sobre cada documento.
4. Cada objeto gerado na etapa anterior é agora enviado para a base de dados. A comunicação entre o web crawler e a base não é realizada diretamente, mas sim com o auxílio de uma aplicação que encapsula os dados de entrada e modulariza as requisições, otimizando assim o fluxo de informações e aumentando a simplicidade na comunicação entre ambos os agentes
5. O objeto gerado a partir do conteúdo extraído de cada URL (ou seja, um arquivo do tipo PDF, CSV, dentre outros) é inserido na base de dados.

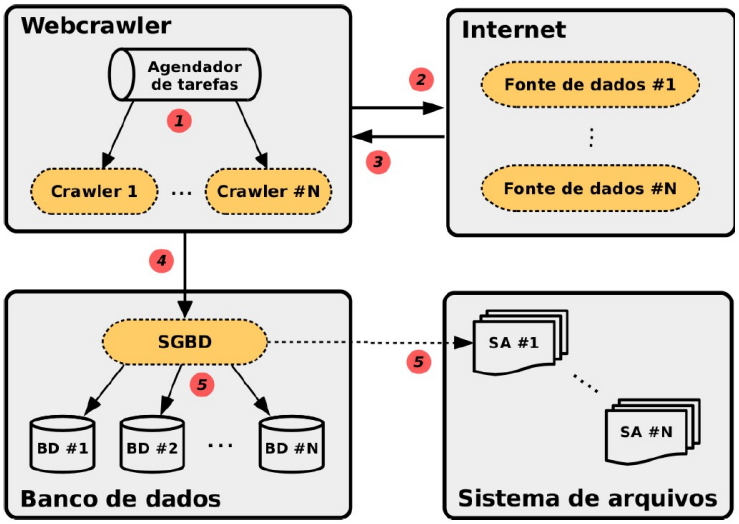

Figura 3 – Metodologia de funcionamento dos web crawlers

2.3 Módulo de filtragem e confiabilidade de dados

A Seção 2.2 teve como objetivo apresentar os processos relacionados à extração e armazenamento das informações de maior relevância na plataforma Miran. No entanto, como essas informações são obtidas – em sua grande maioria – por meio de fontes da internet, elas podem algumas vezes apresentar problemas em seus conteúdos, tais como erros de formatação, ausência de valores, dentre outros. Por esse motivo, torna-se desejável e necessário que seja desenvolvido um módulo onde esses eventuais problemas possam ser identificados, avaliados e corrigidos, de modo que as informações a serem apresentadas para o usuário final tenham maior confiabilidade. Com base nisso, esse módulo foi segmentado em duas funcionalidades: *processamento e interpretação dos dados coletados*, e *detecção e correção de anomalias*.

2.3.1 Processamento e interpretação dos dados coletados

Os algoritmos que realizam atividades de extração de dados, denominados *parsers*, são também conhecidos como analisadores de sintaxe. A análise sintática é o processo de análise de uma sequência de símbolos, seja em linguagem natural ou em linguagem de computador, conforme as regras de uma gramática formal. Esse termo tem significados ligeiramente distintos em diferentes ramos de linguística e ciência da computação. Dentro da ciência da computação, o termo é usado na análise de linguagens computacionais, referindo-se à análise sintática do código de entrada em suas partes componentes, a fim de facilitar a redação de compiladores e intérpretes. O termo também pode ser usado para descrever uma divisão ou separação.

Nesse contexto, os algoritmos de *parsing* são empregados com o intuito de extrair as informações desejadas através da análise sintática destes documentos (ex.: os dados procurados estão em um conjunto de células com conteúdo no formato “ponto-flutuante” da planilha “Contratos” cuja coluna contém o rótulo “Energia contratada em MWh”). Assim, como quase sempre estes documentos contém informação de forma não estruturada e os textos que ajudam a localizar a informação nestes documentos costumam sofrer pequenas alterações ao longo do tempo, os algoritmos de *parsing* precisam localizar a informação de forma inteligente e flexível (para garantir que pequenas alterações nestes documentos não causarão problemas na extração da informação) e, uma vez encontrada, organizá-la dentro da base de dados em forma de arquivos ou de séries temporais.

2.3.2 Detecção e correção de anomalias

A detecção de anomalias tem por objetivo identificar (ou filtrar) possíveis problemas nos dados de séries temporais que são capturados. De modo geral, o termo “anomalia” pode ser definido como qualquer comportamento não usual apresentado por um conjunto de dados (7). Essas anomalias podem ser induzidas nos dados por uma variedade de razões, como por exemplo, na alteração repentina do padrão estrutural dos dados, na presença de ruídos ou simplesmente devido à má formatação desses dados. Outras definições de “anomalia” podem também ser encontradas na literatura, como ruído, outlier, exceção e desvio (8).

Um aspecto chave na detecção de anomalias consiste na determinação da natureza dos dados que serão analisados. Os dados podem ser classificados em função de um conjunto de atributos, como por exemplo, a sua dimensão (univariável ou multivariável), a sua amostragem (discreta ou contínua), dentre outros. Outro aspecto importante que deve ser considerado na detecção de anomalias são as maneiras pelas quais elas podem ocorrer. De modo geral, a ocorrência de anomalias pode ser classificada de três formas: anomalia pontual, em que é retratada pela presença de uma amostra com valor muito destoante dos demais; anomalia contextual, que é definida pela presença de uma não continuidade em um padrão bem definido nos dados; e anomalia coletiva, que ocorre somente quando há padrões sazonais nos dados. Nesse contexto, ela é identificada como uma alteração na forma de um determinado padrão que se repete periodicamente nos dados.

2.4 Módulo de análise e evolução textual

Em virtude do crescimento contínuo do volume de dados disponíveis, técnicas de extração de conhecimento automatizadas têm se tornado cada vez mais necessárias para valorizar a gigantesca quantidade de dados existentes. Como as técnicas para mineração de dados foram desenvolvidas para dados estruturados, técnicas específicas para análise textual (ou mineração de texto) têm sido desenvolvidas para processar uma parte importante da informação disponível que pode ser encontrada na forma de dados não-estruturados.

As aplicações de análise textual fornecem uma nova dimensão das informações disponíveis para os clientes, uma vez que elas possibilitam monitorar a ocorrência e a evolução de termos específicos, como por exemplo dados estratégicos do mercado de energia elétrica, CNPJs, dentre outros. A plataforma Miran provê um serviço *web* de análise e monitoramento textual denominado *Miran Monitor*. O serviço foi desenvolvido para realizar o monitoramento textual para todos os objetos na base de dados da plataforma Miran, levando em consideração a delimitação de uma sintaxe SOLR para realizar as buscas. A Figura 4 mostra o painel principal do sistema, com a lista de monitoramentos textuais já configurados no sistema, na forma de “cartões”. Cada monitoramento tem em destaque, uma imagem gráfica das ocorrências do monitoramento atual, o título do monitoramento e o estado do monitoramento, a sua descrição, o período do monitoramento, e a data que foi criada. Ainda em destaque na Figura 4 constam os controles para ativar/desativar a visualização do monitoramento no aplicativo móvel, ativar/desativar o filtro de anomalias, e o menu de ações com os demais controles.

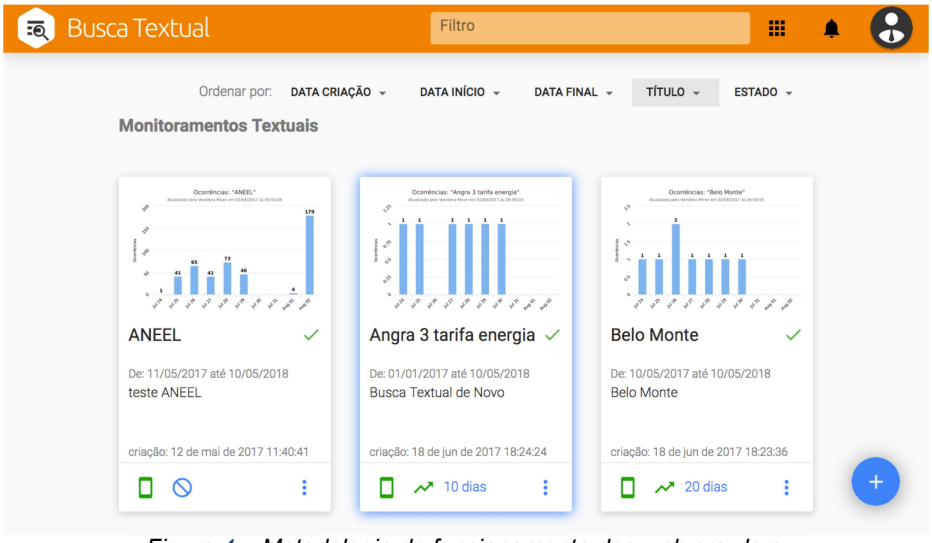

Figura 4 – Metodologia de funcionamento dos web crawlers

Os gráficos dos históricos de evolução temporal das ocorrências do monitoramento textual podem também ser visualizados no aplicativo móvel da plataforma Miran. Em cada um dos cartões de monitoramento textual do painel principal, existe um controle para que os resultados do monitoramento sejam ou não visualizados nos aplicativos móveis. Uma vez ativada a visualização no aplicativo móvel, será possível também ativar o filtro de anomalias para os termos monitorados, conforme especificado na Seção 2.3. Em outras palavras, torna-se possível realizar o monitoramento em tempo real de termos específicos na base de dados, de modo que, caso o

serviço detecte algum padrão irregular, o usuário seja avisado dessa ocorrência, conforme ilustrado na Figura 5.

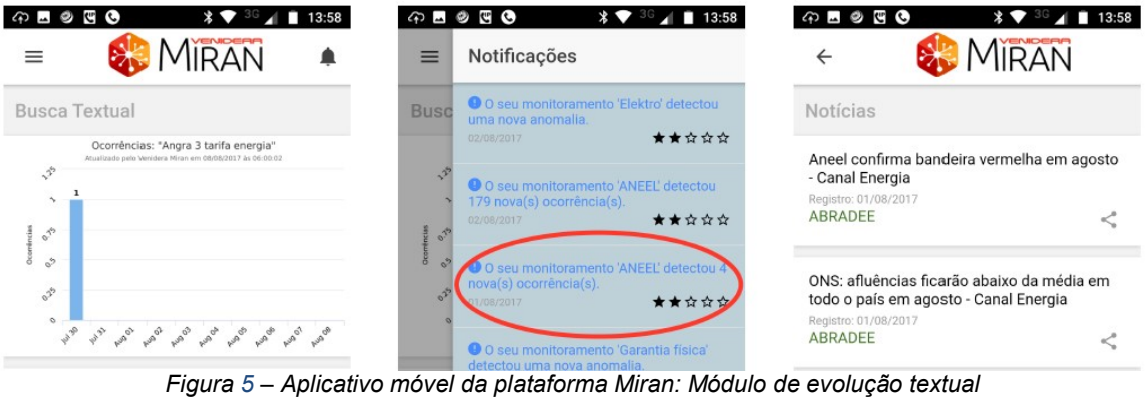

Figura 5 – Aplicativo móvel da plataforma Miran: Módulo de evolução textual

3.0 - CONCLUSÃO

Este artigo apresentou uma plataforma inovadora de ciência de dados que foi projetada como solução para a aquisição, armazenamento, indexação e análise de dados voltados ao setor elétrico brasileiro. A plataforma em questão, denominada Miran, é formada por um conjunto de serviços orquestrados de forma a fornecer o suporte para a coleta em massa de dados, o armazenamento otimizado e a leitura e exportação com alto desempenho, não sofrendo impactos sensíveis com o crescimento elevado do conjunto de dados armazenados. A plataforma também provê o suporte a aplicações diversas de tratamento e uso dos dados, como por exemplo, para a execução de análises estatísticas, detecção de anomalias e simulações em geral. Os resultados apontam a entrega de valor estratégico aos usuários de órgãos públicos ou privados, provendo agilidade de aquisição e utilização dos dados do setor elétrico brasileiro para as mais diversas finalidades.

4.0 - REFERÊNCIAS BIBLIOGRÁFICAS

(1) HAYASHI, Chikio. What is data science? Fundamental concepts and a heuristic example. In: Data Science, Classification, and Related Methods. Springer, Tokyo, p. 40-51, 1998.

(2) DONOHO, David. 50 years of data science. Journal of Computational and Graphical Statistics, v. 26, n. 4, p. 745-766, 2017.

(3) DE MAURO, Andrea; GRECO, Marco; GRIMALDI, Michele. What is big data? A consensual definition and a review of key research topics. In: AIP conference proceedings, p. 97-104, 2015.

(4) FAYYDAY, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37, 1996.

(5) HAN, Jing; LE, Guan; DU, Jian. Survey on nosql database. In 6th international conference on Pervasive computing and applications (ICPCA), p. 363-366, 2011.

(6) CHEONG, Fah-Chun. Internet agents: spiders, wanderers, brokers, and bots. New Riders Publishing, 1996.

(7) CHANDOLA, Varun, BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey. ACM computing surveys (CSUR), v. 41, n. 3, p. 15, 2009.

(8) PATCHA, Animesh; PARK, Jung-Min. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, v. 51, n. 12, p. 3448-3470, 2007.

5.0 - DADOS BIOGRÁFICOS

André Emilio Toscano é Mestre pela Faculdade de Engenharia Elétrica e de Computação da UNICAMP e graduado Faculdade de Tecnologia de Taquaritinga – FATEC-TQ. É cofundador da Venidera e participou no time de pesquisa e desenvolvimento do Venidera Miran.

Marcos de Almeida Leone Filho é Doutor e Mestre pela Faculdade de Engenharia Elétrica e de Computação da UNICAMP e graduado em Engenharia Mecânica pela UNICAMP. É diretor executivo e coordena as atividades de pesquisa e desenvolvimento da Venidera Pesquisa e Desenvolvimento.

Makoto Kadowaki é Doutor pela Faculdade de Engenharia Elétrica e de Computação da UNICAMP, Mestre em Engenharia Elétrica pela Escola de Engenharia de São Carlos – USP. É graduado em Engenharia Elétrica pela UNESP. É sócio da Venidera e integra o time de pesquisa e desenvolvimento da empresa.

Rafael Giordano Vieira é Mestre pela Faculdade de Engenharia Elétrica e de Computação da UNICAMP e graduado Bacharel em Ciência da Computação pela Universidade do Estado de Santa Catarina (UDESC). É sócio da Venidera e integra o time de pesquisa e desenvolvimento da empresa.

João Borsói Soares é Mestre em Engenharia Mecânica pela USP e graduado em Engenharia de Computação pela UFSCAR. É sócio da Venidera e integra o time de pesquisa e desenvolvimento da empresa.