CENTRO PAULA SOUZA
FACULDADE DE TECNOLOGIA DE FRANCA
“Dr. THOMAZ NOVELINO”

TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

GUILHERME HENRIQUE FERREIRA

LIMPEZA DE DADOS UTILIZANDO FERRAMENTAS POWER BI E TABLEAU

FRANCA/SP
2020

GUILHERME HENRIQUE FERREIRA

LIMPEZA DE DADOS UTILIZANDO FERRAMENTAS POWER BI E TABLEAU

Trabalho de Graduação apresentado à Faculdade de Tecnologia de Franca - “Dr. Thomaz Novelino”, como parte dos requisitos obrigatórios para obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador: Prof. Me. Claudio Eduardo Paiva.

FRANCA/SP

2020

Ficha catalográfica

F237l Ferreira, Guilherme Henrique
Limpeza de dados utilizando ferramentas Power Bi e Tableau / Guilherme Henrique Ferreira / [s.n], 2020

37 f.; 30 cm; il
Trabalho de Graduação (Curso Superior de Análise e Desenvolvimento de Sistemas) Fatec - Faculdade de Tecnologia "Dr. Thomaz Novelino".
Orientador: Prof.Me. Claudio Eduardo Paiva

1. Análise de dados. 2.Base de dados. 3. ETL. 4.Limpeza de dados. I. Autor. II. Título.

CDD – 005.74

GUILHERME HENRIQUE FERREIRA

LIMPEZA DE DADOS UTILIZANDO FERRAMENTAS POWER BI E TABLEAU

Trabalho de Graduação apresentado à Faculdade de Tecnologia de Franca – “Dr. Thomaz Novelino”, como parte dos requisitos obrigatórios para obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Trabalho avaliado e aprovado pela seguinte Banca Examinadora:

Orientador(a)..... : _____
Nome..... : Prof. Me. Cláudio Eduardo Paiva
Instituição ..... : Faculdade de Tecnologia de Franca – “Dr. Thomaz Novelino”

Examinador(a) 1 : _____
Nome..... : Profa. Drª Jaqueline Brigladori Pugliesi
Instituição ..... : Faculdade de Tecnologia de Franca - "Dr. Thomaz Novelino"

Examinador(a) 2 : _____
Nome..... : Profa. Me. Maria Luísa Cervi Uzun
Instituição ..... : Faculdade de Tecnologia de Franca - "Dr. Thomaz Novelino"

Franca, 04 de dezembro de 2020

Dedico o presente Trabalho de Graduação primeiramente a Deus, aos meus familiares, em especial à minha mãe Lourdes, a meu Pai Carlos e ao meu Irmão Vitor, a minha Namorada Karlla e aos Professores do Curso.

Sem dados você é só uma pessoa com uma opinião.

William Edward Deming

RESUMO

A limpeza de dados é uma etapa essencial do processo de análise de dados e implica na execução de metodologias e práticas de modificação de bases de dados originais a fim de melhorar sua qualidade. Nos dias atuais, as atividades de estudo de dados coletados de diversas fontes vêm aumentando em empresas públicas e privadas que têm, cada vez mais, visto a importância de se avaliar e analisar dados coletados nas suas tomadas de decisões. Muitas vezes, os *softwares* utilizados nas empresas não foram preparados para fornecer dados para análises e geram dados com baixa qualidade, com anomalias e erros, o que pode exigir um processo de limpeza de dados mais intenso, com grandes modificações da massa de dados original. Uma razão para estudar ferramentas para limpeza de dados, ocorre do fato de que, quando os dados originais possuem vários tipos e formatos diferentes, muitas vezes também eles apresentam redundâncias e inconsistências, o que pode prejudicar sobremaneira a interpretação dos resultados das análises. Esse trabalho propõe estudar, por meio de testes práticos, algumas das principais ferramentas disponíveis no mercado e suas técnicas para limpeza de dados no processo de extração, transformação e carregamento de dados. Os testes permitiram comparar situações em que dada ferramenta se torna mais eficaz e ou apropriada em relação às outras e tiveram seus resultados documentados.

Palavras-chave: Análise de dados. Base de dados. ETL. Limpeza de dados.

ABSTRACT

Cleaning data is part of the data analysis process. This is an essential step and implies the implementation of methodologies and practices for modifying original databases, in order to improve their quality for the creation of analyzes. Nowadays, the activities of studying data collected from different sources have been increasing in public and private companies, which have increasingly seen the importance of evaluating and analyzing data collected in their process making decision. Often, the software used in companies is not prepared to provide data for analysis and generates data with low quality, with a large number of anomalies and errors and which requires a more intense data cleaning process, making changes to the original data mass. One reason for studying tools for data cleaning is that, when the original data has several different types and formats, they often also have redundancies and inconsistencies, which can greatly impair the interpretation of the analysis results. This study proposes analyses of tools and techniques of working in data cleaning in the ETL process, analyzing which process becomes easier and more effective, thus we had the quality conclusion that all results obtaining more significant difference in the speed and quantity of processes required in each process, and showing that each study requires an analysis of the problem in order to run cleaning processes.

Keywords: Data analysis. Data base. ETL. Data Cleaning.

LISTA DE FIGURAS

LISTA DE SIGLAS

BI – *Business Intelligence*
CPU – *Central Processing Unit*
CSV –*Comma Separated Values*
DW – *Data Warehouse*
ETL –*Extract Transform Load*
GB – Gigabyte
IBGE – Instituto Brasileiro de Geografia e Estatística
IBM – *International Business Machines Corporation*
ID – Identificação
JSON – *JavaScript Object Notation*
KDD –*Knowledge-Discovery in Databases*
PDF –*Portable Document Format*
RAM – *Random access memory*
SGBD – *Sistema de gerenciamento de banco de dados*
SQL –*Structure query language*
XML – *Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Limpeza de dados é um dos processos que compõe a etapa de extração, transformação e carregamento dos dados (*Extract, Transform and Load* - ETL) que, por sua vez, faz parte de um processo maior de análise de dados. ETL implica buscar meios de oferecer melhorias na qualidade de bases de dados que são usadas para extração de novas informações na descoberta de conhecimento.

Esse trabalho tem o objetivo de estudar algumas ferramentas para limpeza de dados disponíveis no mercado e entender como são as técnicas empregadas por elas nesta etapa da ETL. Para isto, foram criados cenários de testes e aplicados às ferramentas estudadas a fim de se apontar situações de preferência de uso de ferramentas. Também fez parte do estudo, entender os conceitos de descoberta de conhecimento em bases de dados, as formas de se trabalhar com *Data Warehouse*, *Business Intelligence* e a própria ETL.

Atualmente, muitas empresas utilizam a análise de dados para apoiar a tomada de decisão nos seus mais variados setores, tanto na parte estratégica quanto na parte operacional e, problemas relacionados à qualidade dos dados e que são frequentemente encontrados por profissionais de tecnologia da informação, pode ser abrandado com as técnicas de limpeza de dados.

Observa-se uma grande perda de informações em análises de conjuntos de dados devido a problemas decorrentes de coleta de dados ineficientes, dados faltantes, dados redundantes e erros de digitação em campos abertos. Estas informações irregulares dificultam a obtenção de informações com qualidade e impactam nos resultados dos estudos, prejudicando sua confiabilidade e precisão.

A Limpeza de Dados se torna importante nesse cenário pois ajuda a eliminar ou minimizar problemas de qualidades de dados gerados por coletas ineficientes. Estes processos de tratamento dos dados também podem adequá-los para um estudo específico, permitindo que a base de dados original não seja alterada, ou seja, fica acessível a outros estudos.

Modificar os dados por meio da limpeza ajuda que a informação estudada tenha melhoria na qualidade e é importante ressaltar que o profissional que executa tal limpeza deve ter bem claro quais informações são esperadas como resultado do estudo e ter conhecimento sobre sua base de dados conter os dados específicos necessários. Quando o profissional não sabe um desses quesitos, ele pode eliminar

ou modificar dados para um padrão que não é ideal, causando recuperação de informação errada ou imprecisa, consequentemente levando à tomada de decisões incorretas.

Este trabalho está dividido em 4 capítulos. No capítulo 1 é apresentada a introdução sobre o estudo de ETL, sua justificativa e os conceitos fundamentais da limpeza de dados. No capítulo 2 estão informações sobre problemas frequentemente encontrados nos processos de análise de dados. O capítulo 3 apresenta alguns *softwares* que possuem funcionalidades de ETL e estão em uso no mercado. No capítulo 4 estão as etapas e resultados de cenários de testes criados com uso das ferramentas estudadas. Por fim, as considerações finais do trabalho.

2 REFERENCIAL BIBLIOGRÁFICO

Este capítulo apresenta os principais conceitos utilizados neste trabalho, por meio de estudo do referencial bibliográfico.

2.1 PROCESSO DE DESCOBERTA DE CONHECIMENTO EM DADOS

De acordo com Silva (2004), encontrar semelhanças e padrões em conjuntos de dados e agrupá-los de forma que pessoas possam entender mais facilmente quais informações eles expressam é um dos principais objetivos da descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Databases* – KDD).

KDD envolve uma sequência de várias etapas e se preocupa em melhorar a qualidade dos dados estudados, transformando-os de dados de baixa qualidade em dados de alta qualidade a fim de favorecer a criação de novas informações e gerar conhecimento (GOEBEL; GRUENWALD, 1999).

Segundo Thomé (2008), KDD pode ser considerado uma forma de se transformar conhecimento algo que está disperso e inexplorado em diversas bases de dados por meio da construção de padrões. Para o autor o uso de KDD pode melhorar processos e diminuir custos, além de aumentar lucros das empresas.

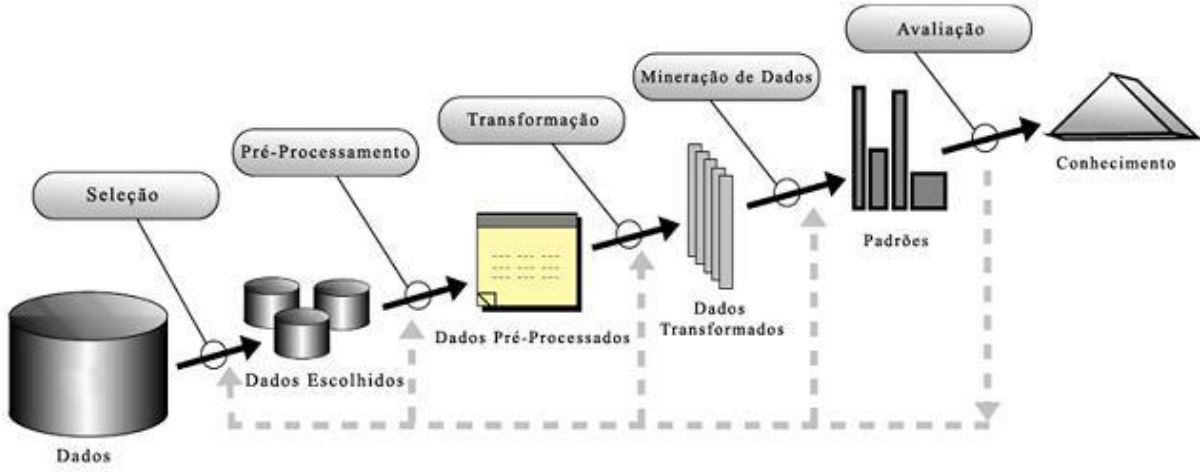
Para Bueno e Viana (2012), KDD tem como finalidade encontrar e interpretar informações de valor a partir da análise dos padrões descobertos nos dados. Os autores afirmam que estas informações ajudam as empresas a ganharem vantagem competitiva, diminuir gastos e melhorar os processos de *marketing* na busca de clientes direcionados. Segundo os autores, o KDD pode ser dividido nas etapas:

- Seleção
- Pré-Processamento
- Transformação
- Mineração de dados
- Avaliação

De acordo com Bueno e Viana (2012), a aplicação de KDD permite que novos conhecimentos com informações de qualidade sejam obtidos, desde que sejam respeitadas as regras e propriedades de cada uma dessas etapas.

A Figura 1 mostra as etapas do processo de descoberta de conhecimento em bancos de dados propostas por Fayyad *et al.* (1996).

Figura 1 – Etapas do processo de KDD



Fonte: Fayyad *et al.* 1996, p. 5.

Fayyad *et al.* (1996) ressalta a importância de se utilizar KDD devido ao aumento rápido do volume de dados gerados e afirma que, para se obter uma aplicação eficiente, é preciso que as ferramentas utilizadas tenham capacidade de extrair e processar informações de maneira precisa e com alta qualidade.

2.2 USO DE DADOS PARA OS NEGÓCIOS

A cada dia se torna mais comum que equipes de gestão e profissionais de posicionamento estratégico façam uso da análise de informações para obter vantagem nos seus negócios. Chamada de inteligência de negócio, ou *Business Intelligence* (*BI*) (MIKROYANNIDIS; THEODOULIDIS, 2010) engloba um conjunto de técnicas e ferramentas que tem por objetivo fornecer às empresas o apoio necessário para a tomada de decisão.

> Na literatura, empresas como a Harrah's Entretenimento, Amazon.com, companhias aéreas Continental e Netflix são exemplos de empresas de destaque que amadureceram nas suas aplicações de BI e, assim, se destacaram em posições de liderança de mercado (DA SILVA *et al.*, 2016, p. 2780)

Grandes empresas também fazem uso da visualização de dados como ferramenta nos processos de tomadas de decisão em áreas estratégicas, gerenciais, comerciais e *marketing*. Nesta mesma linha, empresas de pequeno porte vêm começando a adequar seus processos a fim de se beneficiarem das vantagens oferecidas pelo *BI c*ontudo, às vezes encontram dificuldades em se adaptarem, como nos casos em que contam com sistemas que não foram previamente preparados e alimentados com dados com qualidade.

Segundo Abukari e Jog (2003), para se obter uma implantação bem sucedida de BI é preciso seguir alguns passos:

- Identificar as necessidades a serem endereçadas na solução de BI. Elas devem ser relevantes aos objetivos e estratégias do negócio;
- Identificar as fontes de dados já existentes na organização. As organizações já têm uma infinidade de informações em bancos de dados, planilhas e arquivos;
- Extrair, transformar e carregar os dados para criar uma base multidimensional orientada por assunto (ou fato). Este processo deve garantir que todas as informações relevantes sejam contempladas;
- Ajudar a organização a escolher a ferramenta de apresentação para visualizar e analisar as informações resultantes da etapa anterior;
- Criar relatórios padrões, permitir análises sob demanda e mineração de dados (*Data Mining*) visando obtenção de *insights*;
- Planejar uma implantação de forma abrangente para toda corporação, de forma a garantir que os tomadores de decisão tenham a informação adequada quando e onde eles precisarem.

Outro recurso que considera o uso de dados no ambiente corporativo é o *Data Warehouse*. Segundo Paim (2003), no início da década de 90 foi necessário criar uma solução capaz de atender a necessidade crescente de geração de informações para gestão e decisões de alto nível corporativo. Assim, os *Data Warehouses* foram projetados para que os dados pudessem ser armazenados e acessados de maneira que não ficassem restritos às tabelas e linhas relacionais (DOMENICO, 2001).

2.3 EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DE DADOS

Em um ambiente de análise de dados, os métodos utilizados para extração dos dados de suas bases de origem, sua transformação e carregamento para o ambiente de processamento formam o que é conhecido como ETL. Estas etapas visam gerar informações concretas a partir das grandes bases de dados empregadas que, por sua vez, são formadas pelo uso recorrente dos *softwares* utilizados no contexto empresarial (ABREU, 2008)

Na sua fase inicial, a *ETL* busca extrair ou captar dados dos mais diversos *softwares* e seus recursos de armazenamento de dados e isto inclui aqueles que operam em ambiente *web*, *mobille* ou *desktop*. Nesta etapa há a preocupação em converter os dados para um formato único e homogêneo para que possam ser processados.

A partir daí, passa-se à etapa de transformação dos dados e que é o foco principal deste trabalho. Nessa fase são realizadas correções, modificações e limpeza de dados, padronizações, adequações e tratamento de inconsistências, de modo que os dados passem a ter padrões e possam ser retiradas informações que façam parte das regras de negócios e necessidades da organização que os utiliza.

A terceira etapa da *ETL* executa o carregamento dos dados modificados na base de dados consolidada para serem processados e/ou apresentados em relatórios de análise ou *dashboards*.

Os processos de *ETL* e a aplicação de ferramentas de limpeza de dados podem consumir até um terço do orçamento de um projeto de *Data Warehouse* (FERREIRA *et al.*, 2010).

2.4 TRANSFORMAÇÃO DE DADOS

A fase de transformação de dados da ETL implica na aplicação de metodologias e práticas de modificação de bases de dados a fim de melhorar sua qualidade para análise (ZORZO, 2009).

A transformação de dados se torna importante nos processos de descoberta de conhecimento e análise de dados para negócio, pois nem sempre os *softwares* que geram os dados de entrada estão preparados ou foram projetados com a preocupação de fornecer boa matéria prima para análises de dados.

Problemas recorrentes em dados oriundos de várias fontes são a duplicidade dos dados e a falta de padronização e isto faz com que, algumas vezes, os dados sejam armazenados e apresentados repetidamente, ou seja, guardados de diferentes formatos (um mesmo dado coletado de formas diferentes e armazenado várias vezes). Isto pode acontecer ambientes onde *softwares* utilitários oferecem mais liberdade de escrita aos usuários em campos abertos.

Nos processos de transformações de dados devem ser consideradas as regras de negócios determinadas pela organização que procura por novas informações nos dados. Por isto, a movimentação do mercado, que muda o interesse de organizações por determinadas informações, também precisa ser levada em consideração.

Transformação de Dados é a fase que exige o maior número de recursos na montagem de um *Data Warehouse* devido à grande quantidade de modificações, cálculos e processamentos que faz nos dados (INMON *et al.*, 2008).

Desta maneira, a transformação deve assegurar que os dados foram transformados corretamente segundo as regras de negócio estabelecidas (FERREIRA *et al.*, 2010)

2.5 PROBLEMAS COMUNS DE QUALIDADE DOS DADOS

As operações de limpeza de dados destinam-se a problemas de qualidade existentes nas instâncias dos dados (OLIVEIRA e CARVALHO, 2008).

Motivo da necessidade de transformação de dados, os problemas de qualidades contêm uma variação grande de diferentes situações em que ocorrem. Para referenciar esses problemas utiliza-se o termo “ruído” e, segundo Ferreira *et al.* (2010), eles podem ser causados por erros aleatórios, como incerteza de mediação (perca de dados por dispersão de valores), ou mesmo por problemas de coleta de dados, como erros de digitação.

Segundo Camilo e Silva (2009) o processo de limpeza dos dados geralmente envolve filtrar, combinar e preencher valores vazios, sendo que essas são as ações mais executadas.

As fontes de origem dos dados podem conter diversos formatos e, segundo Olson e Delen (2008), por isto elas podem trazer inúmeras barreiras que impedem sua leitura correta.

Batista (2003) afirma que o tratamento de valores desconhecidos é de grande interesse prático e teórico. Em diversas aplicações é importante saber como proceder quando as informações disponíveis estão incompletas ou quando as fontes de informação se tornam indisponíveis. Para o autor, estas são situações que se repetem dentro de empresas com bases de dados projetadas ou alimentadas de forma desorganizadas.

Segundo Camilo e Silva (2009) algumas situações comuns da limpeza de dados são a remoção de registros com problema, a atribuição de valores padrões e os agrupamentos auxiliares na visão de melhores valores.

Regras de negócios também podem atrapalhar a limpeza de dados, por exemplo, se uma empresa tem um produto identificado por mais de um nome e usuários diferentes alimentam a base de dados, isto impossibilita a atribuição de valores padrão.

Situações com variação de letras minúsculas e maiúsculas para escrita de nomes é outro problema comum e que deve ser resolvido pela transformação de dados por meio da atribuição de um valor padrão para escrita.

Este trabalho manterá seu foco considerando a recorrência dos problemas de coletas de dados e buscará entender dificuldades como:

- Correções de nomes e endereços
- Dados duplicados
- Dados faltantes

2.5.1 Nomes e endereços

A limpeza de nomes e endereços são importantes no contexto empresarial. Este tipo de informação influencia diretamente em processos que envolvem produtos e clientes.

Segundo Paulo (2004) a limpeza de dados de nomes e endereços postais tem sua importância no relacionamento com o cliente. O autor cita algumas ferramentas comerciais que focam na limpeza desse tipo dedados: Idcentric, Pureintegrate, Quick Address, Reunion, Trillium e Vality.

Rahm e Do (2000) explicam que conflitos de nomenclatura surgem quando, em uma mesma base de dados existem dois ou mais nomes iguais para objetivos diferentes.

Christen *et al*. (2002) ressaltam que o aprendizado e a padronização de nomes e endereços são importantes para a integração dos dados, garantido que informações redundantes ou falsas não sejam introduzidas na base de dados de um *Data Warehouse*.

Um exemplo pode ser entendido quando ocorre o cadastro de produtos em sistemas de venda pela *web* e os mesmos produtos são cadastrados manualmente por vendedores em uma segunda base de dados, abrindo precedentes para que estes produtos possam ser nomeados de mais de uma forma. Nestes casos, se houver uma padronização de nomes em uma das bases, os dados padronizados podem ajudar a padronizar a segunda base.

2.5.2 Dados duplicados

Dados duplicados ou redundantes são comuns em processos de ETL. Navathe (2011) afirma que os *softwares* gerenciadores de banco de dados (SGBD) devem controlar a inserção e a criação de dados duplicados, impedindo as inconsistências entre seus arquivos.

Para a detecção de dados duplicados em uma base de dados, cada dado necessita ser comparado com todos os outros dados do arquivo, ou seja, é necessário recorrer a operações de produto cartesiano (PAULO, 2004).

Os métodos de eliminação de duplicatas existentes para limpeza de dados funcionam com base no cálculo do grau de similaridade entre registros apresentados em Lee *et al.* (2000).

2.5.3 Dados faltantes

Segundo Nunes *et al.* (2009) dados faltantes é um problema comum em todo estudo que necessita de bases de dados, principalmente naqueles em que os dados são coletados por *softwares* que permitem mais liberdade ao usuário.

Não é incomum identificar valores nulos em uma tabela durante o processo ETL e isso pode impactar negativamente na precisão dos resultados da análise de dados, explica Ribeiro *et al.* (2011).

Essa imprecisão em quantidades maiores pode impossibilitar a realização do estudo, pois torna ele pouco assertivo e incapaz de gerar análises precisas sobre a situação estudada.

Ezzine e Benhlima (2018) ressaltam os inúmeros métodos e abordagens que vêm sendo usados em bancos de dados relacionais para lidar com dados faltantes, sendo muitos deles adaptados para *big datas*. Esses métodos e abordagens têm conseguido resultados positivos e têm gerado bases de dados de maior qualidade.

Segundo Münzberg *et al.* (2018) a ausência de dados faltantes ou incorretos aumenta a confiança do usuário dos dados.

Trujillo e Luján-Mora (2003) citam uma prática comum para tratar dados faltantes: a substituição de um valor nulo por um valor padrão. Os autores apresentam um exemplo de substituição dos dados faltantes de estados em endereços pelo valor ‘desconhecido’, gerando um novo padrão a ser utilizado como métrica para estudo.

3 FERRAMENTAS DE TRANSFORMAÇÃO DE DADOS

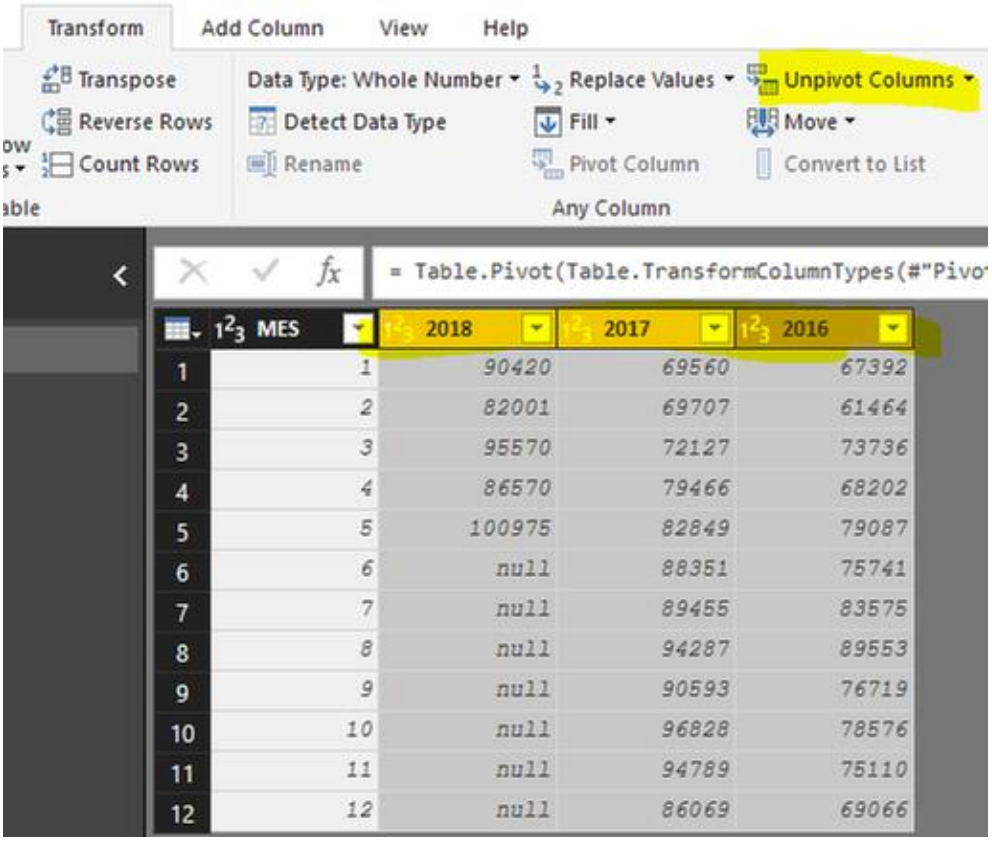
Este capítulo apresenta, de forma resumida, algumas ferramentas utilizadas pelo mercado para realizar o processamento de dados para análise, destacando-se a parte destinada ao processo de transformação de dados.

3.1 POWER B.I

O *software Power BI* (Figura 2) tem como principal função a apresentação de *dashboards*, além disso, ele também tem o processo de transformação de dados como uma das funcionalidades da ferramenta, oferecendo opções ao usuário para executar tarefas como:

- Eliminar dados que não serão utilizados no estudo ou que não têm padrão de informação para o estudo.
- Modificar dados coletados com caracteres que impedem sua leitura correta ou dados fora do padrão correto para o tipo esperado.

Figura 2 –Apresentação de Tabela *Power Bi*

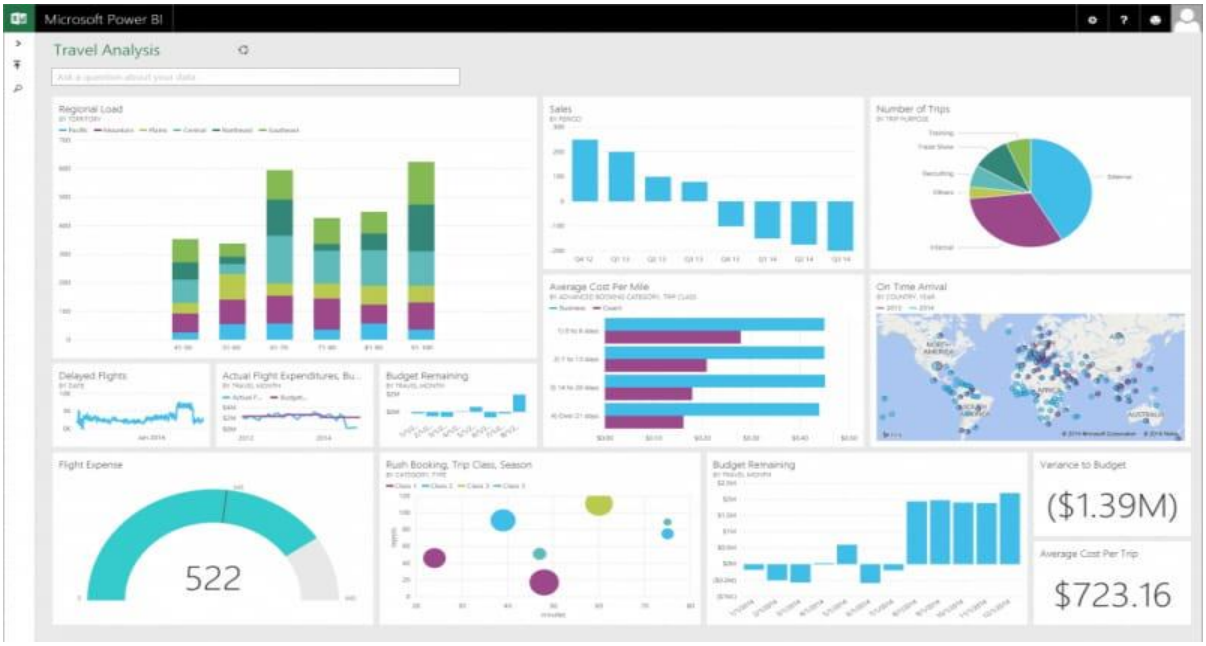


Fonte: *PowerBI.com*

Power Bi possui integrações com ferramentas de gerenciamento de banco dados e linguagem utilizadas em ciências de dados, como: Banco de Dados *Mysql*, *Sql Server* e linguagem de programação *Python*, além de aceitar formatos de arquivo como *XML*, *PDF*, *Json*, planilhas *Excel*, *CSV* entre outros. Todos estes recursos possibilitam que seus usuários recolham dados de várias fontes de dados e enriqueçam seus estudos com um número maior de dados.

Power BI oferece uma visão clara e é considerado de fácil utilização, não requerendo conhecimento da base de dados para análise (SOARES, 2017, p. 7). A Figura 3 apresenta um exemplo de *dashboard* criado no *Power BI*.

Figura 3 – Exemplo de *Dashboard* criado no *Power BI*



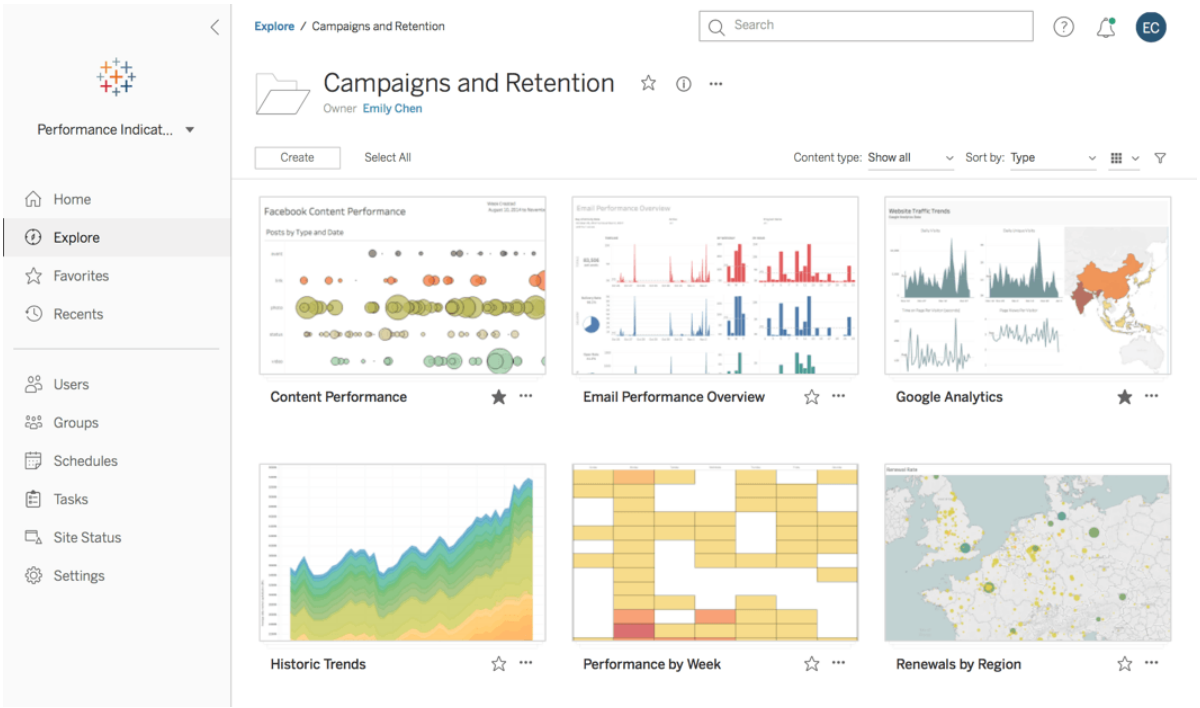
Fonte: *Power Bi.com*

Esta ferramenta possui grande quantidade de funcionalidades e permite modificações rápidas dos dados com uma interface intuitiva. Algumas vantagens do *Power BI* são a acessibilidade, a visualização personalizada, as integrações, as boas comunidades de consulta, as opções de acessos a fontes de dados e a possibilidade de integração com *softwares* sob medida.

3.2 TABLEAU

Tableau é um *software* de criação de *dashboards* que possibilita o carregamento de diversos formatos de dados e oferece ferramentas para limpeza dos dados. Ele também permite realizar o carregamento de dados e gera *dashboards* para análise visual. A Figura 4 apresenta um exemplo de sua interface.

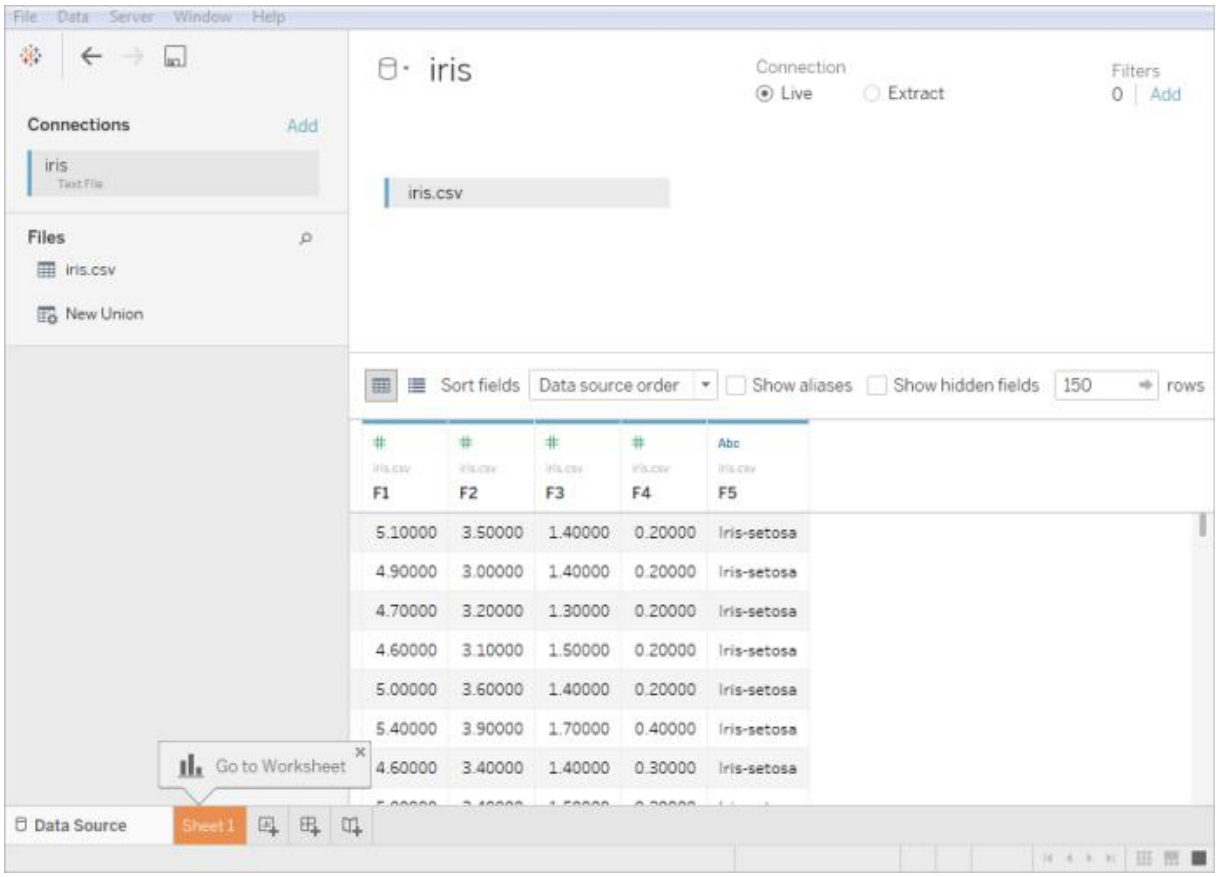
Figura 4 – Interface Tableau



Fonte: Tableau.com

O Tableau Data Engine é fornecido como parte integrante do Tableau 6.0 e se destina aos ambientes de *desktop* e servidor Wesley *et al.* (2011). Na figura 5 observa-se como o Tableau apresenta os dados.

Figura 5- Apresentação de dados



Fonte: Help Tableau

Segundo Heer *et al.* (2008) o *Tableau* registra e visualiza histórias de interação, suporta a análise de dados e a comunicação de descobertas e contribui com novos mecanismos para apresentar, gerenciar e exportar histórias. Sua interface, diferentemente do Power BI, é voltada para analistas e pouco adotada por usuários com pouco ou nenhuma experiência.

A comparação do Tableau com *Power Bi* gera uma divisão entre especialistas da área. Uma das maiores diferenças é a licencia de *software*: enquanto o *Power BI* é gratuito para uso *offline*, o Tableau tem somente um período de teste grátis. Algumas vantagens do Tableau são a sua fácil integração, a exibição de dados com resultados no foco, a facilidade com que trabalha com inúmeras plataformas e a manipulação de grandes bases de dados. Enquanto algumas desvantagens podem ser avaliadas como seu alto custo de aquisição e o fato de que é voltado somente para analistas, dificultando a utilização por usuários com pouco ou nenhum conhecimento.

4. COMPARAÇÃO DAS FERRAMENTAS ESTUDADAS

Para explorar as funcionalidades dos *softwares* estudados, foram criados cenários de testes e seus resultados serão apresentados neste capítulo.

Utilizou-se um notebook com sistema operacional *Windows* 10 64 *Bits, Intel Core I5*-7200 *CPU* @ 2.50 *GHZ* e 8 *GB* de Memória *Ram* para processar dados de estudos da Covid-19.

Foram testados os *softwares Power BI* e *Tableau* utilizando-se suas ferramentas para modelagem de relacionamentos e modelagem do estudo e limpeza dos dados.

Os dados utilizados são dados da Covid-19 do período de maio de 2020 a novembro de 2020 e podem ser acessados baixados no formato CSV em Especial Covid 19 (2020). O arquivo CSV contendo os dados está dividido em 11colunas: data, estado, cidade, tipo de lugar, casos confirmados, números de mortos, ordem para o registro do local, se é última atualização, estimativa de população em 2019, estimativa de população e código do IBGE.

Os dados foram organizados de forma para serem analisados em uma única tabela, dispensando a necessidade de se criar relações entre tabelas.

Os testes dos problemas apresentados nesse texto foram aplicados aos 2 *softwares* estudados dentro de suas propriedades. Alguns dados foram modificados para simular cenários específicos de testes, como aqueles que demandaram dados incompletos ou faltantes.

Como preparação do ambiente de estudo, caracteres foram adicionados em campos de nome, campos foram copiados para simulação de erros de dados duplicados. Os registros modificados foram escolhidos aleatoriamente.

Para o estudo de dados duplicados alguns registros foram adicionados à base original. Os dados duplicados por estados, cidades e datas foram:

- Acrelândia – Acre 03/08
- Assis Acre – Acre 05/07
- Brasiléia – Acre 22/09
- Capixaba – Acre 03/08
- Cruzeiro do Sul – Acre 04/10

- Feijó – Acre 11/08
- Manoel Urbano – Acre 15/10
- Marechal Thaumaturgo – Acre 11/11
- Marechal Thaumaturgo – Acre 22/06
- Porto Acre – Acre 03/11
- Rio Branco – Acre 18/10
- Amapá – Amapá 29/07
- Cutias – Amapá 29/10
- Ferreira Gomes – Amapá 17/09
- Itaubal– Amapá 01/10
- Laranjal do Jari – Amapá 21/10
- Macapá – Amapá 01/11
- Branquinha – Alagoas 12/07
- Cacimbinha – Alagoas 04/10(2)
- Cajueiro– Alagoas 30/09
- Igaci – Alagoas 06/11
- Acopiara – Ceara 23/04
- Aiuba – Ceara 11/04
- Potengi – Ceara 17/05
- Potiretama – Ceara 19/10
- Quixeramobim – Ceara 08/10
- Soure – Pará 26/07
- Sapacuia – Pará 09/06

Foram adicionados caracteres em alguns registros para simular erros de digitação, como:

- Santana do Araguaia 30/08 (erro no nome da cidade)
- Santa Maria do Parai (erro no nome do estado)
- Martinho Campos 21/10 (erro no nome da cidade)
- Corumbiara 07/09 (erro em número)
- Umbuzeiro (erro de número)
- Vieras 29/09 (erro no nome da cidade)

No *software Power BI* os dados são carregados por meio de função de carregamento que permite a leitura de vários formatos. Para formato CSV utilizado, o *software* permite selecionar a variável de separação das colunas de campos, dando opções como dois pontos, vírgula, ponto e vírgula, espaço, tabulação ou sinal de igualdade e permite a personalização.

Nos testes executados o delimitador de separação utilizado foi a vírgula. Foi configurada a linguagem do *software* em UTF-8 para que o arquivo com acentuações pudesse ser lido. Em seguida o arquivo de dados foi processado na ferramenta Power Query, para visualização da separação das colunas e filtros que permitem a separação de padrões de uma coluna.

O primeiro teste verificou dados faltantes na coluna cidades. Foi detectado que os campos que obtém essa característica somente nessa coluna eram registros referentes aos números totais de estados, sendo assim esses registros foram excluídos a fim de evitar duplicidade de informações, já que é possível e mais correto se contabilizar em visualizações a soma dos dados individuais de registro de cada cidade que recebeu no seu campo estado, a sigla do estado equivalente.

A Figura 7 apresenta como estavam os registros ao serem carregados no *software*:

Figura 7 – Apresentação de dados Faltantes

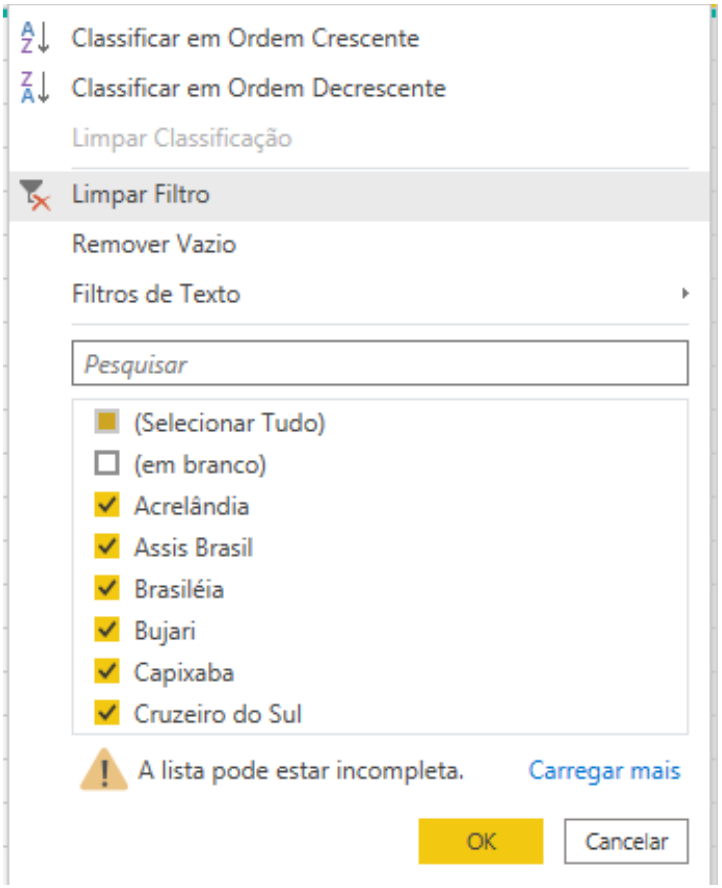
	date	state	city	place_type	confirmed	deaths
1	2020-11-12	AC		state	32515	707
2	2020-11-11	AC		state	32113	705
3	2020-11-10	AC		state	31926	704
4	2020-11-09	AC		state	31788	701
5	2020-11-08	AC		state	31707	699
6	2020-11-07	AC		state	31575	699
7	2020-11-06	AC		state	31326	697
8	2020-11-05	AC		state	31248	697
9	2020-11-04	AC		state	31218	696
10	2020-11-03	AC		state	30962	695
11	2020-11-02	AC		state	30954	693
12	2020-11-01	AC		state	30951	693
13	2020-10-31	AC		state	30796	693
14	2020-10-30	AC		state	30758	693
15	2020-10-29	AC		state	30638	692
16	2020-10-28	AC		state	30545	691
17	2020-10-27	AC		state	30380	690
18	2020-10-26	AC		state	30317	689
19	2020-10-25	AC		state	30304	687
20	2020-10-24	AC		state	30217	686
21	2020-10-23	AC		state	30121	686
22	2020-10-22	AC		state	30028	685
23	2020-10-21	AC		state	29925	682
24	2020-10-20	AC		state	29865	682
25	2020-10-19	AC		state	29765	679
26	2020-10-18	AC		state	29718	679

Fonte: O Autor

No *Power Bi* essa exclusão pode ser feita pela extração da seleção de campos com valor vazio naquela coluna, excluindo assim esses registros da base visualizada durante os estudos.

A Figura 8 apresenta como é a caixa de seleção e o campo já excluído da visualização dos dados.

Figura 8 – Caixa de Seleção

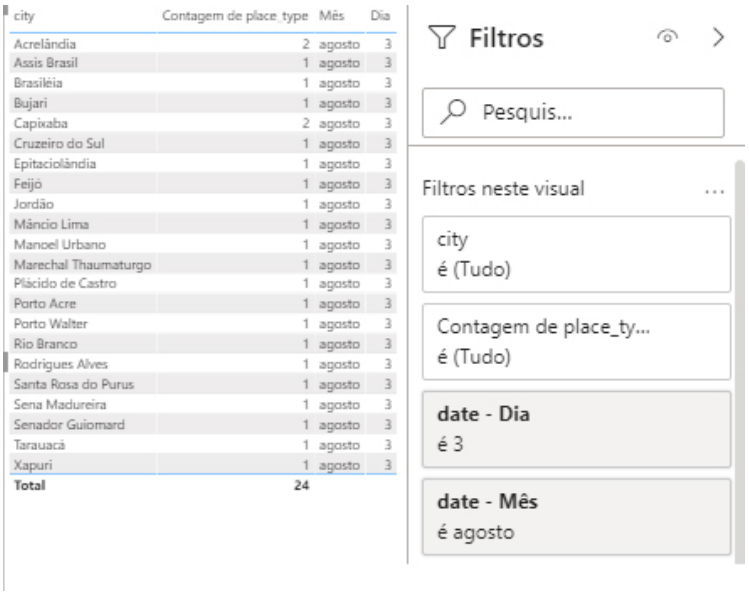


Fonte: O Autor

Nessa parte dos testes foram retirados 6.516 registros com campos vazios, eliminando a duplicidade na contagem de dados de números de casos, mortes e população. Restaram 950.739 registros para as etapas seguintes dos testes.

O cenário que considerou os dados duplicados exigiu uma sequência de filtragens dentro da estrutura do *Power BI*. Uma delas foi a separação de filtros por estado e depois por meses e dias, adicionando-se como parâmetro a contagem de campos de tipo de registro, para chegarmos à unificação do registro identificando repetições pela numeração de identificação. A figura 9 apresenta um exemplo dessas adequações.

Figura 9–Identificação de dados duplicados



Fonte: O Autor

Nota-se na filtragem do dia 03 de agosto que as cidades de Acrelândia e Capixaba identificaram dois registros, enquanto o padrão esperado é um registro máximo por cidade em uma determinada data. Esses registros devem ser eliminados, reconduzindo-os a sua identificação correta, ou seja, no caso desse estudo, o registro duplicado deve ser apagado da base de dados.

Nessa etapa foram identificados, por meio de filtragens, os 31 registros duplicados. A retirada foi feita primeiramente pela substituição de um campo do registro para torná-lo único e depois a retirada dos registros que serão apresentados. Depois dessa etapa restaram 950.708 registros.

A última etapa de teste foi identificar campos de registros que estavam com dados fora do padrão, como registro de nomes, locais e registros com algum caractere indevido.

Nessa etapa os filtros foram feitos por cidades utilizando-se o sistema de padronização por seleção do *Power BI*. Foi permitida a retirada de todas as irregularidades conseguindo assim colocar todos os campos no padrão.

No *Software Tableau* foram carregados os dados do mesmo arquivo .*CSV* de origem. O *software* permite que os dados sejam carregados diretamente da base de dados e sejam apresentados em formato de tabela. Ele permite uma hierarquia entre pastas e planilhas semelhante ao que é encontrado no *Excel* e isto permite a organização de várias tabelas utilizadas em um estudo.

Os testes foram feitos na mesma ordem dos testes executados no *software Power BI* e o primeiro cenário tratou a questão dos dados faltantes. Foi identificada a necessidade de exclusão das linhas com problema, sendo que, neste caso, o padrão são campos que têm valores nulos na coluna cidade.

O *Tableau* permite a utilização de 4 formas de exclusão:

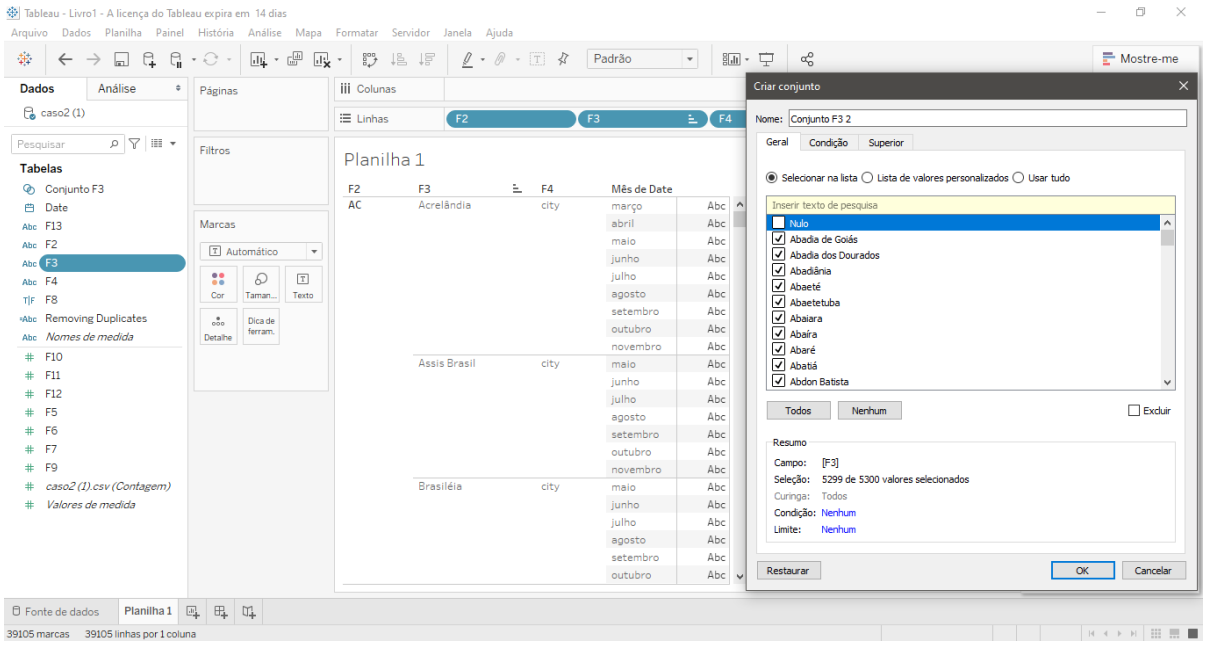
1. Usar um conjunto como filtro
2. Filtro de exclusão que usa apenas valores relevantes
3. Usar um parâmetro como um Filtro
4. Usar uma ação de conjunto

Foram escolhidos para o teste somente os dois primeiros, pois eles permitem alterações em dados que não estão aparecendo na visualização, mas compõe a tabela. Como a tabela tem vários registros, dados não aparecem na visualização.

Inicialmente foram selecionados os dados da coluna cidade e em seguida criou-se um uma caixa de seleção com valores selecionados dos campos, excluindo-se apenas os campos com valores *NULL*.

Assim, o *software* retorna como resultado a retirada de 6.516 registros que possuem valores nulos e, então, o problema de duplicidades desses registros foi retirado. A figura 10 mostra o resultado após os testes ao lado a caixa *dropdown* utilizada na solução testada.

Figura 10 – Solução e Apresentação de retirada de dados nulos *Tableau*

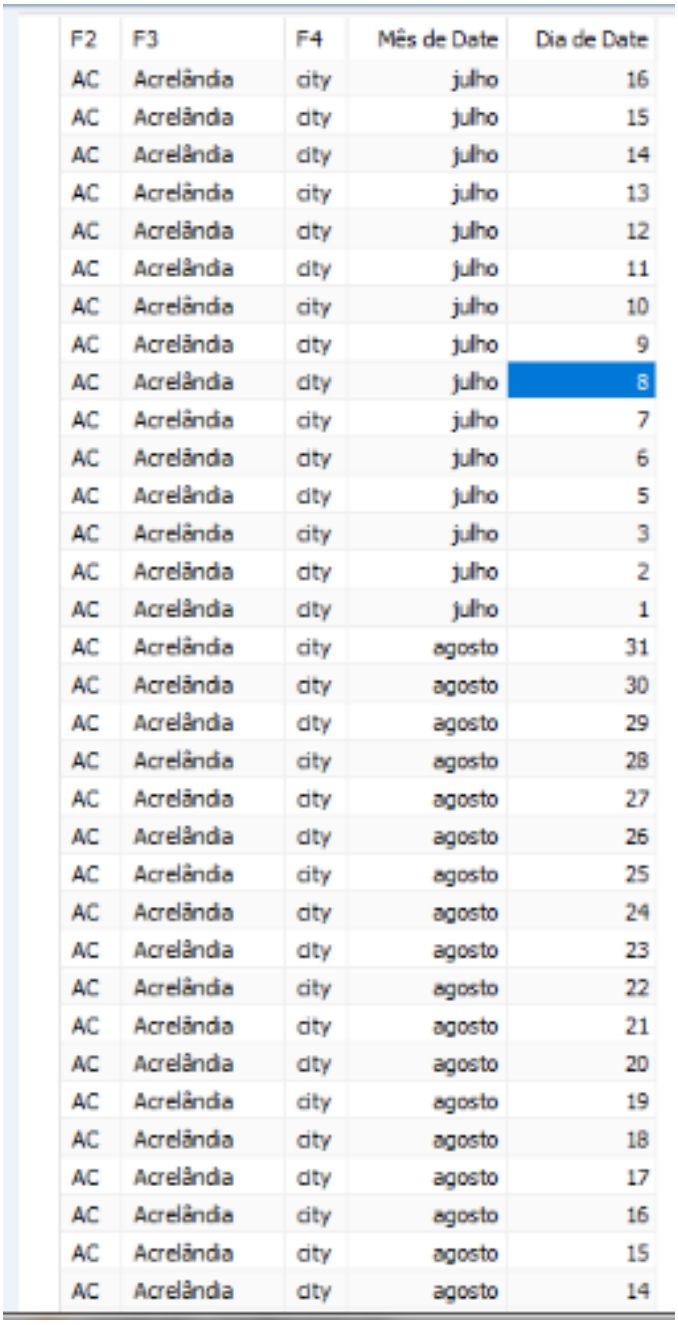


Fonte: O Autor

No segundo teste com dados duplicados o *Tableau* permitiu a filtragem por um comando de código onde foram selecionadas as colunas de interesse. Houve a utilização do comando denominado *min* para algum campo do registro que apresentava duplicidade.

O resultado foi a remoção dos 31 registros duplicados, restando assim somente os dados com registros únicos. A Figura 11 apresenta a retirada de dados duplicados.

Figura 11 – Dados Duplicados Tableau

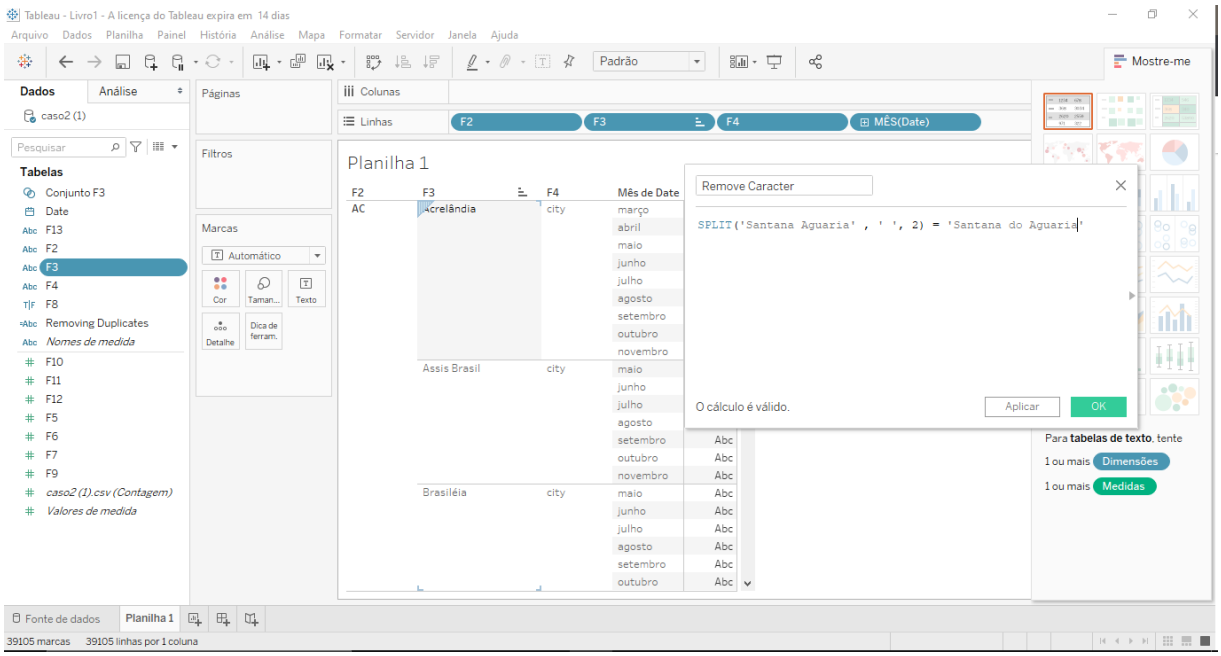


F2	F3	F4	Mês de Date	Dia de Date
AC	Acrelândia	city	julho	16
AC	Acrelândia	city	julho	15
AC	Acrelândia	city	julho	14
AC	Acrelândia	city	julho	13
AC	Acrelândia	city	julho	12
AC	Acrelândia	city	julho	11
AC	Acrelândia	city	julho	10
AC	Acrelândia	city	julho	9
AC	Acrelândia	city	julho	8
AC	Acrelândia	city	julho	7
AC	Acrelândia	city	julho	6
AC	Acrelândia	city	julho	5
AC	Acrelândia	city	julho	3
AC	Acrelândia	city	julho	2
AC	Acrelândia	city	julho	1
AC	Acrelândia	city	agosto	31
AC	Acrelândia	city	agosto	30
AC	Acrelândia	city	agosto	29
AC	Acrelândia	city	agosto	28
AC	Acrelândia	city	agosto	27
AC	Acrelândia	city	agosto	26
AC	Acrelândia	city	agosto	25
AC	Acrelândia	city	agosto	24
AC	Acrelândia	city	agosto	23
AC	Acrelândia	city	agosto	22
AC	Acrelândia	city	agosto	21
AC	Acrelândia	city	agosto	20
AC	Acrelândia	city	agosto	19
AC	Acrelândia	city	agosto	18
AC	Acrelândia	city	agosto	17
AC	Acrelândia	city	agosto	16
AC	Acrelândia	city	agosto	15
AC	Acrelândia	city	agosto	14

Fonte: O Autor

Já o teste com dados com irregularidades, o procedimento no *software* foi utilizado uma função *SPLIT* que permite a substituição do valor com erro por um valor padrão correto. O resultado foi a correção completa de todos os dados que continham irregularidades, sendo o seu padrão um número ou uma *string.* A Figura 12 apresenta este cenário de testes.

Figura 12 – Nomes e Endereços Tableau



Fonte: O Autor

Após as aplicações dos testes os *softwares* obtiveram sucesso em produzir uma base de dados com qualidade. Removendo todos os campos que tinham registros vazios, assim se mostrando eficaz em situações de retirada de registros e identificação de campos vazios.

No teste com dados duplicados ambos os *softwares* tiveram sucesso, porém diferente do primeiro teste, por conter uma quantidade de campos menor apresentando essa irregularidade, o *Software Power BI* exigiu uma utilização maior de funcionalidade e processos para individualizar os dados e apresentar ao usuário os campos que necessitam de alguma adequação.

Já o *Tableau* apresentou uma solução direta que utiliza comandos de programação, exigindo conhecimento do profissional para identificação dos registros com irregularidade e sua remoção.

No cenário de teste com escritas irregulares o *Power* BI apresentou uma solução com mesma necessidade de individualizar os padrões, porém este *software* uma ferramenta simples de correção do problema, apenas com a digitação de um campo. Já o *Tableau* exigiu uma solução lógica para que fosse substituído um registro diretamente, podendo causar erros se os campos com as irregularidades não estiverem bem identificados. Como a não localização do mesmo pelo software.

CONSIDERAÇÕES FINAIS

Um grande acervo de dados é gerado diariamente nas empresas, sejam elas privadas ou públicas e isto implica o uso de formas eficientes para gestão de dados, para que as tomadas de decisões empresariais possam ser apoiadas por estas informações. Contudo, em muitos casos estes dados possuem diferentes fontes e estruturas diferentes, necessitando que, antes de serem analisados, eles necessitem passar por processos de limpeza, padronização ou transformação.

Assim, evidencia-se a importância da limpeza de dados. Esta é um dos processos da *ETL* e visa melhorar a qualidade de bases de dados originais.

Esse trabalho teve como objetivo realizar testes, analisar e comparar os processos e ferramentas de limpeza de dados presentes no mercado. Para isto, foram elaborados cenários de testes, coletados resultados, analisados processos e comparadas técnicas de duas ferramentas, obtendo uma análise comparativa entre funcionalidades e resultados apresentados por tais ferramentas.

Foi desenvolvida uma pesquisa bibliográfica sobre temas relacionados à limpeza de dados e estudados *softwares* de análise de dados que oferecem recursos para limpeza de dados.

Os principais desafios surgiram em elaborar um conjunto de testes onde fosse possível compreender o funcionamento e a aplicação da limpeza de dados, em todos os sistemas testados. E elaborar uma comparação precisa entre a eficácia dos *softwares*.

Futuramente poderão ser realizados novos estudos sobre outras funcionalidades dos *softwares* Power Bi e Tableau no que diz respeito a erros e problemas encontrados na limpeza de dados.

REFERÊNCIAS

ABREU, Fábio Silva Gomes da Gama e. **Desmistificando o Conceito de ETL.** Revista de Sistemas de Informação da FSMA n. 2, 2008. Disponível em; http;//www.fmsa.edu.br/si/Artigos/FSMA_SI_2008_Pincipal_1.html Acesso em: 8 de out. 2020

ABUKARI, K.; JOG, V. Business Intelligence in action. CMA Management, 2003.

Batista, Gustavo Enrique de Almeida Prado. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Diss. Universidade de São Paulo, 2003.

BUENO, M F; VIANA, M R. **Mineração de dados: aplicações, eficiência e usabilidade. Instituto Nacional de Telecomunicações** – Inatel, 2012.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: **Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), p. 1-29, 2009.

Christen, P.; Churches, T. e Zhu, J.X. **Probabilistic name and address cleaning and standardization.** The Australian Data Mining Workshop. 2002.

Da Silva, Rafaela Alexandre, Fernando Cesar Almeida Silva, Carlos Francisco Simões Gomes. **O uso do Business Intelligence (BI) em sistema de apoio à tomada de decisão estratégica.** *Revista GEINTEC-Gestão, Inovação e Tecnologias* 6.1 (2016): 2780-2798

DOMENICO, J. A. **Definição de um Ambiente Data Warehouse em uma Instituição de Ensino Superior**. Florianópolis, 2001. 137 f. Dissertação (Mestrado em Engenharia de Produção) - Departamento de Engenharia de Produção, Universidade Federal de Santa Catarina.

ESPECIAL COVID-19 - Dados por Município. Brasil.IO, 15 mar. 2020. Disponível em: https://www.brasil.io/covid19/. Acesso em: 14 set. 2020

Ezzine, Imane; Benhlima, Laila. "**A study of handling missing data methods for big data**." *2018* IEEE 5th International Congresson Information Science and Technology 2018.

FAYYAD, U M; Piatetsky-shapiro, G; Smyth, P. **From knowledge discovery and data mining:** AI Magazine, Volume 17 Number 3 (1996).

FERREIRA, João *et al.* **O Processo ETL em Sistemas Data Warehouse**. INFORUM, 2010. Simpósio de Informática.

GOEBEL, M; GRUENWALD, L. **A survey of data mining and knowledge discovery software tools.** Nova York. SIGKDD Explorations Volume 1: ACM, 1999.

Heer, J.; Mackinlay, J.; Stolte, C.; Agrawala, M. **“Graphical histories for visualization: Supporting analysis, communication, and evaluation”**, IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), vol. 14, 2008, pp. 1189–1196.

Inmon, Bill; Strauss, Derek; Neushloss, Genia. **DW 2.0: The Architecture for the Next Generation of Data Warehousing.** Morgan Kaufmann, 2008.

Lee, M. L.; Ling, T. W. e Low, W. L. IntelliClean: **A Knowledge-Based Intelligent Data Cleaner.** In Proceedings of the ACM SIGKDD. Boston, EUA. 2000

MIKROYANNIDIS, Alexander; THEODOULIDIS, Babis. **Ontology management and evolution for business intelligence.** International Journal of information Management [s. l.], v. 30, ed. 6, p. 559-566, 2010.

Münzberg, A., Sauer, J., Hein, A., & Rösch, N.. **The use of ETL and data profiling to integrate data and improve quality in food databases**. 14th International Conference on Wireless and Mobile Computing, Networking and Communications *(Wi Mob)* (p. 231-238). IEEE., 2018

NAVATHE, Shamkant B. ELMASRI, Ramez. **Sistema de banco de dados**. 6. Ed São Paulo: Pearson Education, 2011.

Nunes, Luciana Neves, Mariza Machado Klück, and Jandyra Maria Guimarães Fachel. **Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos**. *Cadernos de Saúde Pública* 25 (2009): 268-278.

OLIVEIRA, R. R; CARVALHO, C. L. **Algoritmos de agrupamento e suas aplicações**. Technical report, Universidade Federal de Goiás, 2008.

OLSON, D. L; DELEN, D. **Advanced Data Mining Techniques**. Springer, 2008

PAIM, Fábio Rilston Silva. **Uma metodologia para definição de requisitos em sistemas Data Warehouse**. Orientador: Jaelson Freire Brelaz de Castro. 2003. Dissertação (Mestrado Ciências da Computação) - Universidade Federal de Pernambuco Centro de Informática, [S. l.], 2003.

PAULO, MARCELO VICENTE DE. **Explorando o potencial da plataforma Lattes como fonte de conhecimento organizacional em ciência e tecnologia.** (2004).

RAHM, E. E Do, H. H. **Data Cleaning: Problems and Current Approaches**. IEEE Bulletin of the Technical Committee on Data Engineering, 24(4). 2000.

Ribeiro, Lívia de S., Ronaldo R. Goldschmidt, and Maria Cláudia Cavalcanti. **Complementing data in the ETL process**. *International Conference on Data Warehousing and Knowledge Discovery*. Springer, Berlin, Heidelberg, 2011

SILVA, M P S. **Mineração de dados – conceitos, aplicações e experimentos com WEKA.** Em Escola Regional de Informática RJ/ES, v.1, pp.19-21. Sociedade Brasileira de Computação, 2004

Soares, Ana Camila Fonseca. **Análise de Ferramentas de Business Intelligence com destaque dos serviços de BI na Cloud Computing**. (2017)

THOMÉ, A C G. **Redes neurais – uma ferramenta para kdd e data mining. Inteligência computacional**, 2008.

Trujillo, J. e Luján-Mora, S., **A UML Based Approach for Modeling ETL Processes in Data Warehouses Conceptual Modeling** - ER 2003 (Vol. 2813/2003, pp. 307-320). Berlin Heidelberg: Springer -Verlag, 2003.

Wesley, R.; Eldridge, M.; Terlecki, P. **"An analytic data engine for visualization in tableau"**. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2011, pp. 1185–1194.

ZORZO, André Luis. **ETL 2.0 – Uma proposta de extensão ao processo de extração, transformação e carga voltada à integração de dados estruturados e não estruturados**. 2009. Monografia (Graduação em Sistemas de Informação) - UNIVERSIDADE FEDERAL DE SANTA CATARINA, [*S. l.*], 2009.