Evento: XXIV Jornada de Pesquisa

APLICAÇÃO DE BIG DATA ANALYTICS ÀS SMART GRIDS: UMA REVISÃO BIBLIOGRÁFICA[1]

BIG DATA ANALYTICS APPLICATION TO SMART GRIDS: A LITERATURE REVIEW

Ivan E. M. Kühne[2], Jonas F. Schreiber[3], Airam T. Z. R. Sausen[4], Maurício De Campos[5], Paulo S. Sausen[6]

[1] Pesquisa realizada no Programa de Pós-Graduação em Modelagem Matemática da Unijuí com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001;
[2] Aluno de mestrado do Programa de Pós-Graduação em Modelagem Matemática da Unijuí, bolsista CAPES, oitentaetres@gmail.com;
[3] Aluno de doutorado do Programa de Pós-Graduação em Modelagem Matemática da Unijuí, jonasfs@gmail.com;
[4] Professora Doutora do Programa de Pós-Graduação em Modelagem Matemática da Unijuí, airam@unijui.edu.br;
[5] Professor Doutor do Programa de Pós-Graduação em Modelagem Matemática da Unijuí, campos@unijui.edu.br;
[6] Professor Doutor do Programa de Pós-Graduação em Modelagem Matemática da Unijuí, sausen@unijui.edu.br.

Resumo - Os Sistemas Elétricos de Potência são responsáveis pelo suprimento de energia elétrica aos consumidores, tendo aumentando consideravelmente a sua complexidade ao longo dos anos. Entretanto, em muitos casos a sua infraestrutura ainda é concebida de acordo com padrões que não incorporam as tecnologias modernas de comunicação e computação, o que os torna vulneráveis a ameaças de diversas naturezas. O conceito de Smart Grid incorpora essas tecnologias às redes de distribuição, acrescentando funcionalidades que otimizam a distribuição de energia, minimizam as perdas e proporcionam a capacidade de auto-regeneração. Entretanto, as abordagens tradicionais de análise de dados não conseguem lidar de maneira adequada com o volume, velocidade, variedade, veracidade e valor associados aos dados gerados pelos sensores das Smart Grids. Uma maneira de superar essa limitação é o paradigma de Big Data Analytics, definido como a aplicação de técnicas avançadas de análise a conjuntos grandes de dados.Nesse artigo é apresentada uma revisão de literatura sobre essa abordagem, incluindo as técnicas de análise e as ferramentas computacionais disponíveis, bem como discute o conceito mais amplo de Descoberta de Conhecimento em Bancos de Dados. Acredita-se que essa revisão possa servir de base para outros trabalhos em que haja efetivamente a aplicação das técnicas de Big Data Analytics na análise de dados gerados pelas Smart Grids.

Palavras-Chave: Big Data; Big Data Analytics; Descoberta de Conhecimento em Bancos de Dados; Mineração de Dados; Smart Grid.

Evento: XXIV Jornada de Pesquisa

Abstract - The Electric Power Systems are responsible for supplying electrical energy to consumers, having considerably increased its complexity over the years. In many cases their infrastructure is still designed according to standards that do not incorporate modern communication and computing technologies, which makes them vulnerable to many types of threats. The Smart Grid concept incorporates these modern technologies into distribution grids, adding features that optimize power distribution, minimize losses and provide self-regeneraton ability. However, the traditional data analysis approaches can not adequately handle the volume, velocity, variety, veracity and value associated with data generated by Smart Grids sensors. One way to overcome this limitation is the Big Data Analytics paradigm, defined as applying advanced analysis techniques to large data sets. In this paper is presented a literature review on this approach, including the analysis techniques and computational tools available, and discusses the broader concept of Knowledge Discovery in Databases. It is believed that this review may serve as a basis for other works that effectively apply Big Data Analytics techniques to analyze data generated by Smart Grids.

Keywords: Big Data; Big Data Analytics; Data Mining; Knowledge Discovery in Databases; Smart Grid.

1 - INTRODUÇÃO

Os Sistemas Elétricos de Potência (SEP) são responsáveis pelo suprimento de energia elétrica aos consumidores, abrangendo fases como a geração, transmissão e a distribuição. Conforme Amin e Wollenberg (2005), a infraestrutura responsável por esses processos está cada vez mais interconectada, o que a torna vulnerável a determinados tipos de ameaças que podem se propagar sob a forma de falhas de grande impacto, capazes de afetar as operações de diversos setores da economia que dependem diretamente de um suprimento energético seguro e confiável.

Segundo Amin e Wollenberg (2005), os princípios e elementos principais das operações interconectadas dos Sistemas Elétricos de Potência foram estabelecidos ainda na década de 1960, antes da adoção massiva de computadores e de redes de comunicações. Dessa forma, apontam os autores, uma grande parte da coordenação das operações ainda é realizada sem a utilização desses recursos, em alguns casos ainda ocorrendo através de ligações telefônicas, inclusive, ou especialmente, durante as emergências.

Como forma de superar essas vulnerabilidades e limitações, foi proposto o conceito de Smart Grid. Conforme o Electric Power Research Institute (2008), esse conceito pode ser entendido como "a sobreposição de um sistema unificado de comunicação e controle ao sistema existente de distribuição de energia". Assim, as informações podem ser disponibilizadas às entidades interessadas, como os operadores dos sistemas e os consumidores, dentro de um espaço de tempo adequado para que as decisões corretas possam ser tomadas.

Como resultados dessa sobreposição, o Electric Power Research Institute (2008) aponta a disponibilização ao sistema de distribuição de funcionalidades como a otimização do fornecimento

e da entrega de energia, a minimização das perdas, a capacidade de autorregeneração. Para que esses objetivos possam ser alcançados, Amin e Wollenberg (2005) apontam a necessidade de que cada componente do SEP, como uma subestação ou uma usina, seja capaz de funcionar como um agente independente capaz de se comunicar e de cooperar com os demais componentes. Com isso, é formada uma grande plataforma de computação distribuída.

Além disso, conforme Amin e Wollenberg (2005), cada componente deve estar conectado a uma série de sensores, de forma que possa avaliar as suas próprias condições de funcionamento através da análise dos dados colhidos por esses sensores, bem como comunicar essas condições aos componentes vizinhos. Entretanto, trabalhos mais recentes revelam que os dados obtidos a partir da aplicação massiva de sensores nas Smart Grids possuem características que fazem com que não possam ser analisados adequadamente quando são utilizadas técnicas tradicionais.

Devido à grande quantidade de sensores e à quantidade de parâmetros, ocorre a geração de uma quantidade significativa de dados. A velocidade com que esses dados são coletados e armazenados também é elevada em relação a sistemas tradicionais. Além disso, as características de funcionamento das Smart Grids fazem com que possam haver falhas de leitura e de comunicação, gerando muitas vezes inconsistências nos bancos de dados utilizadas.. Assim, trabalhos como os de Stimmel (2014), Sagiroglu *et al.* (2016) e Zhou, Fu e Yang (2016) apontam para a necessidade de aplicação do paradigma de Big Data Analytics, definido por Russom (2011) como a aplicação de técnicas avançadas de análise a conjuntos grandes de dados.

Nesse trabalho é apresentada uma revisão bibliográfica sobre a aplicação do paradigma de Big Data Analytics na análise dos conjuntos de dados gerados pelas Smart Grids. Essa revisão abrange os conceitos de Big Data e de Big Data Analytics, as possibilidades de aplicação desses conceitos às Smart Grids, algumas ferramentas computacionais que podem ser utilizadas na condução desse processo e, por fim, é apresentado o paradigma de Descoberta de Conhecimento em Bancos de Dados, defendido por Fayyad, Piatetsky-Shapiro e Smyth (1996) como necessário para que as informações encontradas sejam efetivamente transformadas em conhecimento relevante.

2 - METODOLOGIA

Esse trabalho se caracteriza como uma pesquisa exploratória, destinada a um levantamento inicial sobre as possibilidades de aplicação do paradigma de Big Data Analytics na análise das Smart Grids. Esse levantamento inicial se destina a subsidiar trabalhos futuros onde existe a necessidade de trabalhar com os bancos de dados do setor elétrico que tornam-se cada vez maiores em termos de volume e quantidade de informações com o crescimento das Smart Grids.

Como forma de se atingir esse objetivo, foi analisada a literatura científica disponível em revistas como as da Association for Computing Machinery (ACM), da Elsevier e do Institute of Electrical and Electronic Engineers (IEEE). Para o estudo das ferramentas computacionais disponíveis, foram consultadas fontes como as páginas eletrônicas das organizações responsáveis pelo desenvolvimento de cada uma delas, uma vez que a literatura científica tradicional pode ter

dificuldade de acompanhar o dinamismo dessa área.

Foi dada uma atenção especial aos trabalhos publicados pelo IEEE. Essa opção se justifica pelo fato do IEEE ser uma entidade de referência tanto na Engenharia Elétrica quanto na Ciência da Computação, sendo responsável pela elaboração de diversos padrões que são utilizados atualmente na computação e nas comunicações. Além disso, o IEEE conta com grupos de trabalho destinados a debater a evolução das Smart Grids, como o IEEE Working Group on Big Data Analytics, Machine Learning & Artificial Intelligence in the Smart Grid (IEEE SMART GRID, 2019).

Dentro da pesquisa sobre a parte conceitual do paradigma de Big Data Analytics foram utilizados algumas referências mais antigas, uma vez que a compreensão da sua evolução está relacionada ao conhecimento de trabalhos seminais como os de McCulloch e Pitts (1943) e de Cox e Ellsworth (1997). Entretanto, para a análise das possibilidades de aplicações do paradigma de Big Data Analytics às Smart Grids foram utilizados apenas trabalhos publicados dentro dos últimos dez anos, como forma de se acompanhar o Estado da Arte dessa abordagem. Também de acordo com esse objetivo, foram priorizados trabalhos que abordam a utilização de dados reais.

3 - RESULTADOS E DISCUSSÃO

Nessa Seção são apresentados os resultados da revisão bibliográfica realizada. Na Subseção 3.1 são apresentados os conceitos de Big Data e de Big Data Analytics. Na Subseção 3.2 são apresentadas as possibilidades de aplicação de Big Data Analytics às Smart Grids. Na Subseção 3.3 são apresentadas as técnicas de análise disponíveis. Na Subseção 3.4 são apresentadas as ferramentas computacionais disponíveis para a execução do processo de Big Data Analytics. Por fim, na Subseção 3.5 é apresentado o conceito de Descoberta de Conhecimento em Bancos de Dados, que propõe a incorporação do paradigma de Big Data Analytics num contexto mais amplo como forma de descoberta de conhecimento relevante.

3.1 Conceitos de Big Data e Big Data Analytics

Uma das primeiras menções ao termo Big Data é encontrada no trabalho de Cox e Ellsworth (1997), que utilizaram o termo para se referir ao grande volume de dados utilizado em Visualização Científica. Conforme os autores, os conjuntos de dados utilizados nessa área geralmente são grandes, esgotando a capacidade da memória principal, do armazenamento local e mesmo do armazenamento remoto. Dessa forma, os autores apontavam para a necessidade de desenvolvimento de algoritmos e de técnicas de gerenciamento de memória capazes de superar essas limitações.

Hoje essas técnicas de análise massiva de dados são aplicadas em diversas áreas do conhecimento, abrangendo tanto aplicações puramente acadêmicas quanto aquelas voltadas para a resolução de problemas na indústria e no setor de serviços. Assim, podem ser encontradas, por exemplo, aplicações voltadas para a agricultura, para a descoberta de padrões nas preferências dos consumidores, para a análise de mercado, para a segurança cibernética e para a medicina.

Essas aplicações são baseadas na coleta e análise de grandes volumes de dados, que são analisados com objetivo de que sejam encontradas informações relevantes.

Entretanto, devido a essa diversidade de aplicações e ao fato de ser uma área relativamente incipiente, é difícil estabelecer um consenso sobre as definições de Big Data e de Big Data Analytics. São encontradas variações quando são analisados os trabalhos acadêmicos e a forma como as organizações comerciais definem os serviços que oferecem, principalmente quando essas fontes não são relacionadas à mesma área do conhecimento. Em certos casos é adotada a terminologia Mineração de Dados (Data Mining), à qual é atribuída um sentido análogo ao que outras fontes atribuem aos conceitos de Big Data ou Big Data Analytics.

No trabalho de He *et al.* (2017), por exemplo, é proposta a definição de Big Data como uma abordagem cognitiva baseada em dados, que procura descobrir as correlações estatísticas indicadas por parâmetros de alta dimensão através de um modelo não-paramétrico. Entretanto, conforme os autores, ainda não existe uma definição formalizada para essa abordagem. No trabalho em questão e em trabalhos anteriores dos mesmos autores, é proposta uma formalização matemática da definição apresentada.

Para esse trabalho, foi adotada a definição de Big Data Analytics proposta por Russom (2011) e adotada pelo comitê IEEE Smart Grid no relatório técnico Big Data Analytics in the Smart Grid (IEEE, 2018). Segundo essas fontes, Big Data Analytics é a aplicação de técnicas avançadas de análise a conjuntos grandes de dados. Entre as técnicas disponíveis, Russom (2011) e o IEEE Smart Grid (2018) citam Análise de Dados (Data Analytics), Análise Estatística, Aprendizado de Máquina (Machine Learning), consultas SQL, Inteligência Artificial, Mineração de Dados, Processamento de Linguagem Natural e Visualização de Dados.

Conforme foi sendo desenvolvido o paradigma de Big Data e de Big Data Analytics, ele foi associado a três "vês". Esses "vês" representam características relacionadas à maneira como os conjuntos de dados são gerados e aos desafios relacionados ao seu processo de análise: Volume, Velocidade e Variedade. Posteriormente, foram associadas outras duas características, associadas à importância desses dados: a Veracidade e o Valor.

Segundo Sagiroglu *et al.* (2016), o Volume está associado ao número de registros e à capacidade de armazenamento necessária; a Velocidade está associada à frequência da geração, transferência ou coleta dos dados; a Variedade está associada à diversidades de fontes e formatos dos dados, bem como aos campos multidimensionais; a Veracidade está associada à confiabilidade e à qualidade dos dados; e o Valor está associado à possibilidade de descoberta de padrões e *insights* úteis.

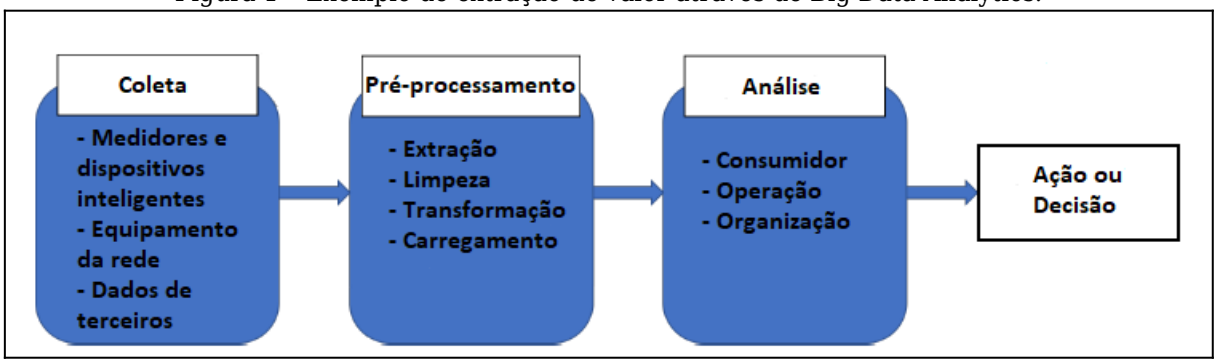
3.2 - Aplicação de Big Data Analytics às Smart Grids

Conforme Sagiroglu *et al.* (2016), existe uma conformidade total entre as características das Smart Grids e o modelo de cinco "vês" apresentado na Subseção 3.1. O Volume está relacionado à grande quantidade de dados que é gerada pelos medidores inteligentes e pela combinação de

outras fontes; a Velocidade está relacionada à necessidade de análise em tempo real devido à grande geração de dados e o tempo requerido para o seu processamento; a Variedade está relacionada ao fato de que os dados obtidos podem estar representados de formas estruturada, semi-estruturada e não-estruturada; a Veracidade está relacionada à necessidade de que os dados sejam coletados adequadamente , de forma que não causem instabilidade nos sistemas; e o Valor está relacionado ao fato de que, conforme Zhou, Fu e Yang (2016), não é suficiente que haja a coleta de dados se eles não forem utilizados de forma a auxiliar na tomada de decisões.

Na Figura 1 é apresentado um exemplo de como o Big Data Analytics pode ser utilizado para extrair valor a partir dos dados gerados pelas Smart Grids. Assim, segundo Stimmel (2014), os dados de fontes como medidores e dispositivos inteligentes são coletados, pré-processados e analisados, gerando informações que subsidiam ações e tomadas de decisões.

Figura 1 – Exemplo de extração de valor através de Big Data Analytics.



Fonte: adaptado de Stimmel (2014).

Segundo o IEEE Smart Grid (2018), existem quatro categorias de análise que podem ser realizadas nas Smart Grids através do paradigma de Big Data Analytics: descritiva, diagnóstica, preditiva e prescritiva. A seguir são explicadas cada uma dessas categorias de análise, de acordo com essa fonte e com Dang-Ha, Olsson e Wang (2015).

a) Análise descritiva: possui o objetivo de fornecer informação sobre o que já ocorreu e se constitui do primeiro passo na tentativa de identificação de dados e informações uteis para processamento adicional. Pode incluir visualização dos dados, mineração de dados e compilação de relatórios;

b) Análise diagnóstica: possui o objetivo de entender as causas de eventos e o comportamento do sistemas, de forma a identificar desafios e oportunidades;

c) Análise preditiva: é utilizada para fazer predições probabilísticas de forma a identificar tendências, com objetivo de determinar o que pode ocorrer no futuro;

d) Análise prescritiva: é aplicada na identificação do melhor resultado possível de eventos, dados os parâmetros do sistema, e na elaboração de estratégias para a gestão de eventos similares no

futuro. Utiliza ferramentas como técnicas de simulação e suporte de decisões para explorar estratégias ótimas para que se possa aproveitar uma oportunidade futura ou mitigar um risco futuro.

3.3 - Técnicas Disponíveis

Através da revisão da literatura, foram identificadas, preliminarmente, algumas técnicas de Big Data Analytics aplicáveis na análise dos conjuntos de dados gerados pelas Smart Grids, como o Aprendizado de Máquina, as Árvores de Decisão, a Modelagem Matemática, as Redes Neurais Artificiais e a Visualização de Dados. Embora todas essas técnicas se mostrem promissoras para a utilização no conexto das Smart Grids, a limitação de tamanho desse trabalho impede que sejam abordadadas cada uma delas. Dessa forma, foram selecionadas duas técnicas para serem apresentadas: as Redes Neurais Artificiais e a Visualização de Dados. São apresentados os seus conceitos e algumas de suas possibilidades aplicação no contexto das Smart Grids. Posteriormente são apresentados dois trabalhos baseados em abordagens que, devido a suas peculiaridades, não podem ser enquadradas nas categorias de análise citadas.

As Redes Neurais Artificiais são sistemas que buscam imitar o funcionamento do cérebro humano, através da utilização de nós interconectados baseados no funcionamento dos neurônios. Assim, esses sistemas podem ser utilizados na resolução de problemas complexos, realizando generalizações e inferências sobre os dados que são processados através deles. Atualmente, são utilizados em tarefas como a classificação, o reconhecimento de padrões, o processamento de imagens e o reconhecimento facial.

A criação do conceito de Redes Neurais Artificiais é atribuída ao neurofisiologista Warren McCulloch e ao matemático Walter Pitts, que propuseram um modelo matemático capaz de representar o funcionamento do cérebro humano no artigo A Logical Calculus of the Ideas Immanent in Nervous Activity (MCCULLOCH; PITTS, 1943). Embora de forma não imediata, o modelo matemático proposto serviu de base para a criação de circuitos eletrônicos capazes de simular as atividades do cérebro humano, cuja complexidade aumentou de acordo com a evolução do hardware e software disponíveis.

As interconexões entre os nós possuem, cada uma, uma importância relativa, denominada peso, que é ajustada de forma iterativa até que a Rede Neural Artificial, de acordo com determinados dados de entrada, seja capaz de produzir dados de saída que simulem o sistema real com uma precisão considerada suficiente. Esse ajuste é guiado por abordagens como o Aprendizado Supervisionado, em que os dados de saída são previamente conhecidos, e o Aprendizado Não-Supervisionado, em que não existe esse conhecimento.

Conforme apresentado por Gurney (2014), os nós que constituem as Redes Neurais Artificiais são elementos que recebem um sinal de entrada e, baseados na avaliação do valor desse sinal, produzem ou não um sinal de saída. Esse sinal de saída, quando ocorre, é propagado para outros nós, de acordo com a forma como as conexões são organizadas, podendo também ser propagado

para a entrada do mesmo nó. Quando é usado o algoritmo *backpropagation*, por exemplo, os sinais referentes ao erro em relação à saída desejada são enviados para os nós anteriores como parte do processo de ajuste dos parâmetros da rede.

Os nós das Redes Neurais Artificiais são organizados em camadas, sendo que a primeira é chamada de "camada de entrada", por receber os sinais de entrada, enquanto a última é chamada de "camada de saída", por fornecer os valores de resposta. Essa é menor quantidade de camadas necessária para que a rede possa ser constituída, podendo ser necessária a adição de camadas intermediárias, denominadas "camadas ocultas", para que problemas mais complexos sejam resolvidos. Como forma de se referir a Redes Neurais Artificiais compostas por muitas camadas, pode ser encontrado na literatura o termo Aprendizagem Profunda (Deep Learning).

No trabalho de Mosaddegh, Cañizares e Bhattacharya (2018) é abordada a utilização de Redes Neurais Artificias como forma de se encontrar a resposta ótima nos alimentadores dos SEP em relação à demanda a que estão submetidos. Os autores adotaram uma abordagem em que uma Rede Neural Artificial é utilizada para modelar um sistema de controle de carga, de forma que possam ser encontrados os parâmetros que minimizam as cargas de pico e o consumo de energia ao longo do sistema de distribuição. O modelo desenvolvido pelos autores foi validado através da utilização de dados reais.

A Visualização de Dados é uma técnica de análise baseada na representação de conjuntos de dados através de meios visuais, como gráficos, diagramas e mapas, de forma que o seu sentido possa ser comunicado. De acordo com Defanti e Brown (1991), cientistas lidam com grandes conjuntos de dados, advindos de fontes como supercomputadores, satélites, veículos espaciais, sistemas médicos de diagnóstico e conjuntos de instrumentos relacionados a eventos geológicos. Entretanto, o cérebro humano não é capaz de interpretar esses grandes volumes de dados, de forma que uma parte considerável dos dados é desperdiçada.

Assim, conforme os autores, os cientistas precisam de uma alternativa às representações numéricas, de forma que os dados possam ser interpretados de forma efetiva e as descobertas possam ser comunicadas a outras pessoas. Os autores apontam o uso da Visualização de Dados como um método computacional capaz de gerar uma representação visual de dados complexos, permitindo que os cientistas visualizem padrões que permaneceriam ocultos caso os dados fossem representados de outras formas (DEFANTI; BROWN, 1991). Segundo Mani e Fei (2017), conforme citado por Fahad e Yahya (2018), a Visualização de Dados é uma forma poderosa e segura de análise de quantidades massivas de dados, de encontro de tendências, de reconhecimento de similaridades e conexões acidentais e de apresentação de informações para outras pessoas, conforme apresentado a seguir.

A análise de quantidades massivas de dados está relacionada ao fato da Visualização de Dados possibilitar que os tomadores de decisão, quando são apresentados às representações gráficas, poderem perceber de forma quase imediata o que está implicado naqueles dados. Isso ocorre de forma muito mais imediata do que quando a análise é feita, por exemplo, com base na

apresentação dos dados em planilhas. O encontro de tendências está relacionada com o fato de que, embora elas estejam representadas de forma implícita quando os dados são apresentados na forma de sequências temporais, não são facilmente reconhecíveis dessa forma. Assim, a aplicação da Visualização de Dados permite que essas tendências sejam encontradas de forma mais óbvia.

O reconhecimento de similaridades e conexões acidentais está relacionado à possibilidade de realização de diversas operações sobre os conjuntos de dados, como a inclusão e exclusão de conjuntos, mudança de escalas, eliminação de *outliers* e mudança da foma de representação visual. Dessa forma, podem ser realizadas análises mais criteriosas, capazes de revelar as percepções que pode ser obtidas a partir dos dados. Por fim, a apresentação das informações para outras pessoas é descrito como um aspecto da Visualização de Dados que muitas vezes é negligenciado, mas que permite que as tendências encontradas sejam comunicadas de forma imediata, clara e altamente eficiente (MANI; FEI, 2017 apud FAHAD; YAHY, 2018).

Além das técnica apresentadas, foram identificadas na literatura outras possibilidades de aplicações de ferramentas matemáticas avançadas na análise dos dados advindos das Smart Grids. No trabalho de Moghaddass e Wang (2018) foi proposto um framework hierárquico para a detecção de anomalias através da utilização dos dados coletados pelos medidores inteligentes presentes nas instalações dos consumidores. Os autores desenvolveram um modelo matemático que leva em consideração não só os dados relativos a cada a instalação, mas também aqueles de outras instalações, de forma que a detecção de anomalias passa a ter mais acurácia.

Como diferencial de trabalho semelhantes, Moghaddass e Wang se propuseram a desenvolver uma ferramenta capaz de prever a ocorrência de anomalias antes da sua ocorrência e capaz de trabalhar com vários níveis de anomalias, como falhas de média e grave complexidade. Assim, a companhia responsável pode desenvolver estratégias de resposta adequadas de acordo com o nível da anomalia detectada. Cabe ressaltar que nesse trabalho não foram utilizados dados reais no desenvolvimento e validação do modelo matemático, mas um conjunto sintético de dados criado com base na Distribuição de Poisson (MOGHADDASS; WANG, 2018).

No trabalho de He *et al.* (2017) foi proposta uma arquitetura para análise de Smart Grids através das técnicas de Big Data. A arquitetura proposta é baseada na teoria matemática das Matrizes Aleatórias. Conforme os autores, os métodos tradicionais de análise são incapazes de lidar com as características intrínsecas ao Big Data, como volume, velocidade e variedade, por trabalharem com modelos simplificados. Dessa forma, os autores propuseram essa abordagem como uma forma mais eficiente de execução de análises de alta dimensionalidade.

3.4 - Ferramentas Computacionais Disponíveis

Através da revisão da literatura, foram identificadas algumas ferramentas computacionais aplicáveis na execução do processo de Big Data Analytics. A seguir são apresentadas as ferramentas MATLAB, Python, R e Waikato Environment for Knowledge Analysis (WEKA), bem como algumas de suas possibilidades de aplicação no contexto de Big Data Analytics.

a) MATLAB: conforme a MathWorks (2019a), empresa responsável pelo seu desenvolvimento e comercialização, é uma combinação entre uma linguagem de programação capaz de trabalhar diretamente com matrizes e vetores e de um ambiente de desenvolvimento voltado para a análise iterativa. Além disso, conta com pacotes complementares ("toolboxes") que são desenvolvidos profissionalmente e são rigorosamente testados e documentados. Esses pacotes expandem as possibilidades de utilização em campos como o Aprendizado de Máquina, o Processamento de Sinais e Visão Computacional.

De acordo com a MathWorks (2019b), o MATLAB oferece diversas funcionalidades que permitem a sua aplicação de acordo com o paradigma de Big Data Analytics. Entre essas funcionalidades, estão o acesso a diversos tipos de fontes de dados, como arquivos, bancos de dados, armazenamento em nuvem, além de fontes em tempo real como dados adquiridos por hardware e relatórios financeiros; capacidade de pré-processamento do dados de forma rápida através de funções de alto nível; e uma grande variedade de modelos de classificação e regressão dos dados, que podem ser comparados e ajustados de forma a se adequarem aos dados analisados.

A aplicação do MATLAB é exemplificada no trabalho de Lü *et al*. (2019), onde foi empregada na criação de um modelo destinado à previsão em curto prazo da carga a que um Sistema Elétrico de Potência vai estar submetido. Foram combinadas técnicas como a Análise de Componentes Principais, os Algoritmos Genéticos e as Redes Neurais Artificiais na criação do modelo, de forma a gerar uma maior acurácia e um menor tempo de processamento, além de se evitar que os resultados pudessem ficar presos a pontos ótimos locais. Como resultado dessa abordagem, os autores concluem que atingiram os objetivos propostos, gerando uma boa correspondência entre os os dados reais, utilizados na construção e validação do modelo, e os dados que foram simulados posteriormente.

b) Python: linguagem de programação mantida pela Python Software Foundation, uma entidade independente e sem fins lucrativos que detém o seu *copyright* desde a versão 2.1. A Python Software Foundation (2019) descreve o Python como uma linguagem de alto nível de propósito geral, capaz de ser aplicada a diversas classes diferentes de problemas, e possuidora de uma sintaxe extremamente simples e consistente. Além disso, possui bibliotecas voltada diretamente para o trabalho com conceitos relacionados ao Big Data Analytics, como a Scikit-Learn, que oferece funcionalidades para o pré-processamento, a classificação e a clusterização dos dados, a seleção de modelos e a aplicação de algoritmos de regressão (SCIKIT-LEARN, 2019).

O Python é reconhecidamente uma boa linguagem para programadores iniciantes, devido a características que fazem o estudante focar em aspectos fundamentais das habilidades iniciais que precisa adquirir. Assim, o estudante não tem a sua evolução atrasada por questões complexas adicionais que, embora importantes, podem ser compreendidas em uma etapa posterior do seu aprendizado (PYTHON SOFTWARE FOUNDATION, 2019). Diversos cursos abordando a utilização de Python para Big Data Analytics, sejam eles pagos ou gratuitos, são oferecidos por plataformas virtuais de aprendizagem como a Coursera, a Data Science Academy, a Udacity e a Udemy.

No trabalho de Fahad e Yahya (2018) é descrita a utilização do Python para a técnica de Visualização de Dados, apontando a confiabilidade dessa linguagem como um fator que a faz ser adotada por desenvolvedores que trabalham com análises de dados diversas especialidades. Conforme os autores, existem diversas bibliotecas do Python destinadas a criar representações gráficas e cada uma delas costuma possuir características nativas distintas daquelas apresentadas pelas demais, que precisam ser conhecidas para que seja feita a opção por uma delas. Baseado nisso, é apresentada no trabalho a comparação de características e funcionalidades de algumas bibliotecas como Altair, Bokeh, Ggplot, Pygal e Seaborne.

c) R: conforme a R Foundation (2019), o R é uma linguagem e um ambiente de desenvolvimento projetados para o trabalho com Estatística Computacional e geração de gráficos. Ambos estão disponíveis abertamente sob a Licença Pública Geral GNU, publicada pela Free Software Foundation (2019). Como linguagem de programação, o R oferece uma grande variedade de técnicas estatísticas e gráficas, além de ser altamente extensível. As técnicas estatísticas disponíveis incluem a criação de modelos lineares e não lineares, os testes estatísticos clássicos, a análise de séries temporais, a classificação e a clusterização (R FOUNDATION, 2019).

Entre outras ferramentas voltadas para a análise de dados, o R oferece a interface gráfica Rattle, desenvolvida pela Togaware. Conforme Williams (2009), em artigo que descreve a utilização dessa ferramenta, as interfaces gráficas oferecem necessariamente uma versão básica das técnicas de Mineração de Dados. Entretanto, de acordo com o autor, o Rattle foi desenvolvido justamente de modo a facilitar a transição dessas técnicas básicas para uma análise de dados mais sofisticada através do uso direto da linguagem R.

Conforme a Togaware (2019), o Rattle possibilita que o usuário realize diversos tipos de operações sobre os dados que está analisando. Como exemplos dessas operações, podem ser citados a apresentação de sumários estatísticos e visuais, a transformação para facilitar a criação de modelos, a geração de Aprendizado de Máquina, tanto supervisionado quanto não-supervisionado, e a apresentação gráfica do desempenho dos modelos criados.

Uma das características principais do Rattle é que todas as interações executadas através da interface gráfica são capturadas como uma sequência de instruções da linguagem que R, que pode voltar a ser executada de forma independente. Assim, o Rattle pode ser utilizado como uma ferramenta para o aprendizado do R e para a geração dos modelos iniciais de Big Data Analytics. Em um momento posterior, esses modelos podem ser ajustados através das opções consideravelmente mais poderosas oferecidas pela linguagem R (TOGAWARE, 2019}.

d) WEKA: conforme a Universidade de Waikato, o WEKA foi desenvolvido como parte de seu projeto relacionado ao Aprendizado de Máquina. Como objetivos desse projeto, são citados a disponibilização ampla das técnicas de Aprendizado de Máquina, a aplicação dessas técnicas a problemas reais que interessam à indústria na Nova Zelândia, o desenvolvimento e disponibilização de novos algoritmos de Aprendizado de Máquina e a contribuição na criação de um framework teórico nessa área.

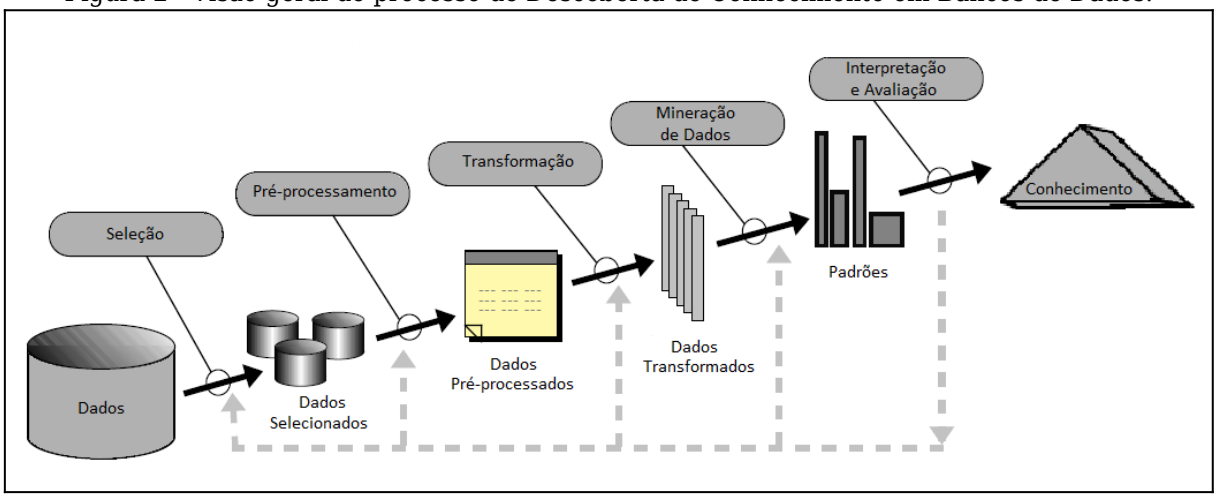
A partir desses objetivos, várias técnicas de Aprendizado de Maquina foram reunidas na solução de software WEKA. Essa solução é distribuída abertamente sob a Licença Pública Geral GNU, publicada pela Free Software Foundation (2019), e permite que um especialista em uma determinada área utilize essas técnicas para encontrar conhecimento relevante em bancos de dados que são grandes demais para que possam ser analisadas manualmente. São oferecidas ferramentas para preparação, classificação, regressão e clusterização dos dados, bem como para a descoberta de regras de associação e visualização (UNIVERSIDADE DE WAIKATO).

3.5 - Descoberta de Conhecimento em Bancos de Dados

O conceito de Big Data Analytics apresentado na Subseção 3.1 e as técnicas apresentadas na Subseção 3.3 podem ser integrados dentro do paradigma mais abrangente de Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases - KDD). Autores como Fayyad, Piatetsky-Shapiro e Smyth (1996) defendem que essa integração é necessária para que as informações encontradas através da aplicação das técnicas de análise sejam transformadas em conhecimento relevante.

Embora Fayyad, Piatetsky-Shapiro e Smyth (1996) utilizem o termo Mineração de Dados, eles atribuem a ele o mesmo significado que Russom (2011) propõe para o termo Big Data Analytics, ou seja, a aplicação de técnicas avançadas de análise a conjuntos grandes de dados. Os autores defendem que a Mineração de Dados deve ser vista como uma das fases do processo de Descoberto de Conhecimento em Bancos de Dados, conforme demonstrado na Figura 2. Assim, a etapa da Mineração de Dados, caracterizada pela aplicação da técnica escolhida, deve ser precedida pela seleção, pré-processamento e transformação dos dados e seguida por uma etapa de interpretação e avaliação dos resultados encontrados.

Figura 2 - Visão geral do processo de Descoberta de Conhecimento em Bancos de Dados.



Fonte: adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), fatores a preparação, a seleção e o tratamento dos dados, a utilização de conhecimentos anteriores e a interpretação adequada dos resultados encontrado são essencial para garantir que conhecimento relevante seja obtido a partir dos dados analisados. Por outro lado, a aplicação cega de técnicas de Mineração de Dados pode ser uma atividade perigosa, que pode levar facilmente ao encontro de padrões inválidos e sem sentido.

Questões semelhantes são debatidas por Cao (2010) ao abordar a aplicação da Mineração de Dados no mundo corporativo. Conforme o autor, existem inúmeros desafios que dificultam que as informações encontradas pela Mineração de Dado sejam aplicadas de forma efetiva na resolução de problemas ou no apoio à tomada de decisões. Assim, segundo ele, os problemas reais dos negócios costumam ser envolvidos por ambientes e fatores complexos, enquanto muitos modelos matemáticos trabalham com abordagens simplificadas. Além disso, muitas vezes a descoberta de padrões é guiada por critérios meramente técnicos, sem levar em conta os interesses empresariais, e há dificuldade na apresentação das informações de maneira que ela possa ser compreendida e utilizada de forma efetiva pelos tomadores de decisões.

4 - CONSIDERAÇÕES FINAIS

Como objetivo desse trabalho, foi proposto um levantamento inicial sobre o paradigma de Big Data Analytics, tendo por base a sua possibilidade de aplicação na análise dos dados gerados pelas Smart Grids. Esse objetivo foi alcançado através da revisão realizada na literatura, que permitiu que se conheçam os conceitos fundamentais desse contexto, as possibilidades de aplicação do paradigma de Big Data Analytics às Smart Grids, as técnicas disponíveis, as ferramentas computacionais disponíveis e o conceito de Descoberta de Conhecimento em Bancos de Dados. Foi possível identificar como o conceito de Big Data Analytics vem sendo desenvolvido ao longo dos anos e hoje não há um consenso sobre a sua definição. Entretanto, foi possível identificar um ponto convergente entre os diversos cocneitos, que é a aplicação de técnicas avaçadas de análise a conjuntos grandes de dados, em busca de conhecimento relevante que possa orientar a tomada de ações e decisões.

A partir da análise dos trabalhos de Fayyad, Piatetsky-Shapiro e Smyth (1996) e Cao (2010), ficou clara a necessidade de que o paradigma de Big Data Analytics não seja entendido como a simples aplicação de ferramentas de análise sobre conjuntos de dados. Ele deve ser entendido como sendo parte de um contexto mais amplo, onde as técnicas matemáticas e computacionais são aplicáveis na resolução de problemas reais e na descoberta de conhecimento relevante para os tomadores de decisões. Acredita-se que, dadas as características de um determinado conjunto de dados e os objetivos que se pretende alcançar com a análise desse conjunto, uma determinada técnica seja mais ou menos eficiente em relação às demais. Dessa forma, essa Revisão Bibliográfica pode ser utilizada como uma forma de subsidiar a decisão sobre a técnica a ser aplicada em trabalhos futuros, em que haja de forma efetiva a análise de dados gerados por Smart Grids em busca de padrões relevantes.

Como limitação desse trabalho, não foram abordadas todas as técnicas de análise associadas ao paradigma de Big Data Analytics, devido ao seu número máximo de páginas. As técnicas que foram identificadas, mas não foram abordadas nesse trabalho, podem ser melhor exploradas em tabalhos futuros. Também não foram abordadas as questões relacionadas à segurança e à privacidade dos dados dos consumidores durante a sua coleta, transmissão e armazenamento pelas Smart Grids, por estarem fora do escopo da pesquisa conduzida. Entretanto, cabe ressaltar que essas questões são importantes no projeto e operação das Smart Grids e são assunto de diversos trabalhos científicos.

REFERÊNCIAS

AMIN, S. M.; WOLLENBERG, B. F. **Toward a Smart Grid: Power Delivery for the 21st Century**. IEEE power and energy magazine, IEEE, v. 3, n. 5, p. 34–41, 2005.

CAO, L. **Domain-Driven Data Mining: Challenges and Prospects**. IEEE Transactions on Knowledge and Data Engineering, IEEE, v. 22, n. 6, p. 755–769, 2010.

COX, M.; ELLSWORTH, D. **Application-Controlled Demand Paging for Out-of-Core Visualization**. In: IEEE. Proceedings of Visualization'97. [S.l.], 1997. p. 235–244.

DANG-HA, T.-H.; OLSSON, R.; WANG, H. **The Role of Big Data on Smart Grid Transition**. In: IEEE. 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity). [S.l.], 2015. p. 33–39.

DEFANTI, T. A.; BROWN, M. D. **Visualization in Scientific Computing**. In: Advances in Computers . [S.l.]: Elsevier, 1991. v. 33, p. 247–307.

ELECTRIC POWER RESEARCH INSTITUTE. **The Green Grid - Energy Savings and Carbon Emissions Reductions Enabled by a Smart Grid**. 2008. Disponível em:. Acesso em: 27 jul. 2019.

FAHAD, S. A.; YAHYA, A. E. **Big Data Visualization: Allotting by R and Python with GUI Tools**. In: IEEE. 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). [S.l.], 2018. p. 1–8.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, v. 17, n. 3, p. 37–37, 1996.

FREE SOFTWARE FOUNDATION. **GNU General Public License**. 2019. Disponível em: <https://www.gnu.org/licenses/gpl-3.0.en.html>. Acesso em: 27 jul. 2019.

GURNEY, K. **An Introduction to Neural Networks**. [S.l.]: CRC Press, 2014.

HE, X. et al. **A Big Data Architecture Design for Smart Grids Based on Random Matrix**

Theory. IEEE Transactions on Smart Grid , IEEE, v. 8, n. 2, p. 674–686, 2017.

IEEE SMART GRID. **Big Data Analytics in The Smart Grid**. IEEE, 2018. Disponível em: <https://resourcecenter.smartgrid.ieee.org/publications/white-papers/SGWP0003.html>. Acesso em: 27 jul. 2019.

IEEE SMART GRID. **IEEE Smart Grid**. Disponível em: <https://smartgrid.ieee.org/>. Acesso em: 29 jul. 2019.

LÜ, Y.-C. et al. **Research on Short-Term Load Forecasting Approach for Smart Grid**. In: IEEE. 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). [S.l.], 2019. p. 602–605.

MATHWORKS. **MATLAB**. 2019. Disponível em: <https://www.mathworks.com/products/matlab.html>. Acesso em: 27 jul. 2019.

MATHWORKS. **Data Science**. 2019. Disponível em: <https://www.mathworks.com/solutions/data-science.html>. Acesso em: 27 jul. 2019.

MCCULLOCH, W. S.; PITTS, W. **A Logical Calculus of the Ideas Immanent in Nervous Activity**. The Bulletin of Mathematical Biophysics, Springer, v. 5, n. 4, p. 115–133, 1943.

MOGHADDASS, R.; WANG, J. **A Hierarchical Framework for Smart Grid Anomaly Detection Using Large-Scale Smart Meter Data**. IEEE Transactions on Smart Grid, IEEE, v. 9, n. 6, p. 5820–5830, 2018.

MOSADDEGH, A.; CAÑIZARES, C. A.; BHATTACHARYA, K. **Optimal Demand Response for Distribution Feeders with Existing Smart Loads**. IEEE Transactions on Smart Grid, IEEE, v. 9, n. 5, p. 5291–5300, 2018.

PYTHON SOFTWARE FOUNDATION. **Welcome to Python.org**. 2019. Disponível em: <https://www.python.org/>. Acesso em: 27 jul. 2019.

R FOUNDATION. **The R Project for Statistical Computing**. 2019. Disponível em: <https://www.r-project.org/>. Acesso em: 27 jul. 2019.

RUSSOM, P. **Big Data Analytics.** TDWI Best Practices Report, v. 19, n. 4, p. 1–34, 2011. Disponível em: <https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf>. Acesso em: 27 jul. 2019.

SAGIROGLU, S. et al. **Big Data Issues in Smart Grid Systems**. In: IEEE. 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA). [S.l.], 2016. p. 1007–1012.

SCIKIT-LEARN. **Scikit-Learn: Machine Learning in Python**. 2019. Disponível em: <https://scikit-learn.org/stable/>. Acesso em: 27 jul. 2019.

STIMMEL, C. L. **Big Data Analytics Strategies for The Smart Grid**. [S.l.]: Auerbach Publications, 2014.

TOGAWARE. **Rattle: A Graphical User Interface for Data Mining Using R**. 2019. Disponível em: <https://rattle.togaware.com/>. Acesso em: 27 jul. 2019.

UNIVERSIDADE DE WAIKATO. **Machine Learning at Waikato University in New Zealand**. s.d. Disponível em: <https://www.cs.waikato.ac.nz/ml/>. Acesso em: 27 jul. 2019.

WILLIAMS, G. J. **Rattle: a Data Mining GUI for R**. 2009. Disponível em: <https://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf>. Acesso em: 27 jul. 2019.

ZHOU, K.; FU, C.; YANG, S. **Big Data Driven Smart Energy Management: from Big Data to Big Insights**. Renewable and Sustainable Energy Reviews, Elsevier, v. 56, p. 215–225, 2016.