UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA METALÚRGICA E DE MATERIAIS

CURSO DE GRADUAÇÃO EM ENGENHARIA METALÚRGICA

SARAH SANTIAGO FAÇANHA

USO DE TÉCNICAS E FERRAMENTAS DE CIÊNCIA DE DADOS PARA REALIZAÇÃO DE ANÁLISES PREDITIVAS SOBRE MEDIÇÕES DE VAZÃO, TEMPERATURA E PRESSÃO DE ETENO NA INDÚSTRIA DE PETRÓLEO E GÁS

FORTALEZA

2019

SARAH SANTIAGO FAÇANHA

USO DE TÉCNICAS E FERRAMENTAS DE CIÊNCIA DE DADOS PARA REALIZAÇÃO DE ANÁLISES PREDITIVAS SOBRE MEDIÇÕES DE VAZÃO, TEMPERATURA E PRESSÃO DE ETENO NA INDÚSTRIA DE PETRÓLEO E GÁS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia Metalúrgica do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia Metalúrgica.

Orientador: Prof. Dr. Marcelo José Gomes da Silva

Coorientador: MSc. Marcus Davi do Nascimento Forte

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

F123u Façanha, Sarah Santiago.
USO DE TÉCNICAS E FERRAMENTAS DE CIÊNCIA DE DADOS PARA REALIZAÇÃO DE ANÁLISES PREDITIVAS SOBRE MEDIÇÕES DE VAZÃO, TEMPERATURA E PRESSÃO DE ETENO NA INDÚSTRIA DE PETRÓLEO E GÁS / Sarah Santiago Façanha. – 2019.
46 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia Metalúrgica, Fortaleza, 2019.
Orientação: Prof. Dr. Marcelo José Gomes da Silva.
Coorientação: Prof. Me. Marcus Davi do Nascimento Forte.

1. Ciência de Dados. 2. Análise Preditiva. 3. Visualização de Dados. 4. Indústria de Óleo de Gás. I. Título.

CDD 669

SARAH SANTIAGO FAÇANHA

USO DE TÉCNICAS E FERRAMENTAS DE CIÊNCIA DE DADOS PARA REALIZAÇÃO DE ANÁLISES PREDITIVAS SOBRE MEDIÇÕES DE VAZÃO, TEMPERATURA E PRESSÃO DE ETENO NA INDÚSTRIA DE PETRÓLEO E GÁS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia Metalúrgica do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia Metalúrgica.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Marcelo José Gomes da Silva (Orientador)
Universidade Federal do Ceará (UFC)

MSc. Marcus Davi do Nascimento Forte (Coorientador)
Univesidade Federal do Ceaŕa (UFC)

Prof. MSc. Raimundo Carlos Martins Leite
Universidade Federal do Ceará - UFC

Dedico este trabalho a Deus,
Aos meus pais, Rogério e Silvani,
Aos meus avós Helano e Lanelde,
Ao meu irmão Yuri,
Às minhas tias Sônia, Silvana e Ana
Ao companheiro incansável, Marcus

AGRADECIMENTOS

Primeiramente a Deus, por me abençoar imerecidamente.

Aos meus pais, Rogério e Silvani, por terem sido exemplo de determinação e garra.

Ao meu orientador, Prof. Dr. Marcelo José Gomes da Silva por ter disponibilizado seu tempo e conhecimento para me instruir.

Ao estimado amigo, Marcus Paulo, que mesmo morando no Canadá, ofereceu imensurável ajuda na área de Ciência de Dados e no estudo de cada uma das análises que realizei.

Aos meus colegas de trabalho, Gabriel Falconieri, João Paulo e Marcel Melo sem os quais não seria possível a realização deste projeto.

Aos coordenadores, professores e alunos do Videira Intensive, que me proporcionaram valioso crescimento de caráter.

À minha amiga, Yasmim Chaves, que pacientemente me suportou durante esse processo.

Ao meu namorado, amigo e parceiro, Marcus Davi, que dedicou longas horas revisando este trabalho, me motivando, me corrigindo e me amando, ilimitavelmente.

“Porquanto somente Eu conheço os planos que determinei a vosso respeito!’, declara Yahweh, ‘planos de fazê-los prosperar e não de lhes causar dor e prejuízo, planos para dar-vos esperança e um futuro melhor.”

(Jeremias 29:11)

RESUMO

Este trabalho apresenta um estudo baseado na teoria de ciência de dados para elaboração de análise preditiva sobre medições de vazão, temperatura e pressão de tubulações de gás eteno sobre um conjunto massivo de dados obtidos de uma planta petroquímica. Este conjunto possui 11 *Tags* com aproximadamente 2 milhões de dados por *Tag*, totalizando cerca de 24 milhões de dados a serem processados. Esses dados possuem medições de diversos sensores presentes nos gasodutos cujas medidas são utilizadas para elaboração de uma análise sobre o comportamento dinâmico do sistema. As consultas obtidas são então utilizadas para verificação de falhas ou medições errôneas nos sensores; estimação do volume total a ser transportado na planta química em até um ano; verificação da correlação entre as medidas das *Tags* e validação das técnicas e ferramentas de ciência de dados utilizadas para realização das análises.

Palavras-chave: Visualização de dados. Ciência de Dados. Análise Preditiva. Indústria de Óleo e Gás

ABSTRACT

This paper presents a study based on Data Science theory to elaborate analyzes on a massive set of data obtained from a petrochemical plant. The set has 12 tags with approximately 5.3 million data per Tag. The dataset has measurements of several sensors present in the petrochemical plant, measurements used to analyze a predictive analysis of the system's dynamic behavior. Statistics are used for fault checking or incorrect sensor measurements; taxation of the total volume of being transported in the chemical plant within one year; verification of the correlation between the measures tags and validation of the techniques and tools Data Science used to perform the analyzes.

Keywords: Data visualization. Data Science. Predictive analysis. Oil and Gas Industry

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ACF	*Autocorrelation Function (Função de Autocorrelação)*
AR	*Autoregressive (Autoregressivos)*
ARIMA	*Autoregressive integrated moving-average*
ARMA	*Autoregressive moving-average*
BI	*Business Intelligence*
ETL	*Extract, Transform and Load*
MA	*Moving Average (Média Móvel)*
MDA	*Multiple Discriminant Analysis*
OLAP	*Online Analytical Processing*
PACF	*Partial Autocorrelation Function (Função de Autocorrelação Parcial)*
SQL	*Structured Query Language*

SUMÁRIO

1 INTRODUÇÃO

Os recentes avanços tecnológicos resultaram em uma geração diária de massivos conjunto de dados nas indústrias de exploração e produção de petróleo e gás. Foi reportado que gerenciar esses conjunto de dados é uma grande preocupação das indústrias das companhias de óleo e gás. Um relatório pelo (BRULE, 2015) afirmou que engenheiros de petróleo e geo cientistas gastam em torno de metade do tempo deles pesquisando e reunindo dados. *Big Data* refere às novas tecnologias de manuseio e processamento desses massivos conjunto de dados. Esses conjunto de dados são armazenados em diferentes variedades e gerados em largo volume em várias operações de *upstream* e *downstream* da indústria de petróleo e gás.

1.1 Upstream, Midstream e Downstream

Upstream se refere a tudo que é ligado a produção e exploração de óleo e gás natural. Sondas geológicas e qualquer coleta de informações usada para localizar áreas específicas onde os minerais são possivelmente encontrados é comumente chamado de “exploração”. O termo *Upstream* também inclui as etapas envolvidas na perfuração e no processo de levar os recursos de óleo e gás natural para a superfície, se referindo a parte de “produção”.

Um outro importante setor dessa indústria é o *Midstream* que se refere a tudo aquilo que é necessário para transportar e armazenar petróleo bruto e gás natural antes de serem refinados e processados em combustíveis e elementos-chave necessários para fazer uma lista longa de produtos que usamos todos os dias. A *Midstream* inclui oleodutos e toda a infraestrutura necessária para movimentar esses recursos por longas distâncias, como estações de bombeamento, caminhões-tanque, vagões-tanque e caminhões-tanque transcontinentais.

Downstream se refere ao setor final da indústria de óleo e gás natural. Inclui tudo aquilo que envolve a transformação de petróleo bruto e gás natural em milhares de produtos finais que utilizamos diariamente. Alguns produtos bem conhecidos como gasolina, diesel, querosene, asfalto para a construção de estradas. E também aqueles não tão óbvio como borrachas sintéticas, preservativos, fertilizantes, recipientes plásticos. O produtos de petróleo e gás natural também são utilizados para fazer membros artificiais e aparelhos auditivos. De fato, tintas, corantes, fibras e praticamente qualquer coisa que tenha sido fabricada têm alguma conexão com petróleo e gás natural.

Além disso, na maioria dos casos, se os dados forem processados eficientemente,

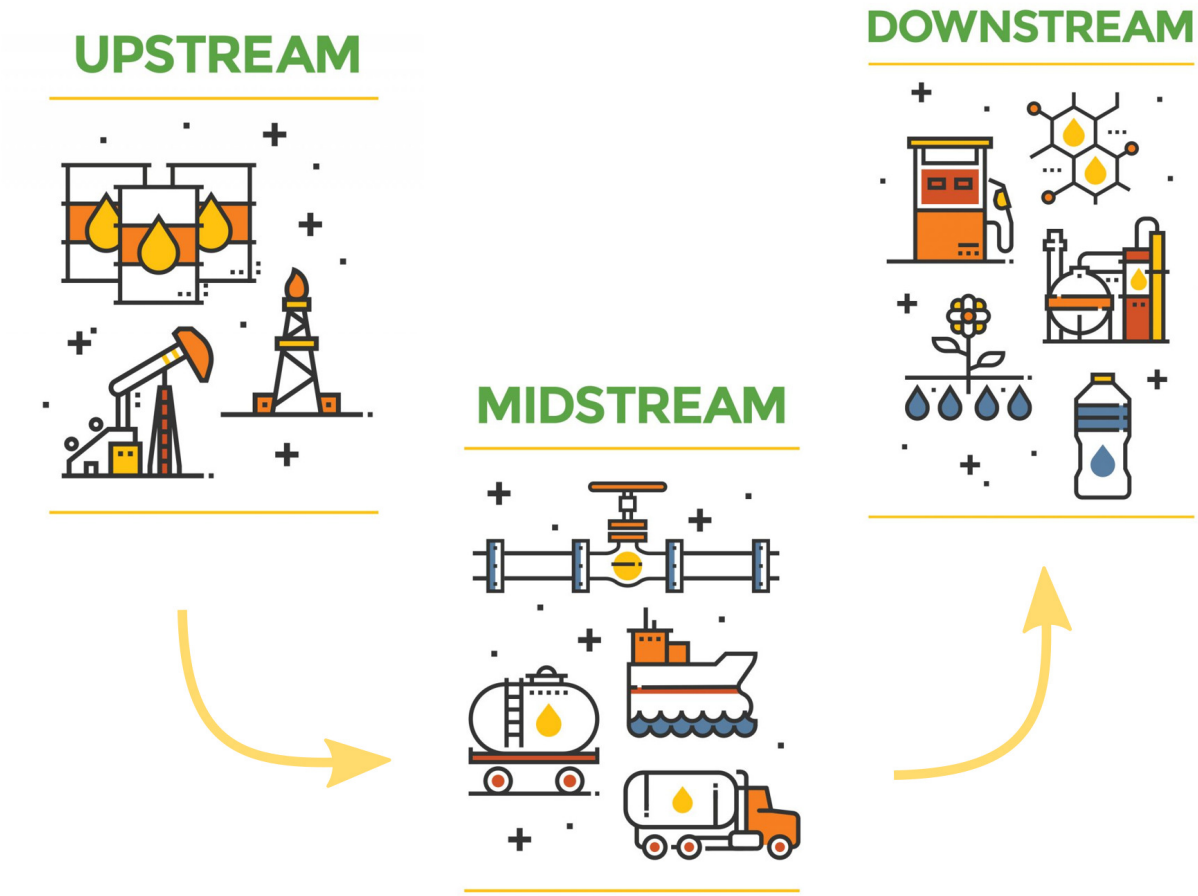

Figura 1 – Fluxo na indústria de petróleo e gás. Fonte : (Upstream?. . . , 2017)

poderão revelar importantes modelos matemáticos que descrevem a dinâmica do problema por detrás dos dados. Baseados nos resultados de pesquisa reportados pelo (MEHTA, 2016), conduzidos pela General Eletric e Accenture, 81% dos executivos consideraram *Big Data* como uma das prioridades top 3 para as companhias de petróleo e gás de 2018. Baseado no artigo deles, a principal razão por trás dessa popularidade é a necessidade de melhorar a exploraçao e produçao de petróleo e gás de forma eficiente. Esse ponto de vista e previsão futura entre executivos para 2018 se tornam mais interessantes quando comparamos a descoberta com as descobertas pelo (FEBLOWITZ, 2013). Baseado em uma pesquisa em 2012 pela IDC Energy, 70% das empresas de petróleo e gás participantes dos EUA não eram familiarizadas com *Big Data* e suas aplicações.

Esse ponto de vista e previsão futura entre executivos para 2018 se tornam mais interessantes quando comparamos a descoberta com as descobertas pelo (FEBLOWITZ, 2013) em 2013. Baseado em uma pesquisa em 2012 pela IDC Energy, 70 % das empresas de petróleo e gás participantes dos EUA não eram familiarizadas com *Big Data* e suas aplicações em *Big Data*.

Volume se refere a quantidade de dados ou informações. Esses dados podem vir de um sensor ou uma ferramenta de gravação de dados. Lidar com essa grande quantidade de dados é considerado um desafio devido a problemas de armazenamento, sustentabilidade e análise (TRIFU, 2014). Muitas empresas estão lidando com um grande volume de dados em seus arquivos; no entanto, eles não têm a capacidade de processar esses dados. A principal aplicação do *Big Data* é fornecer ferramentas de processamento e análise para a crescente quantidade de dados.

O termo velocidade como característica do *Big Data* refere-se à velocidade de transmissão e processamento de dados. Também se refere ao ritmo acelerado da geração de dados. A questão desafiadora sobre o componente velocidade é o número limitado de unidades de processamento disponíveis em comparação com o volume de dados. Recentemente, a velocidade de geração de dados é enorme, pois dados de 5 exabytes são gerados em apenas dois dias. Isso é equivalente à quantidade total de dados criados por seres humanos até 2003.

A característica de velocidade é ainda mais proeminente para a indústria de petróleo e gás devido à natureza complexa de vários problemas de engenharia de petróleo. O processamento de grande quantidade de dados gerados por um indivíduo para um problema complexo é impossível e resulta em atraso e incerteza significativos. Existem muitos casos em que o processamento em tempo real e rápido de dados é crucial na indústria de petróleo e gás. Por exemplo, o processamento rápido dos dados do poço durante a perfuração pode resultar na identificação de chutes e na prevenção eficiente de explosões destrutivas (FEBLOWITZ, 2013).

Variedade refere-se aos vários tipos de dados que são gerados, armazenados e analisados. Os dispositivos e sensores de gravação de dados são diferentes em tipos e, como resultado, os dados gerados podem estar em diferentes tamanhos e formatos.

Veracidade refere-se à qualidade e utilidade dos dados disponíveis para o objetivo da análise e tomada de decisão. Trata-se de distinguir entre dados limpos e sujos. Isso é muito importante logo que os dados ruins podem afetar significantemente a velocidade e a precisão da análise de dados.

1.2 Transporte de petróleo e gás

Devido ao enorme crescimento do consumo de petróleo, o transporte de óleo e gás à longa distância através de tubulações tem se tornado uma das formas essenciais de alocação destes tipos de recurso. Ademais, é necessário que as condições das tubulações sejam prontamente

identificadas, principalmente em caso de vazamento. A identificação de vazamentos através dos sensores é de extrema importância uma vez que as empresas responsáveis por esses estão sujeitas a multas milionárias caso o problema não seja corrigindo a tempo hábil, afetando não somente o meio ambiente como também a imagem da empresa. Para tanto, a identificação de forma rápida e precisa é um dos desafios encarados por este tipo de setor na indústria de gás e óleo. A identificação de vazamentos é normalmente feita através de sensores que mostram a vazão daquele fluido ou a pressão nas tubulações em pontos específicos. (Wei *et al.*, 2009) mostra que esta análise possui geralmente três características

1. Os índices que descrevem o tipo de vazamento estão encontrados na leitura crua dos dados
2. Durante o processo de detecção, o trabalho de analisar são bastante onerosos devido à quantidade massiva de dados e a presença de irregularidades nas leituras
3. Os sensores estão inevitavelmente sujeitos a ruídos e erros de medição

Alguns trabalhos visam propor soluções para este tipo de problema. Em (Wei *et al.*, 2009), os autores propõem o uso de algoritmos *fuzzy cluster* para o manuseio de uma quantidade massiva de dados de pressão de uma tubulação de óleo. O algoritmo seleciona analisa determinadas amostras e extrai as características no domínio de tempo. O algoritmo mencionado faz então um reconhecimento de padrões nos dados para averiguar se houve ou não vazamento na tubulação. Outro trabalho (Jiang ChunLei; Wang Yuan, 2013) usa filtros adaptativos para identificar vazamentos em tubulações de gás, seguida de um estudo de caso real. Este último trabalho reforça a complexidade do problema, citando a dificuldade de averiguar os vazamentos em tempo real devido às longas distâncias.

Nesses casos e em diversos outros, pesquisadores têm buscado solucionar um problema comum às indústrias que lidam com o transporte de óleo ou gás em quaisquer dos setores *up*, *mid* ou *downstream*. A presença de incertezas nas medições dos sensores de vazão ou pressão, juntamente com a grande quantidade de dados apresentam considerável dificuldade para o tratamento adequado dos dados.

Destacam-se outras valiosa utilidade para os dados coletados dos sensores - a possibilidade de gestão das quantidades de gás e óleo que perpassam os dutos de transferência na camada *midstream* da indústria. Com os mesmos conjunto de dados adquiridos de sensores de pressão e fluxo, além de viabilizar a análise de falhas de sensores ou vazamentos, faz-se possível a realização de controle de fluxo e volumetria do fluido passante pelo duto. Esse tipo de dado tem grande importância para o controle de estoque e inventário do fluido. A importância da

análise e previsão de falhas nos sensores é ampliada em função dos custos de manutenção destes tipos de sensores, pois os estes são geralmente acoplados a instalações subterrâneas. Além do custo monetário, avalia-se também o custo gerado pela perda de dados importantes durante um longo período de tempo.

Um outro aspecto de considerável importância no tratamento dos dados em questão é o uso dos mesmos para realização de previsões com base no histórico do sensoriamento. Através de ferramentas de previsão de dados tais como a modelagem ARIMA, é possível estimar o comportamento das curvas dos dados adquiridos. De posse da coleta periódica de determinado sensor ao longo do tempo, é possível constituir modelos matemáticos representativos do sistema do qual o sensor toma suas medidas.

O modelo matemático cuja complexidade é parâmetro de ajuste do sistema de modelagem (ou seja, o projetista pode ajudar o quão complexo o modelo matemático pode ser dependendo do encaixe da curva estimada com as medições reais). Dentre os benefícios desta análise, destacam-se: a possibilidade de se fazer manutenções preventivas sobre os sensores; o conhecimento prévio das medições dos sensores em determinadas épocas do ano.

1.3 Motivação

Diante das dificuldades existentes no processamento das grandes quantidades de dados presentes em várias etapas dos processos de coleta e transporte na indústria de óleo e gás; dos custos envolvidos no processamento manual feito pelos profissionais que operam nessas indústrias quando lidam com tais dados; das incertezas presentes nos sensores podendo levar análise dos dados a diagnósticos equivocados e da carência de análises preditivas para os dados já presentes para estimação do comportamento dos sensores em estações futuras; Percebeu-se uma carência de estudos relacionados ao processamento de conjunto de dados dirigidos a este tipo de indústria.

Além disso, a tendência das indústrias a caminharem em direção ao paradigma de Indústria 4.0 fará com que os sensores estejam cada vez mais integrados ao sistema de modo geral - fornecendo ainda mais dados e exigindo menos de operadores para seu manuseio. Em outras palavras, as soluções do futuro estarão concentradas mais em software do que em hardware. Investimentos em ciência e análise de dados têm se tornado o objeto de estudo de grandes empresas.

A disponibilidade e o avanço de ferramentas de *Big Data* permitem a análise sobre

grandes conjunto de dados para extração de informações relevantes às indústrias de óleo e gás. Uma vez que os dados são devidamente analisados, a indústria poderá utilizar os resultados da análise para elaboração de diversas aplicações que incrementem, de modo geral, à indústria. Aplicações como manutenções preventivas sobre os sensores, por exemplo, podem poupar custos desnecessários à empresa.

1.3.1 Objetivo

O objetivo deste trabalho é a partir do uso de conjuntos massivos de conjunto de dados, a realização da organização, interpretação e análise dos dados adquiridos por meio dos processos de ETL e construção de Cubos OLAP. Com os dados já organizados e interpretados aplica-se o método ARIMA de previsão de linha temporal para que uma análise preditiva construída. Esse método é feito por meio do treinamento do algoritmo usando o conjunto de dados e prevendo dentro de um período estabelecido valores de nível máximo ao mínimo.

2 CONCEITOS E FERRAMENTAS DA CIÊNCIA DE DADOS

Neste capítulo, os aspectos relevantes do tema Ciência de Dados serão abordados em detalhes. Tratar-se-á dos conceitos, definições e termologia deste área, seguida das ferramentas específicas utilizadas para a análise do conjunto de dados obtido de uma industria de óleo e gás.

2.1 Conceitos e Teoria

2.1.1 O que é Ciência de Dados?

A ciência de dados é uma extensão evolutiva das estatísticas capaz de lidar com as grandes quantidades de dados produzidos atualmente. Ela adiciona métodos da ciência da computação para o repertório de estatísticas. Em uma nota de pesquisa de (LANEY, 2012) sobre papel do cientista de dados e da ciência de dados, os autores vasculharam centenas de descrições de cargo para cientista de dados, estatístico e analista de *Business Intelligence* (BI) para detectar as diferenças entre esses títulos. As principais coisas que definem um cientista de dados além de um estatístico, a capacidade de trabalhar com *Big Data* e experiência em aprendizado de máquina, computação e construção de algoritmos. (CIELEN *et al.*, 2016)

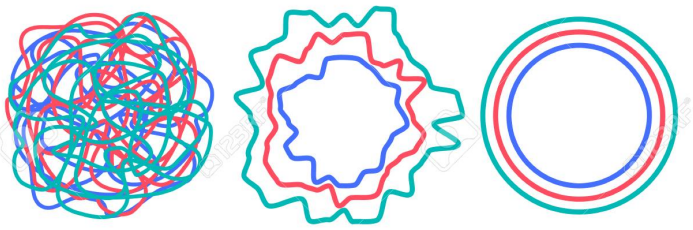

Figura 2 – Progressiva organização de dados. Fonte : (123RF, 2019)

Ciência de Dados é um campo multidisciplinar que utiliza ferramentas, processos e algoritmos para extrair conhecimento e percepções de dados estruturados e não estruturados. Esse conceito inclui conhecimentos de estatística, análise de dados e *Machine Learning*, a fim de compreender e solucionar fenômenos reais a partir dos dados adquiridos.

2.1.2 Categorias dos dados

Em Ciência de Dados e no *Big Data*, encontraremos os mais diversos tipos de dados, dentre os quais podemos citar como os principais:

1	TAG	TIMESTAMP	VALUE	STATUS
2	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:00:36.3	27704.7	Good
3	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:01:36.3	27943.5	Good
4	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:02:36.3	28185.6	Good
5	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:03:36.3	27985.8	Good
6	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:04:36.3	27837.7	Good
7	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:05:36.3	27813.3	Good
8	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:06:36.3	27791.6	Good
9	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:07:36.3	27745.6	Good
10	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:08:36.3	27758.9	Good
11	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:09:36.3	27589.9	Good
12	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:10:36.3	27593.6	Good
13	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:11:36.3	27652.9	Good
14	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:12:36.3	27654.2	Good
15	VAZMP_ETEN_CLIENTE2	01-JAN-14 00:13:36.3	27668	Good

Figura 3 – Uma tabela de Excel é um exemplo de dado estruturado

- Estruturado
- Não Estruturado
- Linguagem Natural
- Gerado por máquina
- Baseado em gráficos
- Áudio, vídeo e imagens
- *Streaming*

Destes acima mencionados, destacam-se os dados Estruturados, Não-Estruturados e Gerados por máquina.

2.1.2.1 Dados Estruturados

Dados estruturados são dados que dependem de um modelo de dados e residem em um campo fixo dentro de um registro. Dessa forma, geralmente é fácil armazenar dados estruturados em tabelas nos bancos de dados ou arquivos do Excel (figura x). *Structured Query Language* (SQL) é uma maneira comumente utilizada para gerenciar e consultar dados que residem em bancos de dados. Dados hierárquicos, como uma árvore genealógica, são um exemplo.

2.1.2.2 Dados Não-Estruturados

Dados não estruturados são aqueles que não são fáceis de ajustar em um modelo de dados pois o conteúdo é contexto específico ou variável. Para exemplificar, podemos citar o e-mail. Embora este contenha elementos estruturados como remetente, assunto e texto do corpo, existem variadas formas de como escrevê-lo. Os milhares de idiomas e dialetos diferentes podem complicar ainda mais isso. Um e-mail escrito por humanos, como mostrado na 4, também é um exemplo perfeito de dados de Linguagem Natural.

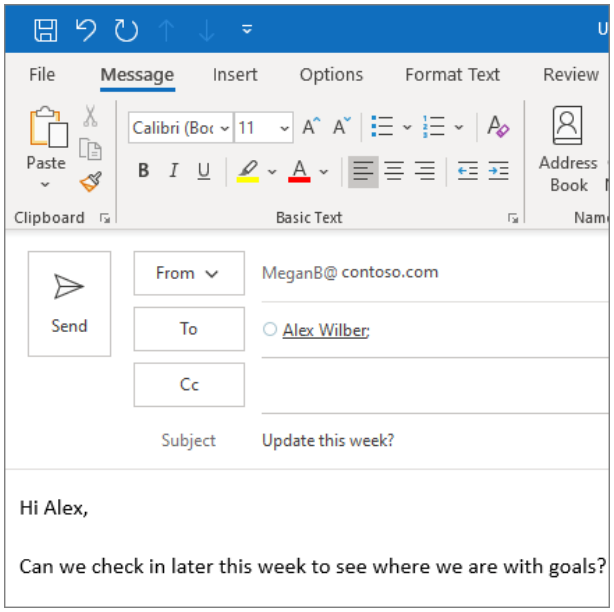

Figura 4 – Exemplo de Dados Não estruturados

2.1.2.3 Dados Gerado por máquina

Dados gerados por máquina são informações criadas automaticamente por um computador, processo, aplicação ou outra máquina sem intervenção humana. Isso significa que os dados inseridos manualmente por um usuário final não são considerados gerados pela máquina. Esses dados cruzam todos os setores que fazem uso de computadores em qualquer uma de suas operações diárias, e os humanos geram cada vez mais esses dados sem saber, ou pelo menos fazem com que sejam gerados pela máquina. Essa categoria está se tornando um recurso importante de dados e continuarão a fazê-lo. A análise dos dados da máquina depende de ferramentas altamente escalonáveis, devido ao seu alto volume e velocidade.

Exemplos de dados da máquina são os inúmeros logs do sistema gerados pelo sistema operacional e outros softwares de infraestrutura no curso normal do dia, bem como a solicitação

de páginas da Web e os logs de fluxo de cliques produzidos pelos servidores da Web.

Consulta_Sarah_2015.csv | Consulta_Sarah_2016.csv | Consulta_Sarah_2017.csv | Consulta_Sarah_2018.csv

```
TEMP ETEN BRASK PVC;17-SEP-16 08:39:33.8;33.5941;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:40:33.8;33.6491;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:41:33.9;33.6858;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:42:33.9;33.7775;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:43:33.9;33.8875;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:44:33.9;34.1076;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:45:33.9;34.2359;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:46:33.9;34.4193;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:47:33.9;34.6027;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:48:33.9;34.7861;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:49:33.9;35.0061;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:50:33.9;35.1712;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:51:34.0;35.3545;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:52:34.0;35.4095;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:53:34.0;35.4462;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:54:34.0;35.4462;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:55:34.0;35.5562;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:56:34.0;35.5746;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:57:34.0;35.5929;Good
TEMP ETEN BRASK PVC;17-SEP-16 08:58:34.0;35.5929;Good
```

Figura 5 – Exemplo de dados gerado pela máquina

2.2 Séries Temporais

Séries temporais são definidas como um conjunto (ou série) de informações cujos dados estão indexados e ordenados pelo tempo. Apesar de uma definição um tanto rebuscada, "série temporal" é um termo estatístico dado a uma coleção de dados que são medidos no tempo, e que a informação de quando foi medida é relevante e armazenada.

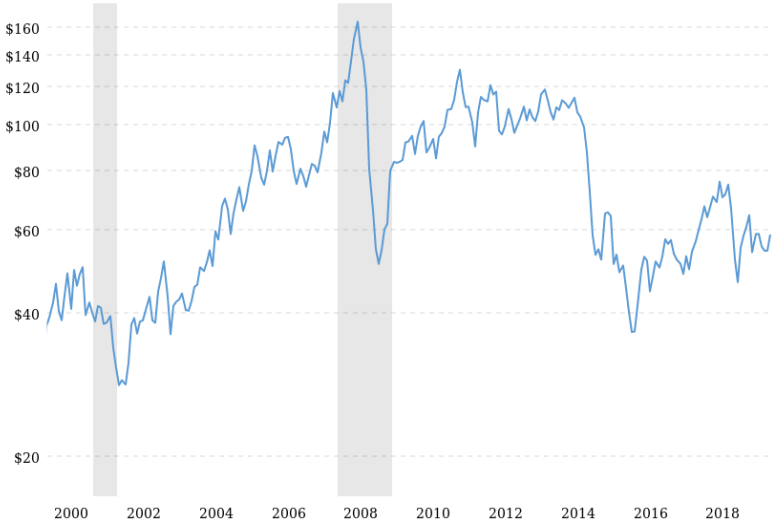

Figura 6 – Gráfico do Preço do Barril de Petróleo em Dólares.Fonte : (MACROTRENDS, 2019)

Um simples exemplo é mostrado no gráfico da Figura 6. Neste gráfico, mostra-se o preço do barril de petróleo em dólares em função do tempo. Os dados são medidos uma vez por hora, e posteriormente é feita uma média ao longo de um mês (Essa técnica chama-se

Roll-Up ou *Pivot Table*) Por ser um conjunto finito (inteiro) de dados, as séries temporais são sistemas caracteristicamente discreto, ou seja, não existem dados entre duas amostras ou dois pontos. Dessa forma, a modelagem e análise matemática feita de tais sistemas deve obedecer aos princípios de sistemas discretos. Uma vantagem de se trabalhar com sistemas discretos é a mais fácil implementação de métodos numéricos computacionais para realização das análises das séries. É sabido, contudo, que a natureza dos sistemas descritos por séries temporais (a maioria) são de fato contínuos, e a discretização ocorre no momento da amostragem.

As séries temporais estão presentes em praticamente todas as áreas da ciência, uma vez que geralmente é desejado que os dados estejam organizados em função do tempo. As cotas das bolsas de valores, o crescimento populacional de uma região, o número de casos de uma doença ou a pressão sanguínea medida de um paciente são poucos dos incontáveis casos em que uma série temporal pode ser usada para descrição para a combinação dos dados. Séries temporais são frequentemente utilizadas nas áreas de reconhecimento de padrões, engenharia de controle, estatística, matemática financeira, dentre outras.

2.2.1 Análise de Séries Temporais

A forma sistemática com a qual alguém responde a perguntas estatísticas e matemáticas relacionadas às séries temporais é chamada de análise de séries temporais (SHUMWAY; STOFFER, 2005). Existem duas abordagens principais : a abordagem pelo domínio do tempo e pelo domínio da frequência.

Sistemas analisados através do domínio da frequência assumem que os dados possuem alguma natureza periódica ou variações sinusoides inerentes aos mesmos. Este comportamento costuma aparecer em sistemas físicos ou biológicos. Por exemplo, quando se deseja compreender a influência da lua sobre as marés (um evento periódico) de forma quantitativa, faz-se uma análise da série temporal (dados coletados da maré) através da abordagem do domínio da frequência (devido à periodicidade da órbita da lua).

Alternativamente, sistemas analisados através do domínio do tempo geralmente são descritos através de equacionamentos que explicitam uma dependência de valores atuais tanto do tempo quanto dos valores passados. É comum usar esse tipo de análise não só para modelar matematicamente uma série temporal, mas também realizar predições futuras com base no histórico da série e nos valores atuais. Em termos de Ciência de Dados, os resultados dos modelos de predição são popularmente usados como ferramentas de previsões dos dados.

Uma classe de modelos desse tipo são conhecidos como modelos ARIMA, que serão mais detalhadamente descritos na seção 2.3

O principal objetivo da análise de séries temporais é desenvolver modelos matemáticos que oferecem uma descrição de um conjunto de dados. Uma vez que os dados possuem natureza caracteristicamente aleatória, define-se as séries temporais como uma coleção de variáveis aleatórias ordenadas pelo tempo na qual foram medidas. Ou seja, dado um conjunto $\{x_1, x_2, x_3\}$, o valor x_t é assumidamente obtido em um tempo anterior ao valor x_{t-1}. Em geral, esse conjunto é também conhecido como um processo estocástico. Um outro e simples exemplo de série temporal puramente aleatória é mostrada na Figura 7.

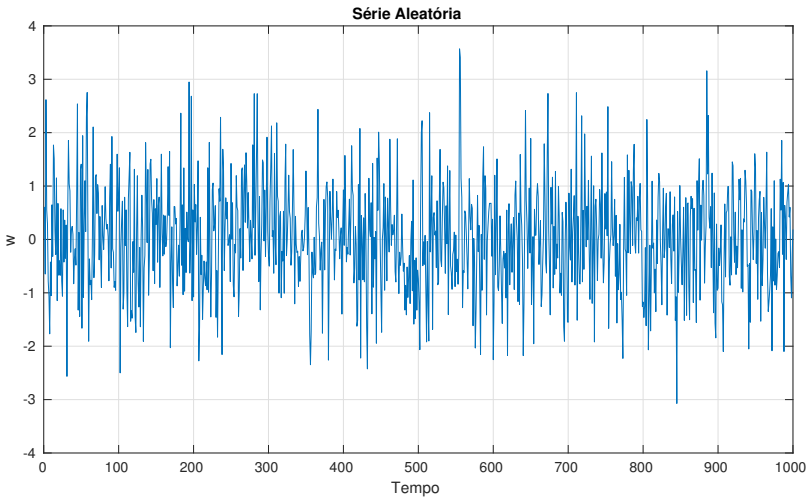

Figura 7 – Ruído Branco

No caso do ruído branco, os valores amostrados independem dos valores anteriores (não-correlacionados) e são estatisticamente descritos por $w_t \sim w_n(0, \sigma_w^2)$. A notação usada indica que a média do sinal aleatório é nula e sua variância é dada por σ^2. Este sinal é bastante utilizado para simulação de ruídos vários sistemas da engenharia. Uma série particularmente útil é o ruído branco Gaussiano, cujas amostras w_t obedecem à distribuição normal, denotadas por $w_t \sim iid\ N(0, \sigma_w^2)$, onde *iid* significa que a série é independente e identicamente distribuída – as variáveis aleatórias possuem a mesma distribuição e são mutualmente independentes.

O uso de ruídos brancos para descrição de sinais permite a utilização de ferramentas estatísticas para análise dos mesmos. Além de este tipo de ruído estar presentes nas amostragens físicas de dados em geral, será visto na seção 2.3 que os modelos ARIMA possuem termos considerados ruído branco no equacionamento.

2.3 Modelos ARIMA

A regressão clássica geralmente não é suficiente na descrição de dados das séries temporais. Isso ocorre porque os dados inseridos em um algoritmo de regressão são entendidos como independentes entre si. Essa natureza introduziu a necessidade de considerar a influência dos valores passados, motivando a criação dos modelos *Autoregressive (Autoregressivos)* (AR), ou auto-regressivos, através de uma dependência linear. Percebeu-se também que as séries poderiam também apresentar uma dependência dos erros de predição passados, motivando a criação dos modelos *Moving Average (Média Móvel)* (MA). A combinação de ambos os modelos resultam em um terceiro tipo, conhecidos por ARMA. Estes últimos são capaz de representar séries temporais fracamente estacionárias [1], que de fato descrevem razoável parte das séries temporais com as quais se trabalham nas mais diversas áreas. Destaca-se que os modelos supramencionados são utilizados para realização de predições, ou *forecast*, uma vez que os modelos tenham descrito bem os sinais adquiridos de uma determinada amostra. Existem índices estatísticos que indicam o quanto o modelo estimado se encaixa no conjunto de dados original para validação do modelo adquirido.

Contudo, a ausência de características estacionárias de alguns processos aleatórios fazem com que modelos ARMA não se apliquem. As leituras de um sensor de vazão de um gaseoduto, por exemplo, varia com o momento do dia. Para este tipo de aplicação, o modelo ARMA não seria um bom representativo da dinâmica da vazão do gaseoduto. Para lidar com este tipo de dados, foi criado o modelo ARIMA. A única diferença entre ARIMA e ARMA é que o ARIMA aplica uma diferenciação (derivada discreta) sobre os dados. Esta operação procura transformar dados não estacionários em dados estacionários, viabilizando a abordagem ARMA.

2.3.1 Modelos Autoregressivos (AR)

Os modelos AR são tipicamente descritos pela equação

$$Y_t = c + \varepsilon_t + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p}, \tag{2.1}$$

onde c é uma constante (geralmente a média do sinal), θ_t é o coeficiente de ponderação de um valor passado e Y_t é a saída do modelo no tempo t. O termo ε representa uma parcela do modelo contento o erro de predição, sendo caracteristicamente um ruído branco.

[1] Processos estacionários não alteram suas propriedades estocásticas com o tempo

Modelos AR possuem dependência explicita de valores amostrados passados e podem ser utilizados para representação de processos aleatórios. A ordem da dependência é parâmetro de entrada para a estimação de uma série temporal, e normalmente se utiliza a notação $AR(p)$, onde p é a ordem do modelo AR.

Perceba que, de posse dos coeficientes do modelo AR, é possível realizar predições a respeito de um conjunto de dados. A obtenção dos coeficientes é feita através de algoritmos (mínimos quadrados por exemplo) que buscam minimizar o erro entre o conjunto de dados e o modelo (2.1). As predições são feitas recursivamente, escrevendo-se

$$Y_{t+1} = c + \theta_1 Y_t + \theta_2 Y_{t-1} + \cdots + \theta_p Y_{t-p+1}, \tag{2.2}$$

sucessivamente, até o número desejado de predições.

2.3.2 Modelos Média Móvel (MA)

Diferentemente dos modelos AR, o MA busca representar um processo aleatório em função dos erros de predição passados. Matematicamente, tem-se

$$Y_t = c + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q}. \tag{2.3}$$

É interessante mencionar que o modelo recebe esse nome devido à interpretação de que o sinal da predição ou estimação Y_t é uma ponderação de médias dos erros passados de predição. Uma vez que as predições "caminham" em direção a um número desejado de previsões, a ponderação dos erros acompanham as predições. Assim como o AR, os modelos MA possuem uma ordem q associada à quantidade de erros passados influenciam na predição, denotados por $M(q)$.

Modelos MA são mais complexos de se obter pois os erros ε_t não são diretamente observáveis, o que impede algoritmos convencionais de regressão de serem utilizados. Para isso, realizam-se métodos numéricos que simultaneamente fazem as predições e computam o erro buscando o "encaixe" do modelo a um conjunto de dados

2.3.3 Modelos Autoregressivos com Média Móvel

A combinação dos modelos mencionados anteriormente constitui o tipo ARMA. Estes modelos são matematicamente descritos por

$$Y_t = c + \varepsilon_t + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q}. \tag{2.4}$$

A notação da ordem de dependência dos erros e predições passadas é dada por $ARMA(p,q)$. Uma ferramenta estatística clássia para representação de processos aleatórios descritos por séries temporais, modelos ARMA são bastante utilizados para adequação de séries temporais em geral por um modelo. Na Figura 8 mostra-se os resultados de um algoritmo em *Python* que encaixa um modelo ARMA(2,2) a um conjunto de dados contendo o número de passageiros aéreos de um dia em um aeroporto. O conjunto de dados foi retirado de (JBROWNLEE, 2019).

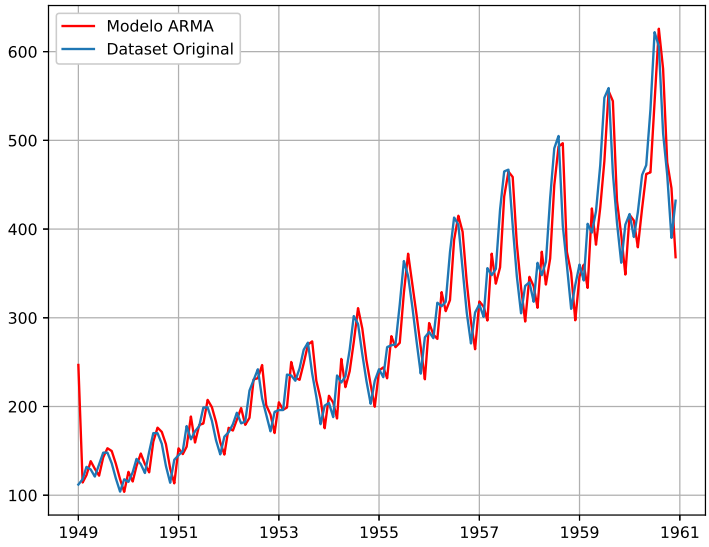

Figura 8 – Exemplo Estimação de Modelo usando ARMA

2.3.4 Modelos Autoregressivos com Média Móvel Integrado

Conforme mencionado anteriormente, modelos ARIMA são apenas um modelo ARMA aplicados em um conjunto de dados que fora derivado, ou diferenciado. A diferenciação em um conjunto de dados é feita simplesmente aplicando

$$Y'_t = Y_t - Y_{t-1} \tag{2.5}$$

para cada ponto do conjunto. Esta simples operação faz com que as tendências dos *datasets* sejam removidas, efetivamente tornando o conjunto efetivamente estacionário. A principal vantagem de abordar conjuntos estacionários é a remoção das tendências dos conjuntos de dados, viabilizando predições futuras com mais acurácia.

A notação de modelos ARIMA é então dada por $ARIMA(p,d,q$, onde p é a ordem a parcela AR, d é a ordem da diferenciação dos dados e q é a ordem da parcela MA.

Matematicamente, escreve-se

$$Y'_t = c + \varepsilon_t + \theta_1 Y'_{t-1} + \theta_2 Y'_{t-2} + \cdots + \theta_p Y'_{t-p} + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q}. \quad (2.6)$$

A determinação das ordens de *AR* e *MA* podem ser auxiliadas pelo uso das funções *Autocorrelation Function (Função de Autocorrelação)* (ACF) e *Partial Autocorrelation Function (Função de Autocorrelação Parcial)* (PACF). Estas funções retornam a influência de valores passados sobre uma observação atual ou duas observações arbitrárias. A ordem da parcela *MA* é definida pela análise da ACF, e a ordem da parcela *AR* é definida pela análise da PACF. A ordem da diferenciação depende principalmente do comportamento do conjunto de dados. Caso este apresente uma tendência linear, tal qual Figura 8, uma diferenciação de primeira ordem é suficiente. Caso exitam tendências quadráticas ou cúbicas, a diferenciação da ordem correspondente à tendência deverá ser utilizada.

2.3.4.1 Exemplo de aplicação do ARIMA em Python

Para efeito de ilustração, traz-se um passo-a-passo de como aplicar o procedimento ARIMA para a maioria dos casos em que se deseja realizar uma previsão de uma série temporal. O exemplo aplicado utiliza a linguagem *Python* e alguns módulos de terceiros gratuitos.

Tomando-se o mesmo conjunto de dados da seção 2.3.3, toma-se primeiramente o valor do logaritmo dos valores originais da planilha. Isso é feito como uma forma de normalização dos dados, com intuito de reduzir o crescimento e facilitar numericamente os algoritmos para identificação de modelos.

A seguir, calcula-se as derivadas para que a série de dados se torne estacionária, mostrados na Figura 10. Conforme mencionado neste capítulo, o uso das derivadas permite que os modelos sejam mais facilmente encaixados através do algoritmo de identificação. Este passo é apenas para visualização, pois o algoritmo ARIMA *Python* já o calcula internamente.

Com o uso do módulo *statsmodels.tsa.arima_model*, a identificação do modelo através do método ARIMA é realizada. Os resultados são mostrados na figura 11.

O código mostrado foi trazido em Apêndice A. Ao final da execução, o programa imprime na tela os coeficientes calculados, juntamente com alguns índices estatísticos. Na Figura 12 uma captura de tela dos resultados são mostrados.

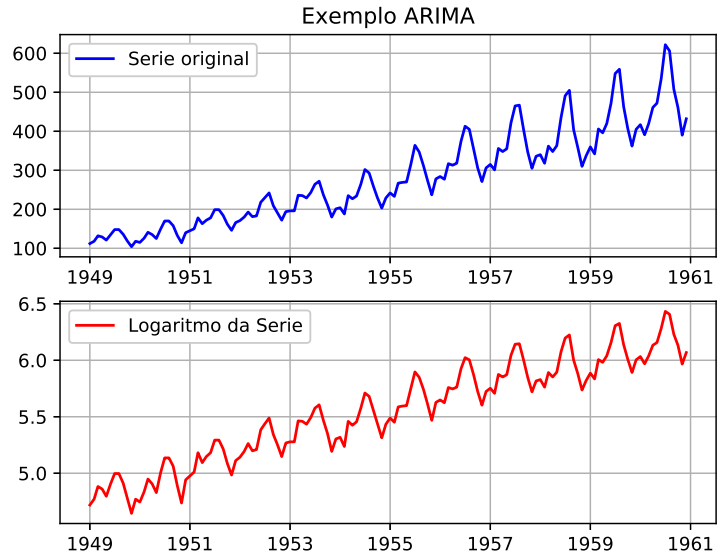

Figura 9 – Cálculo dos logaritmos para normalização dos dados

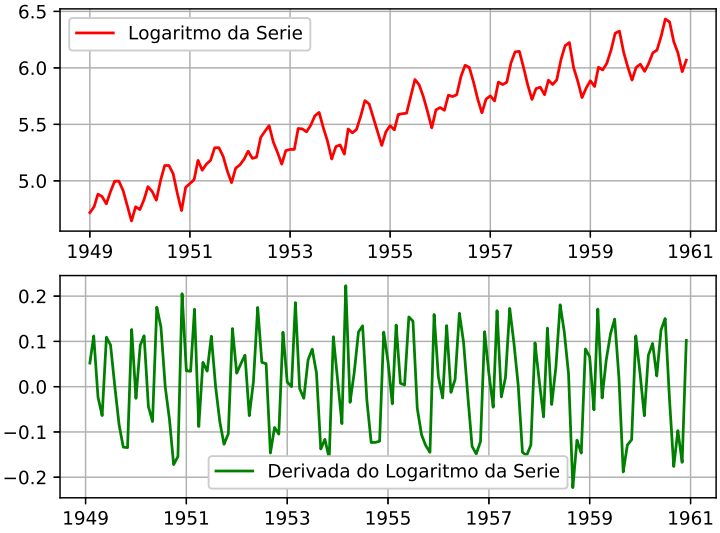

Figura 10 – Cálculo das derivadas para visualização

Os coeficientes encontrados são então dados por

$$Y'_t = 0.0096 + \varepsilon_t + 1.6293Y'_{t-1} - 0.8946Y'_{t-2} + -1.8270\varepsilon_{t-1} + 0.9245\varepsilon_{t-2}. \quad (2.7)$$

2.4 OLAP

O processamento analítico online, ou OLAP, é uma abordagem para responder rapidamente a consultas analíticas multidimensionais (*Multiple Discriminant Analysis* (MDA)) na computação e está inserido dentro de uma categoria ampla de inteligência de negócios, que

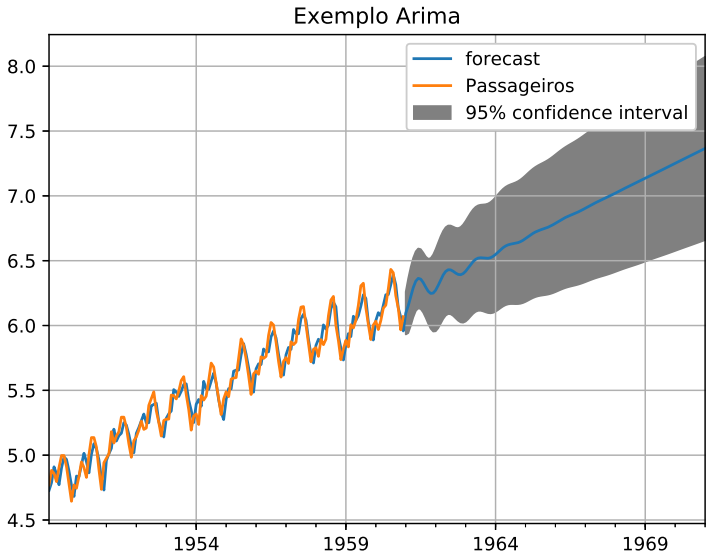

Figura 11 – Cálculo da previsão

ARIMA Model Results

Dep. Variable:	D.Passageiros	No. Observations:	143
Model:	ARIMA(2, 1, 2)	Log Likelihood	149.640
Method:	css-mle	S.D. of innovations	0.084
Date:	Sun, 01 Dec 2019	AIC	-287.281
Time:	10:33:11	BIC	-269.504
Sample:	02-01-1949	HQIC	-280.057
	- 12-01-1960		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0096	0.003	3.697	0.000	0.005	0.015
ar.L1.D.Passageiros	1.6293	0.039	41.868	0.000	1.553	1.706
ar.L2.D.Passageiros	-0.8946	0.039	-23.127	0.000	-0.970	-0.819
ma.L1.D.Passageiros	-1.8270	0.036	-51.303	0.000	-1.897	-1.757
ma.L2.D.Passageiros	0.9245	0.036	25.568	0.000	0.854	0.995

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.9106	-0.5372j	1.0573	-0.0848
AR.2	0.9106	+0.5372j	1.0573	0.0848
MA.1	0.9881	-0.3245j	1.0400	-0.0505
MA.2	0.9881	+0.3245j	1.0400	0.0505

Figura 12 – Resultados do identificador de modelo ARIMA em *Python*

também abrange bancos de dados relacionais, elaboração de relatórios e mineração de dados.

Essa análise permite que gerentes e analistas examinem e manipulem interativamente grandes quantidades de dados detalhados e consolidados de várias perspectivas. O OLAP envolve a investigação de relações complexas entre milhares ou até milhões de itens de dados armazenados em *data warehouses* [2] e *data marts* [3], e outros bancos de dados multidimensionais

[2] Um grande repositório centralizador de dados que contém informações de várias fontes dentro de uma organização, usados para orientar as decisões de negócios por meio de ferramentas de análise, relatório e mineração de dados

[3] Um subconjunto de um *data warehouse* orientado para uma linha de negócios específica que contêm repositórios de dados resumidos coletados para análise em uma seção ou unidade específica de uma organização

para descobrir padrões, tendências e condições de exceção.

Uma sessão OLAP é realizada *on-line* em tempo real, com respostas rápidas às consultas de um gerente ou analista, para que o processo analítico ou tomada de decisão seja realizado de forma interrupta. O processamento analítico online envolve várias operações analíticas básicas, incluindo consolidação, *"drill-down"* e *"slicing and dicing"*.

- **Consolidação**. A consolidação envolve a agregação de dados, que pode envolver agregações simples ou agrupamentos complexos envolvendo dados inter-relacionados. Por exemplo, o acúmulo de volume de gás em um determinado tanque de uma usina a gás podem ser acumulados a nível da usina inteira, fornecendo uma perspectiva geral para controle de inventário.
- ***Drill-down***. O OLAP também pode ir na direção inversa e exibir automaticamente dados detalhados que compreendem dados consolidados. Para esse processo, utiliza-se o termo *drill-down*. Por exemplo, o acesso aos dados gerados por máquina (*Tags*) pelos sensores que compõem a usina podem ser facilmente acessados.
- ***Slicing and dicing***. O termo *"slicing and dicing"* é coloquialmente usado na língua inglesa para se referir ao ato de "fatiar e cortar". Isto concerne à capacidade de vasculhar o banco de dados de diferentes pontos de vista. Uma fatia do banco de dados pode mostrar todos os alarmes de um tipo de tubulação na planta. Uma outra fatia pode mostrar todas as áreas de uma planta por tipo de tubulação subdivididas por tipos de alarmes. "Fatiar e cortar" em cubos é geralmente realizado ao longo de um eixo temporal para analisar tendências e encontrar padrões baseados em tempo nos dados.

2.5 ETL

Na computação, extrair, transformar, carregar (ETL) é o procedimento geral de copiar dados de uma ou mais fontes para um sistema de destino que representa os dados de forma diferente da (s) fonte (s) ou em um contexto diferente da (s) fonte (s). O processo ETL tornou-se um conceito popular na década de 1970 e é frequentemente usado em *data warehousing*.

A extração de dados envolve a extração de dados de fontes homogêneas ou heterogêneas; a transformação de dados processa os dados limpando e transformando-os em um formato/estrutura de armazenamento adequado para fins de consulta e análise; finalmente, o carregamento de dados descreve a inserção de dados no banco de dados de destino final, como

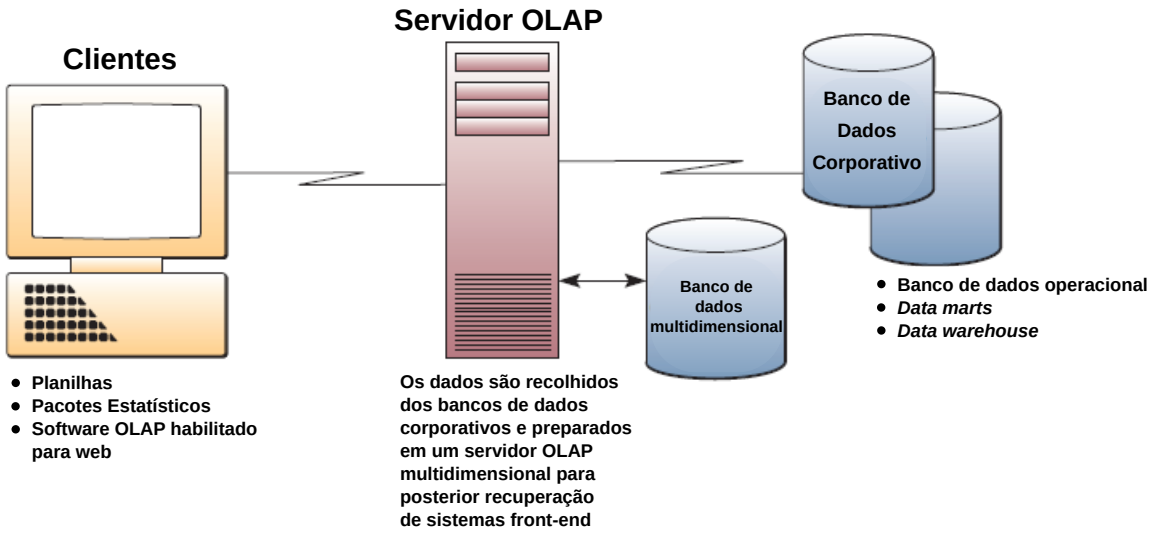

Figura 13 – OLAP pode envolver o uso de servidores especializados e bancos de dados multidimensionais. Adaptada de (O'BRIEN; MARAKAS, 2009)

um armazenamento de dados operacional, um *data mart*, *data lake* ou *data warehouse*.

2.5.1 Extrair

A primeira parte de um processo ETL envolve extrair os dados do (s) sistema (s) de origem. Em muitos casos, isso representa o aspecto mais importante do ETL, pois a extração de dados define corretamente o cenário para o sucesso dos processos subsequentes.

A maioria dos projetos de *data warehousing* combina dados de diferentes sistemas de origem. Cada sistema separado também pode usar uma organização e/ou formato de dados diferente. Formatos comuns de fonte de dados incluem bancos de dados relacionais, XML, JSON e arquivos simples.

Uma parte intrínseca da extração envolve a validação de dados para confirmar se os dados extraídos das fontes têm os valores corretos/esperados em um determinado domínio (como um padrão/padrão ou lista de valores). Se os dados falharem nas regras de validação, eles serão rejeitados total ou parcialmente. Os dados rejeitados são idealmente relatados de volta ao sistema de origem para análises adicionais para identificar e corrigir os registros incorretos.

Em geral, a fase de extração visa converter os dados em um único formato apropriado para o processamento de transformação.

2.5.2 Transformar

Uma função importante da etapa de transformação é a limpeza de dados, que visa passar apenas dados "adequados"para o destino. O desafio quando diferentes sistemas interagem está na interface e comunicação dos sistemas relevantes. Os conjuntos de caracteres que podem estar disponíveis em um sistema podem não estar em outros.

- Selecionar apenas determinadas colunas para carregar: (ou selecionar colunas nulas para não carregar);
- Traduzir valores codificados: (por exemplo, se o sistema de origem codificar masculino como "1"e feminino como "2", mas o banco codificar masculino como "M"e feminino como "F");
- Derivar um novo valor calculado: (por exemplo, quantidade_da venda = qty * preço_ unitário)
- Classificar ou ordenar os dados com base em uma lista de colunas para melhorar o desempenho da pesquisa;
- Transposição ou rotação (transformar várias colunas em várias linhas ou vice-versa);

2.5.3 Carregar

Essa é a última etapa do ETL, quando os dados serão carregados no destino final, que pode ser qualquer armazenamento de dados, incluindo um arquivo simples delimitado ou um *data warehouse*. É um processo que pode variar a depender das necessidades da organização.

Em alguns casos, os *data warehouse* podem sobrescrever informações já existentes por informações cumulativas. Essa atualização dos dados extraídos pode ser feita com frequência diária, semanal ou mensal. Outras partes (ou até mesmo outros *data warehouse*) desse mesmo *data warehouse* podem adicionar novos dados de uma forma histórica em intervalos regulares - por exemplo, a cada hora.

Para entender isso, considere um *data warehouse* necessário para manter registro de quebras de equipamentos do ano passado. Esse *data warehouse* sobrescreve todos os dados anteriores a um ano pelos dados mais recentes. No entanto, a entrada de dados para qualquer janela de um ano é feita de maneira histórica. O momento e o escopo para sobrescrever ou acrescentar são opções de design estratégico, dependendo do tempo disponível e das necessidades da empresa. Em sistemas mais complexos geralmente se mantém um histórico e um registro de

todas as alterações feitas no *data warehouse*.

À medida que a fase de carregamento interage com um banco de dados, as restrições definidas no esquema do banco se aplicam (por exemplo, exclusividade, integridade, referencial, campos obrigatórios), o que também contribui para o desempenho geral da qualidade dos dados do processo ETL.

3 METODOLOGIA

Para alcançar os objetivos do presente trabalho, a metodologia mostrada na Figura 14 foi utilizada.

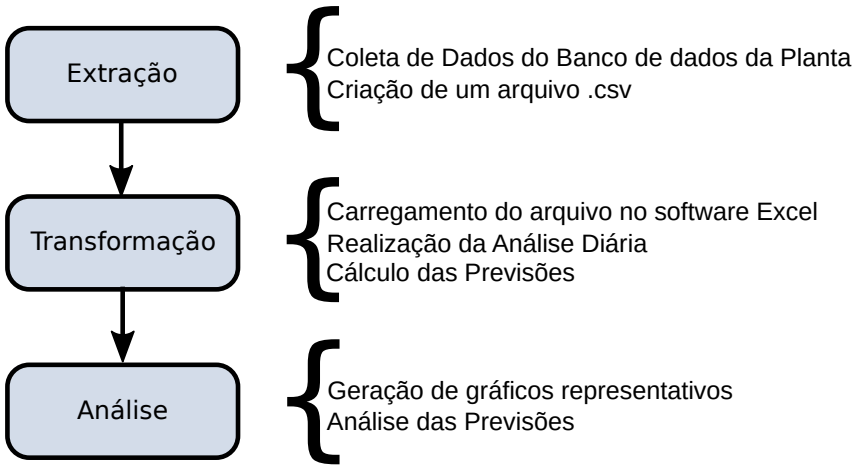

Figura 14 – Fluxograma da metodologia aplicada.

Os detalhes de cada etapa são discorridos detalhadamente neste capítulo. Os conceitos estudados no Capítulo 2 serão utilizados para explicação dos procedimentos e análise realizadas sobre o conjunto de dados em estudo.

3.1 Recuperação de dados dos sensores (Extração)

Os dados são continuamente recuperados dos sensores instalados em gaseodutos de uma indústria petroquímica, coletados de um historiador. Essa recuperação contém informações a partir de setembro 2015 até janeiro de 2019 e foram transferidos para um arquivo do formato *.csv* [1].

As *tags*[2] desses sensores contém informações de *timestamp*[3], valor absoluto da medição e um nome. A coleta foi feita para os seguintes medidas: temperatura, pressão, vazão e massa acumulada. Cada uma dessas *tags* possui unidades de engenharia específicas. A temperatura é medida em graus Celsius (^oC), a pressão é medida em kgf/cm^2, vazão em kg/h e

[1] *Comma Separated Values* - Valores separados por vírgula

[2] "O termo *Tag* refere-se a qualquer variável na base de dados, com seus atributos em tempo real ou com seus dados históricos associados." (SPINENGENHARIA, 2019)

[3] Os dados são "carimbados" com o tempo, contendo o ano, mês, dia, hora, minuto e segundo em que a medição ocorreu

o a massa em toneladas (t).

TAG	Número de Linhas (2014-2019)
PRES_ETEN_CLIENTE1_PVC	2,096,204
PRES_ETEN_CLIENTE2	79,315
TEMP_ETEN_CLIENTE1_PVC	2,469,689
TEMP_ETEN_CLIENTE2	181,144
TOT_ETEN_CLIENTE1_PVC_MP	2,469,678
TOT_ETEN_CLIENTE1_PVC_MS	2,469,678
TOT_ETEN_CLIENTE2_MP	2,469,678
TOT_ETEN_CLIENTE2_MS	2,469,678
VAZCMP_ETEN_CLIENTE1_PVC_MP	2,468,644
VAZMP_ETEN_CLIENTE2	2,469,689
VAZMS_ETEN_CLIENTE2	2,469,689
Total	**22,113,086**

Tabela 1 – Quantidade de dados extraídos inicialmente

Quando se utiliza o termo "linha" no contexto aqui trazido, refere-se a uma linha de um arquivo *.csv*. Esta serve como unidade para cálculo da complexidade do problema. Cada linha é processada individualmente nas ferramentas de ciência de dados e softwares convencionais. Neste trabalho, cada linha está associada a três colunas (Nome/Tag, Valor de leitura, Timestamp). Em termos mais grosseiros, são então $22,113,086\ x\ 3 =$ **66,339,258** células a serem processadas.

3.2 Manipulação dos dados (Transformação)

Para o completo carregamento dos dados do arquivo *.csv* extraídos do servidor, foi necessário utilizar uma ferramenta do Excel chamada *Power Query*[4]. Sem essa ferramenta, o Excel frequentemente ocasiona travamento total na máquina devido ao uso de excessiva memória no computador quando *.csv* é carregado normalmente. Além disso, o suplemento permite o tratamento dos dados que estavam em sua forma bruta e precisavam de alterações nas suas características iniciais. Data e hora foram modificados para somente data e o valor absoluto de leitura da *tag* foi alterado de número inteiro para decimal.

Uma vez carregados, foi aplicada uma análise diária sobre dos dados através do *Power Query*. Essa análise consiste em calcular médias diárias a partir dos dados originais, coletados a cada minuto, reduzindo de forma considerável a quantidade de dados para se trabalhar

[4] O Power Query é um suplemento gratuito para Extrair, Transformar e Carregar dados de diversas fontes de dados e criar relatórios. É a primeira etapa de um processo de BI, extrair os dados e tratar os dados.

sem provocar perda de representatividade. A tabela 2 traz o número de linhas a serem processadas com os dados reduzidos.

TAG	Número de Linhas (2014-2019)
PRES_ETEN_CLIENTE1_PVC	1,458
PRES_ETEN_CLIENTE2	1,458
TEMP_ETEN_CLIENTE1_PVC	2,083
TEMP_ETEN_CLIENTE2	2,083
TOT_ETEN_CLIENTE1_PVC_MP	2,083
TOT_ETEN_CLIENTE1_PVC_MS	2,083
TOT_ETEN_CLIENTE2_MP	2,083
TOT_ETEN_CLIENTE2_MS	2,083
VAZCMP_ETEN_CLIENTE1_PVC_MP	2,083
VAZMP_ETEN_CLIENTE2	2,083
VAZMS_ETEN_CLIENTE2	2,083
Total	**25,829**

Tabela 2 – Quantidade de dados após transformação da análise diária

Segundo a lógica trazida em 3.1, calcula-se então $25,829\ x\ 3 = \mathbf{77487}$ células a serem processadas. É válido ressaltar que tal redução implica também na redução dos dados ocupados em memória de computador. Foi calculado que os dados originais das amostras a cada minuto entre 2014 e 2019 totalizaram uma ocupação de $1,2GB$ de dados na máquina. Após a compressão, esse valor caiu para $416KB$. A tabela 3 mostra os valores transformados e a razão de compressão.

Pré-Processamento	Pós-Processamento	Razão de Compressão
66,339,258 Células	77,487 Células	1:856
1,2 GB	416 KB	1:2885

Tabela 3 – Resultados da transformação dos dados após análise diária

3.3 Análise dos dados (Carregamento)

De posse do novo conjunto de dados, utiliza-se o software *Microsoft Power BI* para produção das análises desejadas. Primeiramente, gera-se um gráfico contendo as médias diárias computadas no passo anterior. A visualização do gráfico das médias diárias auxilia na identificação de possíveis falhas que tenham durante o período de amostragem.

O software possui ferramentas para fabricação de previsões baseados em modelos ARIMA sobre a análise diária. Outras propriedades estatísticas, como tendência e média são

também geradas.

Gerou-se, então, um total de 20 gráficos distribuídos em 7 figuras da seguinte forma

- Figura 16 → Pressão Cliente 1 com Previsão
- Figura 17 → Pressão Cliente 2 com Previsão
- Figura 18 → Temperatura Clientes 1 e 2 com Previsão
- Figura 19 → Massa de Total Cliente 1 (Sensores Primário e Secundário) com Previsão
- Figura 20 → Massa de Total Cliente 2 (Sensores Primário e Secundário) com Previsão
- Figura 21 → Vazão Cliente 1 (Sensor Primário) com Previsão
- Figura 22 → Vazão Cliente 2 (Sensores Primário e Secundário) com Previsão

No capítulo 4 as Figuras são analisadas e comentadas. Para ilustração do procedimento de forma mais prática, traz-se um diagrama na Figura 15. A curva tracejada indica uma opção alternativa e gratuita (Python) ao uso de softwares da Microsoft ™.

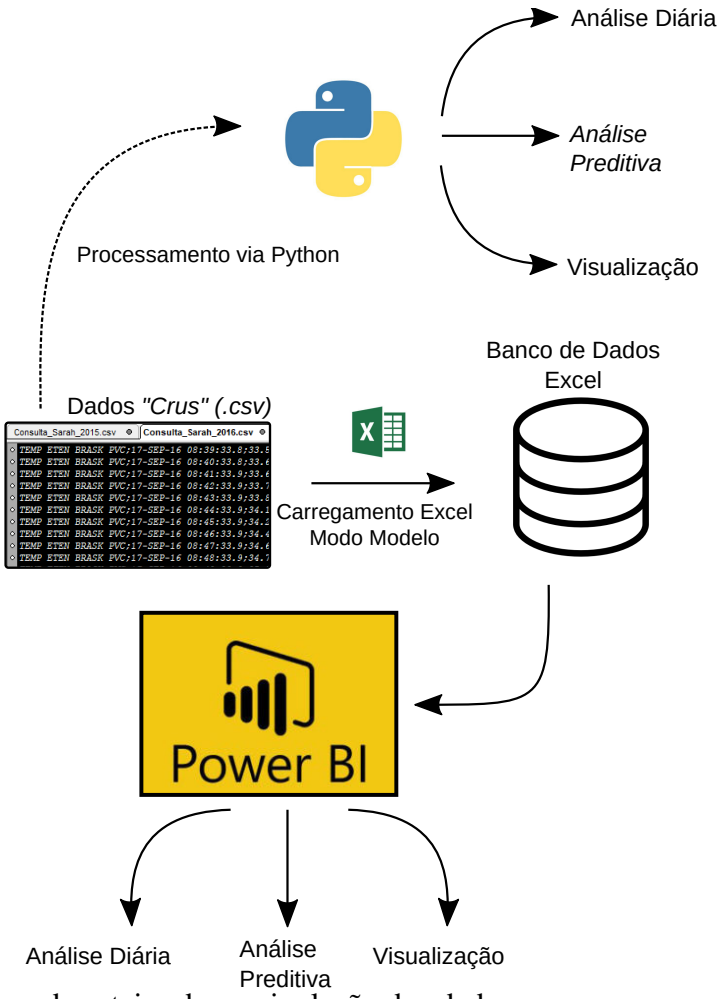

Figura 15 – Diagrama do roteiro de manipulação dos dados

4 RESULTADOS

Nesta seção são apresentados e discutidos os resultados obtidos após a aplicação da metodologia.

4.1 Resultados das Análises Individuais

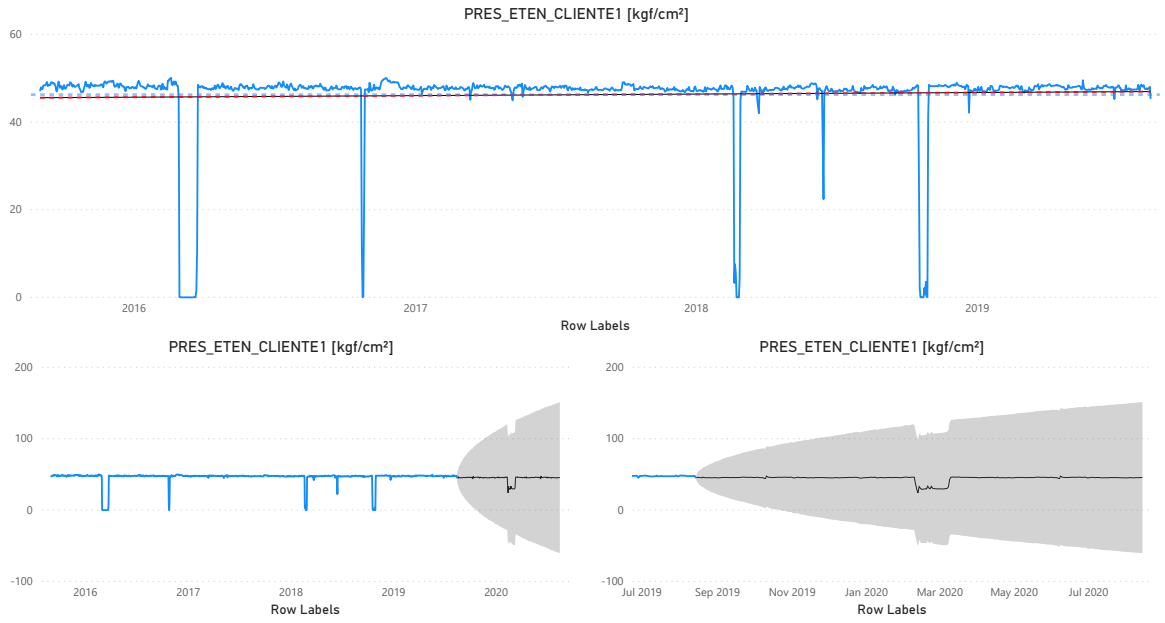

Figura 16 – Pressão Cliente 1 com Previsão

Na Figura 16, pode-se perceber claramente a presença de uma possível falta no sensor de pressão do cliente 1 por volta de março de 2020. Vale ressaltar que esta estimação matemática leva em consideração todo o histórico de 5 anos de dados. Uma manutenção pode ser então previamente agendada para que a falta não seja ocasionada.

Na Figura 17 estima-se que falhas ocorrerão em Maio de 2020 e que uma teria ocorrido em outubro deste ano.

As temperaturas dos gráficos da Figura 18 mostraram um comportamento interessante. Nos sensores de ambos os clientes, os sinais apresentaram uma periodicidade, cujos mínimos ocorrem entre os meses de junho e julho e os máximos no início de cada ano. Este fenômeno provavelmente ocorre devido à influência da estação do ano na temperatura na localidade da instalação.

Na Figura 19 tem-se o acúmulo da massa do cliente 1. O sensor apresentou pequenas

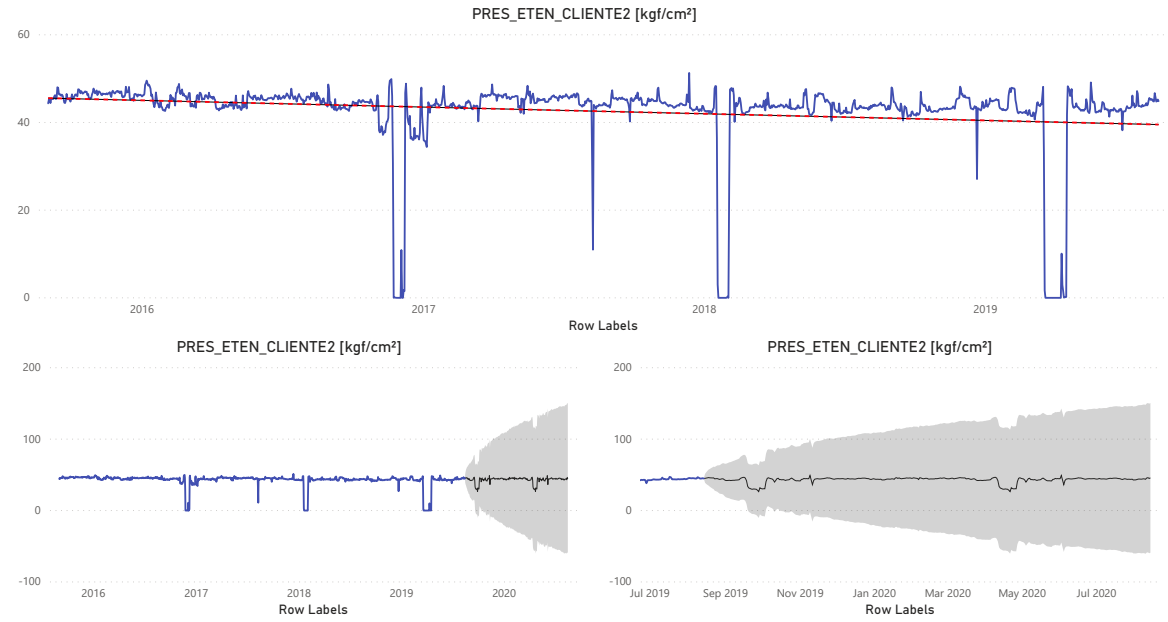

Figura 17 – Pressão Cliente 2 com Previsão

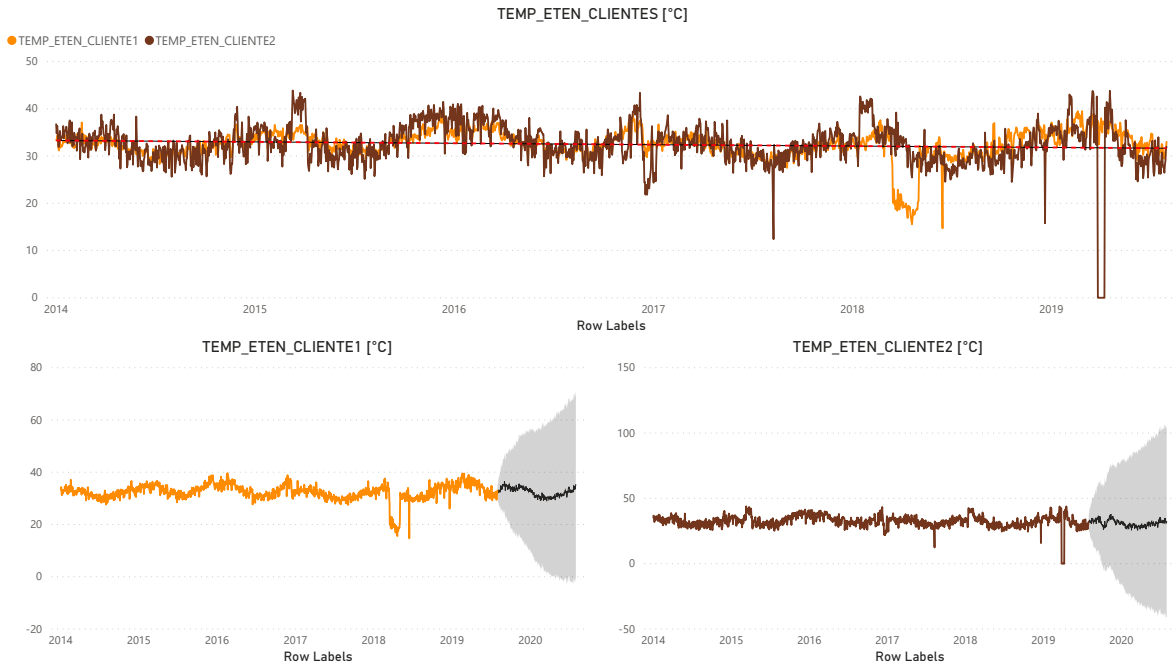

Figura 18 – Temperatura Clientes 1 e 2 com Previsão

falhas de leitura em abril de 2018, mas as leituras rapidamente se recompuseram. O Previsão desta curva pode ser usada para previsão do controle de estoque.

Semelhante à Figura 19, a Figura 20 mostra a mesma medida no cliente 2. Contudo, assume-se que as leituras do sensor zeradas em agosto de 2017. Este comportamento, porém, não prejudicou a análise Previsão.

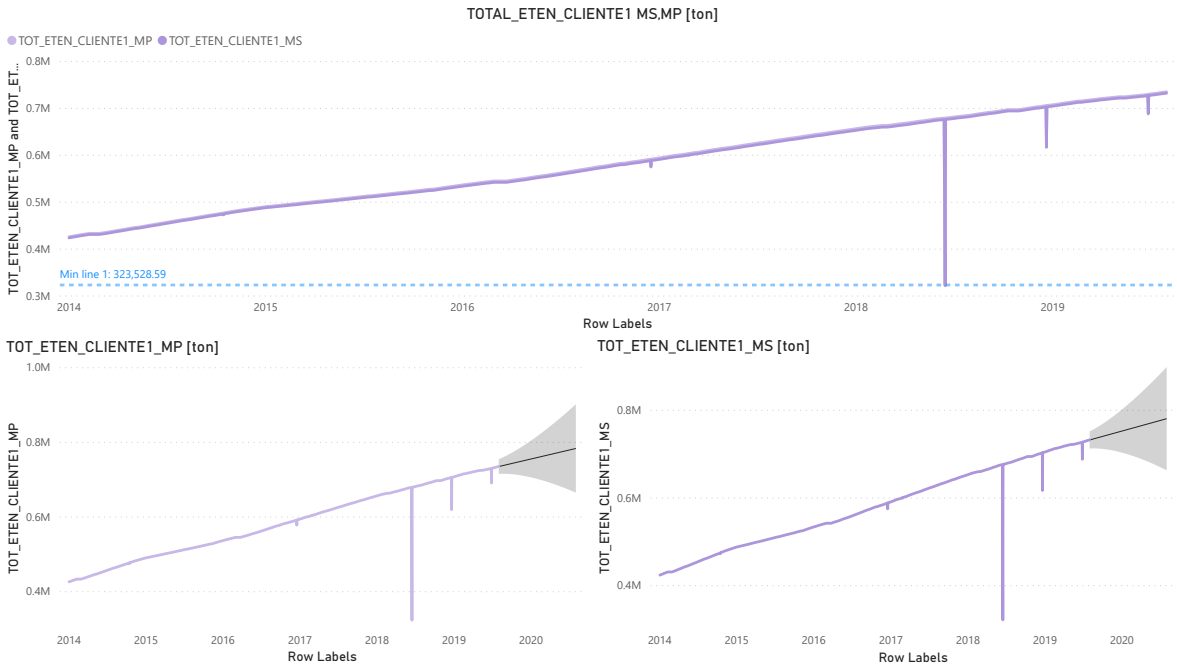

Figura 19 – Massa de Total Cliente 1 (Sensores Primário e Secundário) com Previsão

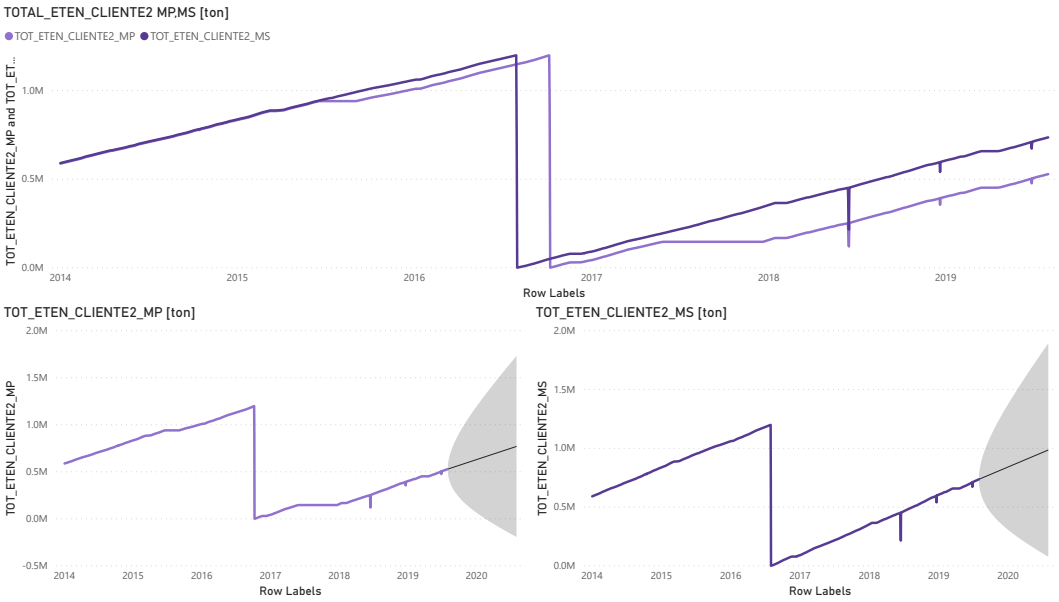

Figura 20 – Massa de Total Cliente 2 (Sensores Primário e Secundário) com Previsão

O sensor de vazão da Figura 21 mostrou um comportamento altamente oscilatório, repleto de picos em zero. As previsões deste sensor se mantiveram em sua média, sem falhas.

Pelos gráficos da Figura 22, percebe-se também um comportamento oscilatório das leituras de vazão. Neste caso, o cliente 2 forneceu leitura de um sensor primário e um secundário para redundância. Os sensores apresentaram leituras semelhantes, com algumas falhas em ambos os sensores. As falhas de um sensor foram compensadas pelo outro.

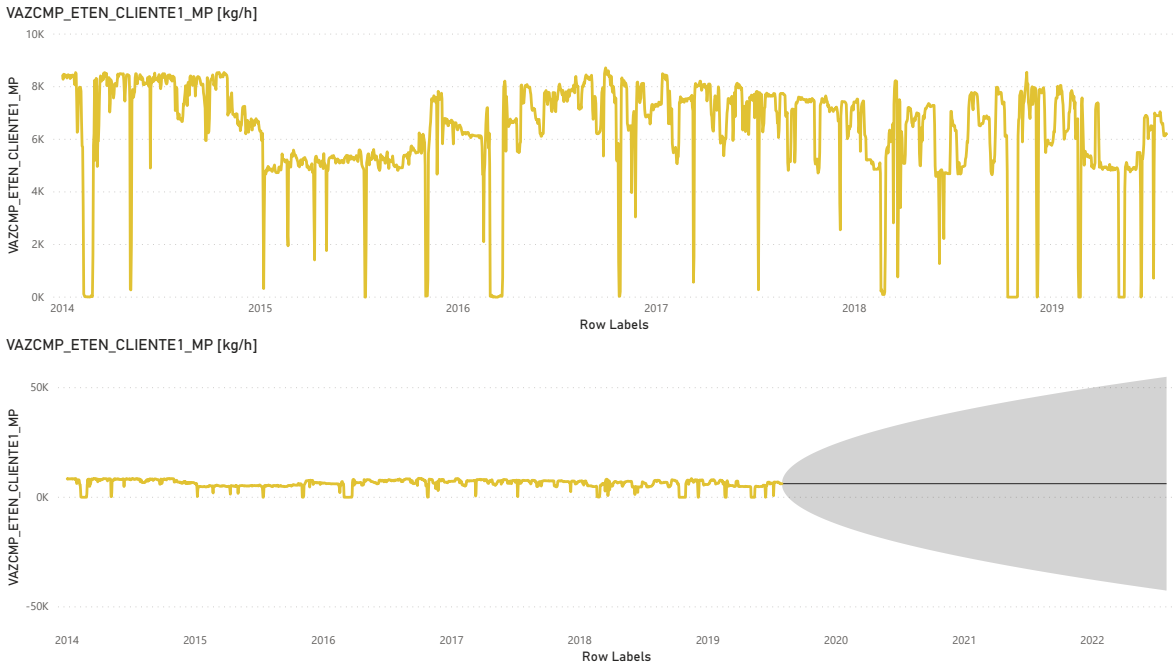

Figura 21 – Vazão Cliente 1 (Sensor Primário) com Previsão

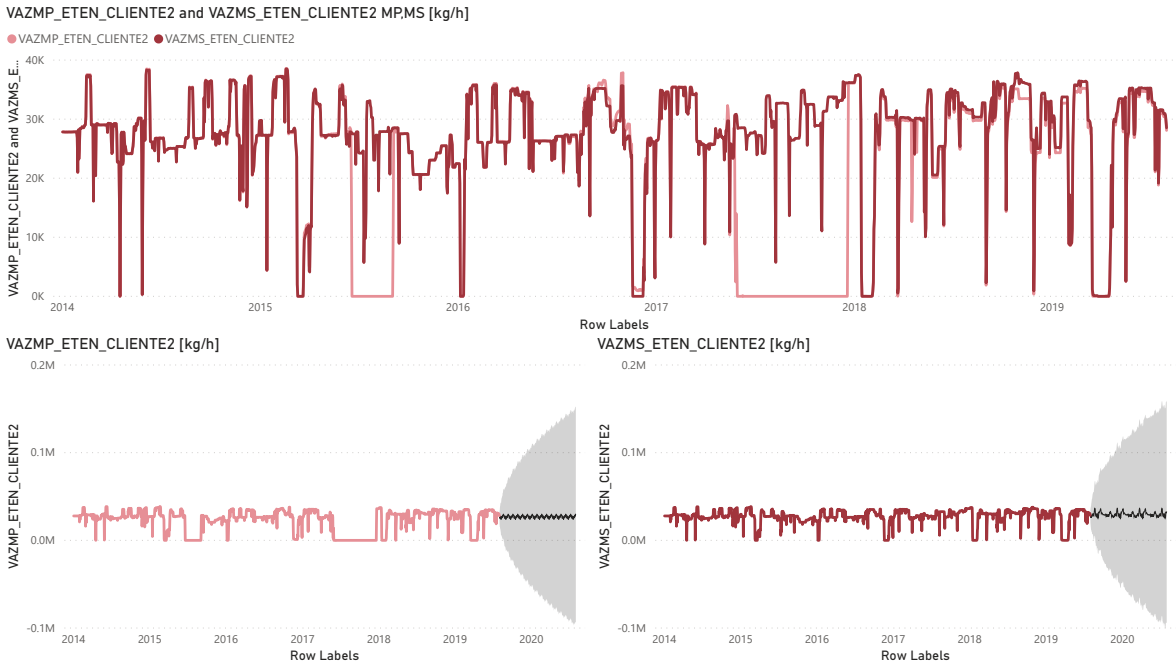

Figura 22 – Vazão Cliente 2 (Sensores Primário e Secundário) com Previsão

5 CONCLUSÕES

Esse trabalho foi criado com o objetivo principal de extrair, organizar e interpretar através de análises dos dados adquiridos, pelo uso do método ARIMA, uma previsão temporal para construção de uma análise preditiva.

Os assuntos relacionados à engenharia metalúrgica e engenharia de software utilizados na elaboração deste trabalho não são limitados à aplicação desenvolvida. O uso dos conceitos aqui utilizados podem ser aplicados aos mais diversos setores da indústria, principalmente nas que utilizam monitoramento por automação.

A quantidade de dados coletadas pelas organizações tem crescido continuamente, e, por isso, existe um aumento na necessidade de estudo na área de Ciência de Dados. Esta tendência foi o principal agente motivador para a redação deste trabalho, no qual se buscou e se aplicou com sucesso teorias e ferramentas desta ciência a um conjunto de dados industriais. Os dados foram extraídos, processados e reduzidos ao ponto de possibilitar análises preditivas de falhas para os sensores.

5.1 Recomendação para Trabalhos Futuros

A partir dos resultados obtidos neste trabalho, percebe-se a necessidade de se aprofundar em alguns temas, os quais são recomendados para estudos em trabalhos futuros. São eles:

- Análise de correlação entre medidas – Foi observado que as medições possuem uma inter-dependência (sensor primário com secundário, ou vazão com pressão). Um trabalho futuro seria quantificar essa relação e realizar previsões mais precisas
- Aplicação de análises preditivas diárias para turnos e/ou jornadas de trabalho na indústria

REFERÊNCIAS

123RF. 2019. Disponível em: <https://www.123rf.com/photo_111597426_stock-vector-big-data-science-vector-illustration-machine-learning-algorithm-for-information-filter-and-anaytic-i.html>.

BRULE, M. R. The data reservoir: How big data technologies advance data management and analytics in e&p. Society of Petroleum Engineers, 2015.

CIELEN, D.; MEYSMAN, A.; ALI, M. **Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools**. 1st. ed. Greenwich, CT, USA: Manning Publications Co., 2016. ISBN 1633430030, 9781633430037.

FEBLOWITZ, J. Analytics in oil and gas: The big deal about big data. Society of Petroleum Engineers, 2013.

JBROWNLEE. 2019. Disponível em: <https://github.com/jbrownlee/Datasets>.

Jiang ChunLei; Wang Yuan. The research of natural gas pipeline leak detection based on adaptive filter technology. In: **Proceedings of 2013 2nd International Conference on Measurement, Information and Control**. [S.l.: s.n.], 2013. v. 02, p. 1229–1233. ISSN null.

LANEY, D. **Defining and Differentiating the Role of the Data Scientist**. 2012. Disponível em: <https://blogs.gartner.com/doug-laney/defining-and-differentiating-the-role-of-the-data-scientist/>.

MACROTRENDS. **Crude Oil Prices**. 2019. Disponível em: <https://www.macrotrends.net/1369/crude-oil-price-history-chart>.

MEHTA. **Tapping the Value From Big Data Analytics**. 2016. Disponível em: <https://pubs.spe.org/en/jpt/jpt-article-detail/?art=2494>.

O'BRIEN, J.; MARAKAS, G. **Management Information Systems**. 9. ed. New York, NY, USA: McGraw-Hill, Inc., 2009. ISBN 0073376760, 9780073376769.

SHUMWAY, R. H.; STOFFER, D. S. **Time Series Analysis and Its Applications (Springer Texts in Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2005. ISBN 0387989501.

SPINENGENHARIA. 2019. Disponível em: <http://www.spinengenharia.com.br/help/an-2014/ActionNETUG/TagsAseetsAndTemplates/TagsAseetsAndTemplates.htm>.

TRIFU, M. **Big Data: present and future**. 2014. Disponível em: <https://pdfs.semanticscholar.org/930a/249978326a0ab83ec446b1f42fbb9e9b6f49.pdf>.

Upstream? Midstream? Downstream? What's the Diffence? 2017. Disponível em: <https://energyhq.com/2017/04/upstream-midstream-downstream-whats-the-difference/>.

Wei, L.; Laibin, Z.; Yingchun, Y. Data mining technology based leak detection method for crude oil pipeline. In: **2009 WRI World Congress on Computer Science and Information Engineering**. [S.l.: s.n.], 2009. v. 3, p. 656–660. ISSN null.

APÊNDICE A – CÓDIGOS-FONTES UTILIZADOS EM PYTHON PARA DEMONSTRAÇÃO DO MODELO ARIMA

```
1  import numpy as np
2  import pandas as pd
3  from matplotlib import pyplot as plt
4  from statsmodels.tsa.stattools import adfuller
5  from statsmodels.tsa.seasonal import seasonal_decompose
6  from statsmodels.tsa.arima_model import ARIMA
7  from pandas.plotting import register_matplotlib_converters
8  register_matplotlib_converters()
9
10 df = pd.read_csv("airline-passengers.csv", parse_dates = ["
       Month"], index_col = ["Month"]) #Carrega arquivo .csv do
        repositorio de Data Science
11
12 df_log = np.log(df) #calcula logaritmo para melhorar o
       shape
13 # Plots
14 plt.subplot(2,1,1)
15 plt.plot(df, label="Serie original", color="blue")
16 plt.title("Exemplo ARIMA")
17 plt.grid()
18 plt.legend()
19 plt.subplot(2,1,2)
20 plt.plot(df_log, label="Logaritmo da Serie", color="red")
21 plt.grid()
22 plt.legend()
23 plt.ion();
24 plt.show();
25
```

```
26 df_log_shift = df_log - df_log.shift() #Calcula derivada
      discreta
27 # Plots
28 plt.figure()
29 plt.title("Exemplo ARIMA")
30 plt.subplot(2,1,1)
31 plt.plot(df_log, label="Logaritmo da Serie", color="red")
32 plt.grid()
33 plt.legend()
34 plt.subplot(2,1,2)
35 plt.plot(df_log_shift, label="Derivada do Logaritmo da
      Serie", color="green")
36 plt.grid()
37 plt.legend()
38 plt.show()
39
40
41 model = ARIMA(df_log, order=(2,1,2)) #aplica modelo arima
      (2,1,2)
42 results = model.fit(disp=-1) #calcula o modelo
43
44
45 results.plot_predict(1,264) #calcula predicoes de 264
      amotras (144 presentes, 120 futuras)
46
47 #Plots
48 plt.grid()
49 plt.ioff()
50 plt.title("Exemplo Arima")
51 print(results.summary())
52 plt.show()
```