UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UMA ANÁLISE EXPLORATÓRIA E PRÁTICA DO USO DO ETL EM PORTAIS DE TRANSPARÊNCIA

Dissertação de Mestrado

Marcus Vinicius Santana Poletti

São Cristóvão - Sergipe
2023

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Marcus Vinicius Santana Poletti

UMA ANÁLISE EXPLORATÓRIA E PRÁTICA DO USO DO ETL EM PORTAIS DE TRANSPARÊNCIA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal de Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação.
Orientador: Prof. Dr. Methanias Colaço Rodrigues Júnior.

São Cristóvão - Sergipe
2023

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

P765a

Poletti, Marcus Vinicius Santana
Uma análise exploratória e prática do uso do ETL em portais de transparência / Marcus Vinicius Santana Poletti ; orientador Methanias Colaço Rodrigues Júnior. – São Cristóvão, SE, 2023.
59 f. : il.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2023.

1. Computação. 2. Armazenamento de dados. 3. Banco de dados. I. Rodrigues Júnior, Methanias Colaço, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: Marcus Vinicius Santana Poletti

Em 30 dias do mês de agosto do ano de dois mil e vinte três, com início às 14h00min, realizou-se na Sala de Seminários do PROCC da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloísio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **Marcus Vinicius Santana Poletti**, que desenvolveu o trabalho intitulado: "***UMA ANÁLISE EXPLORATÓRIA E PRÁTICA DO USO DE ETL EM PORTAIS DE TRANSPARÊNCIA***", sob a orientação do Prof. Dr. **Methanias Colaço Rodrigues Júnior**. A Sessão foi presidida pelo Prof. Dr. **Methanias Colaço Rodrigues Júnior** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Prof. Dr. **Raphael Pereira de Oliveira** (DSI - UFS) e, em seguida, o Prof. Dr. **Rafael Oliveira Vasconcelos** (Procc/UFS). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) APROVADO "*(aprovado/reprovado)*". Atendidas as exigências da Instrução Normativa 05/2019/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 04/2021/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária "*Prof. José Aloísio de Campos*", 30 de agosto de 2023.

Prof. Dr. Methanias Colaço Rodrigues Júnior
(PROCC/UFS)
Presidente

Prof. Dr.Rafael Oliveira Vasconcelos
(PROCC/UFS)
Examinador Interno

Prof. Dr. Raphael Pereira de Oliveira
(DSI - UFS)
Examinador Externo

Marcus Vinicius Santana Poletti
Candidato

Agradecimentos

Primeiramente quero agradecer à minha família, principalmente minha mãe, por todo apoio, paciência e compreensão.

A Lydiane pela paciência e otimismo durante toda a caminhada, sempre me dando confiança e força para seguir em frente.

Ao meu professor e orientador Methanias, por toda paciência e boa vontade em passar seu conhecimento. Quero agradecer também pela confiança e oportunidade de trabalhar no projeto Transparência Traduzida.

Por fim, a todos que estiveram envolvidos nessa caminhada e que contribuíram de alguma forma. O meu mais sincero agradecimento.

Resumo

Contexto: Os portais de dados abertos são construídos com base em processos *ETL* (*Extract, Transform and Load*), os quais aumentam a qualidade e interoperabilidade dos dados, perfazendo um subsistema crítico para estas aplicações, passível de pesquisas avaliativas para melhorias. **Objetivo**: Analisar publicações sobre o uso de ETL em portais de transparência, a fim de caracterizá-las quanto aos seus cenários, impactos, métodos empíricos e dados bibliométricos gerais. A partir dessa caracterização, desenvolver e avaliar um módulo ETL para um portal de transparência, comparando-o qualitativamente com módulos desenvolvidos em duas ferramentas ETL amplamente usadas no mercado. Adicionalmente, foi feita uma análise das eficiências dos procedimentos de carga gerados pelos 3 tratamentos avaliados. **Método:** Utilizando a estratégia PICO (População, Intervenção, Comparação e Resultado), foi realizado um mapeamento sistemático da literatura. Além disso, foi executada uma Pesquisa-Ação para construção de procedimentos *ETL* do Anuário Econômico de Sergipe. As ferramentas avaliadas durante o processo de desenvolvimento foram: (1) *Pentaho Data Integration - Kettle, Open Source,* e (2) *SQL Server Integration Services - SSIS, Closed Source,* contra (3) um código ETL construído na linguagem *Python*. **Resultados**: De um total de 204 publicações pesquisadas, foram selecionados 25 trabalhos, dos quais 40% apresentam, como principal impacto para os portais, a disponibilidade de suporte para construção de cargas por meio de uma interface gráfica, seguida da possibilidade de conectividade entre bases de dados heterogêneos (27%) e capacidade de monitoramento de cargas (22%). Em relação à automação real de cargas e seu controle de qualidade, respectivamente, apenas 8% e 3% dos trabalhos discutiram os impactos dessas características. No que concerne à pesquisa-ação, foram encontradas evidências de destaque da ferramenta *Kettle*, do ponto de vista da usabilidade e eficiência de desenvolvimento por meio de interface gráfica, bem como do ponto de vista da curva de aprendizagem. Na sequência, vieram a linguagem de programação *Python* e a ferramenta *SSIS*. Em relação à eficiência, a mensuração do tempo de carga mostrou um melhor desempenho da linguagem *Python*, seguida do *Kettle* e do SSIS. **Conclusão:** O trabalho mostrou que o uso de ETL em portais de transparência ainda carece de estudos comparativos e de viabilidade. Nesse sentido, um desafio existente é a escassez de pesquisas que realizem replicações para consolidar e validar os trabalhos já publicados, evidenciado pela insuficiência de experimentos controlados na área. Além disso, análises sobre o controle de qualidade das cargas foram uma importante

lacuna identificada. Por fim, definidas as prioridades contextuais de portais de transparência, como, por exemplo, a eficiência das cargas ou a eficiência de desenvolvimento, a avaliação sistematizada de soluções disponíveis, tal como a proposta nesta dissertação, norteia situações de *trade-off* e seleção do melhor custo-benefício.

Palavras-chave: Portais da Transparência, ETL, eficiência, usabilidade, qualidade.

Abstract

Context: Open data portals are built based on ETL processes (Extract, Transform and Load), which increase data quality and interoperability, making a critical subsystem for these applications, subject to evaluative research for improvements. **Objective**: To analyze publications on the use of ETL in transparency portals, in order to characterize them in terms of their scenarios, impacts, empirical methods and general bibliometric data. From this characterization, develop and evaluate an ETL module for a transparency portal, qualitatively comparing it with modules developed in two ETL tools widely used in the market. Additionally, an analysis of the efficiencies of the loading procedures generated by the 3 evaluated treatments was carried out. **Method**: Using the PICO (Population, Intervention, Comparison and Outcome) strategy, a systematic mapping of the literature was carried out. In addition, an Action-Research was carried out for the construction of ETL procedures for the Economic Yearbook of Sergipe. The tools evaluated during the development process were: (1) Pentaho Data Integration - Kettle, Open Source, and (2) SQL Server Integration Services - SSIS, Closed Source, against (3) an ETL code built in the Python language. **Results**: From a total of 204 researched publications, 25 works were selected, of which 40% present, as the main impact for the portals, the availability of support for the construction of loads through a graphical interface, followed by the possibility of connectivity between bases heterogeneous data (27%) and load monitoring capacity (22%). Regarding the actual automation of loads and its quality control, respectively, only 8% and 3% of the works discussed the impacts of these characteristics. With regard to action research, outstanding evidence of the Kettle tool was found, from the point of view of usability and development efficiency through the graphical interface, as well as from the point of view of the learning curve. Next came the Python programming language and the SSIS tool. Regarding efficiency, the load time measurement showed a better performance of the Python language, followed by Kettle and SSIS. **Conclusion**: The work showed that the use of ETL in transparency portals still lacks comparative and feasibility studies. In this sense, an existing challenge is the scarcity of research that carry out replications to consolidate and validate already published works, evidenced by the insufficiency of controlled experiments in the area. In addition, analyzes on the quality control of loads was an important identified gap. Finally, once the contextual priorities of transparency portals are defined, such as load efficiency or development efficiency, the systematic evaluation of available solutions, such as the

one proposed in this dissertation, guides trade-off situations and selection of the best cost -benefit.

Keywords: Transparency Portals, ETL, efficiency, usability, quality

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BI	Business Intelligence
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
DM	Data Mart
DW	Data Warehouse
ETL	Extração, Transformação e Carga
IPC	Índice de Percepção da Corrupção
LAI	Lei de Acesso à Informação
NGOs	Non-Governmental Organizations
OGD	Open Government Data
PICO	Population, Intervention, Comparison and Outcome
SSIS	Server Integration Services
SGBD	Sistema Gerenciador de Banco de Dados ()
SQL	Structured Query Language

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO

Este capítulo pretende realizar uma breve contextualização relacionada ao tema da pesquisa, motivação, problemática, questões, objetivos e suposição que se pretende evidenciar.

1.1 CONTEXTUALIZAÇÃO

O Índice de Percepção da Corrupção (IPC) é uma classificação anual elaborada pela Transparência Internacional desde 1995, que analisa a percepção da corrupção em 180 países e territórios do mundo. Ele atribui uma pontuação de 0 a 100 a cada país, onde 0 representa uma percepção de alta corrupção e 100 representa uma percepção de integridade máxima (International Transparency, 2021a; International Transparency, 2021b). Quanto maior a pontuação, menor a percepção de corrupção no país.

No relatório de 2020, o Brasil ficou na 94ª posição com uma pontuação de 38 pontos. Essa pontuação ficou abaixo da média internacional de 43 pontos e abaixo da média da América Latina e Caribe, que foi de 41 pontos (International Transparency, 2021a; International Transparency, 2021b). Isso indica que a percepção de corrupção no Brasil ainda é considerada alta em comparação com outros países da região e do mundo (International Transparency, 2021b).

Além disso, o relatório de 2020 evidenciou a existência de corrupção ligada à pandemia da Covid-19 no Brasil. Isto incluiu casos de subornos, desfalques, superfaturamentos e falta de controle adequado em licitações e fiscalização. A corrupção nesse contexto compromete a resposta efetiva do país à crise de saúde e afeta negativamente os recursos destinados à compra de equipamentos médicos, remédios, contratação de profissionais de saúde e manutenção de hospitais e clínicas. Como resultado, a parcela mais vulnerável da população é afetada de forma desproporcional (International Transparency, 2021b, International Transparency, 2021c).

Para combater a corrupção e melhorar a transparência, algumas ações são recomendadas, como fortalecer os órgãos fiscalizadores, implementar mecanismos de controle mais rigorosos nas licitações, garantir o direito à fiscalização por parte da

população e promover a publicação de dados detalhados sobre os gastos e distribuição de recursos (International Transparency, 2021b, International Transparency, 2021c).

Apesar do Brasil ter adotado a Lei de Acesso à Informação (LAI) em 2011 (Brasil, 2011), que tem como objetivo promover a transparência por meio de dados abertos, há desafios na prática. Ao acessar os portais de transparência das prefeituras e estados brasileiros, observa-se inadequações nos dados disponibilizados, falta de cumprimento dos princípios do padrão *Open Government Data* (OGD) (Open Knowledge Foundation, 2021), que busca disponibilizar dados de forma completa, primária, oportuna, acessível, processável por máquina, não discriminatória, não proprietária e livre de licenças (Eberhardt e Silveira, 2018; Oliveira e Silveira, 2018; Bachtiar e Muhamad, 2020; Cenci, Fillotrani e Ardenghi, 2017). Essas deficiências nos portais de transparência dificultam a reutilização efetiva dos dados para promover o engajamento da população e o controle social sobre os gastos públicos (Oliveira e Silveira, 2018; Bachtiar e Muhamad, 2020; Muller, Gil-Garcia e Tirelli, 2018).

Em dezembro de 2020, o Governo Federal publicou um Plano Anticorrupção com 142 ações para melhorar a transparência pública e o engajamento da população (Comitê Interministerial de Combate à Corrupção, 2020). No entanto, para que essas ações sejam bem-sucedidas, é necessária a implementação de uma estrutura tecnológica adequada. Um dos pontos críticos é a carga de dados, especialmente em grandes instituições, onde os dados podem estar dispersos em diferentes sistemas com bases de dados heterogêneas.

Para solucionar esse desafio, é fundamental utilizar sistemas *ETL* (Extração, Transformação e Carga). Esses sistemas são responsáveis por extrair os dados de diversas fontes, transformá-los em um formato consistente e carregá-los em bases de dados integradas (Pan, Zhang e Qin, 2018; Diouf e Boly, 2017), que serão acessadas pelos portais de transparência. Essa integração dos dados é essencial para garantir que as informações sejam confiáveis, completas e acessíveis ao público de forma eficaz.

Portanto, a implementação de uma infraestrutura tecnológica adequada, incluindo sistemas *ETL* (Martínez e Galviz-Lista, 2012), é crucial para garantir a efetividade dos portais de transparência, promovendo o controle social, combatendo a corrupção e facilitando o engajamento da população nas questões públicas.

Os portais precisam construir e/ou adquirir esses sistemas, considerando impactos tais como conectividade com todos os tipos de base, facilidade e agilidade de uso por meio de interface gráfica, controle e garantia de qualidade dos dados, automação e monitoramento das cargas (Martínez e Galviz-Lista, 2012). Além das preocupações

citadas acima, os sistemas *ETL* também estão diretamente envolvidos com a criação de metadados, publicação de dados e metadados na web, bem como com a criação de registros de catálogo apropriados (Saranya et al, 2021).

Diante desse contexto, a proposta desta dissertação foi realizar um mapeamento sistemático para caracterizar artigos em relação ao desempenho de ferramentas *ETL* em portais de transparência, identificando impactos, abordagens, cenários, métodos de pesquisa e dados bibliométricos gerais. A partir desta caracterização, foi realizada uma pesquisa-ação, com o objetivo de conceber e avaliar cargas de um módulo *ETL* de um portal de transparência, considerando três cenários diferentes: dois utilizando duas ferramentas de mercado, e o terceiro cenário usando uma implementação específica. Os ciclos da pesquisa-ação foram realizados para identificar problemas, avaliar e evidenciar as melhores soluções para o desenvolvimento e execução de processos *ETL* neste tipo de portal.

O ambiente selecionado para as avaliações foi o Anuário Econômico de Sergipe (Universidade Federal de Sergipe, 2021), o qual necessita de procedimentos *ETL* para o povoamento do seu *Data Warehouse*. No mapeamento supracitado, publicado com o título "*An Exploratory Analysis of the Use of ETL in Transparency Portals*" (Poletti, Colaço Júnior & Nascimento (2023), observou-se a ausência de trabalhos com avaliações de ferramentas *ETL open source* contra *closed source*, bem como contra procedimentos *ETL* codificados manualmente, em portais de transparência. Desta forma, foram avaliados o *Pentaho Data Integration*, *Open Source*, e o *SQL Server Integration Services, Closed Source*, contra procedimentos escritos na linguagem de programação *Python*, a qual atende à demanda dos Departamentos de Economia e Sistemas de Informação da Universidade Federal de Sergipe (UFS), responsáveis pelo portal.

1.2 PROBLEMÁTICA E SUPOSIÇÃO

Para o estudo, o problema maior em questão é averiguar a suposição de maior eficiência das cargas e de maior facilidade de implementação destas, quando do uso de ferramentas, amplamente usadas no mercado, para gerar e automatizar processos *ETL*, em detrimento à codificação específica das cargas em uma linguagem de programação.

Dessa forma, devemos nos atentar aos seguintes questionamentos sobre essa pesquisa:

a) Questão 1: A implementação de processos *ETL*, por meio de ferramentas semiautônomas, aumenta consideravelmente a produtividade?

b) Questão 2: As ferramentas *ETL* de mercado podem facilitar a implementação e reduzir o seu tempo?

c) Questão 3: Qual abordagem gera cargas mais eficientes?

Identificado o problema, faz-se necessária a elaboração de uma suposição passível de investigação deve ser elaborada. A suposição em questão é: para o contexto de portais de transparência, a codificação de módulos ETL alcança melhor eficiência de carga e maior facilidade de implementação do que módulos produzidos em ferramentas ETL do mercado.

1.3 OBJETIVO GERAL

Este projeto tem como objetivo geral realizar uma análise exploratória e prática do uso do ETL em portais de transparência.

1.4 OBJETIVOS ESPECÍFICOS

Para possibilitar a realização do objetivo geral, podemos enumerar os seguintes objetivos específicos:

- Executar Mapeamento Sistemático da Literatura, com a finalidade de identificar e caracterizar os estudos existentes sobre o uso de ferramentas ETL em portais de transparência;
- Construir três cenários ETL em um Portal de Transparência: dois utilizando duas ferramentas de mercado, selecionadas no mapeamento e por participação de mercado, e o terceiro cenário usando uma implementação específica;
- Realizar estudo de caso para avaliar e comparar os três cenários, do ponto de vista da eficiência de carga dos módulos finais e da facilidade de construção e manutenção dos ETLs.

]

1.5 METODOLOGIA

O estudo foi desenvolvido sob uma perspectiva metodológica de natureza aplicada, tendo em vista que o interesse do estudo é a aplicação do conhecimento gerado. A pesquisa aplicada tem como características o interesse na aplicação, utilização e as ações práticas pelo conhecimento (Gil, 2008).

Seu objetivo apresenta caráter exploratório e descritivo. Uma pesquisa exploratória tem como finalidade a familiarização do problema por meio da análise de dados ou de observações empíricas (Marconi & Lakatos, 2003). Quanto ao objetivo descritivo, sua conduta procura a caracterização e a determinação de fenômenos ou populações (Gil, 2008).

Considerando o ponto de vista exploratório, inicialmente, foi realizado um mapeamento sistemático, com o objetivo de identificar lacunas e tendências acerca do desempenho de ferramentas ETL em portais de transparência, descobrindo impactos, abordagens, cenários, métodos de pesquisa e dados bibliométricos gerais.

Do ponto de vista da pesquisa aplicada, e como forma de avaliar as ferramentas ETL, foi utilizada a metodologia pesquisa-ação. Esse método foi escolhido por apresentar caráter colaborativo e integrar prática e teoria, sendo, portanto, utilizado como forma de avaliação ativa de uma determinada área ou fenômeno (Filippo, 2011). A avaliação, na pesquisa-ação, foi feita a partir do desenvolvimento de um módulo ETL para um portal de transparência, comparando-o quali e quantitativamente com módulos desenvolvidos em ferramentas ETL amplamente usadas no mercado.

1.6 ORGANIZAÇÃO DA DISSERTAÇÃO

Este documento está organizado de acordo com a Instrução Normativa Nº 05/2019/PROCC, a qual permite que a Dissertação seja “uma compilação de artigos científicos submetidos ou publicados em veículos com *Qualis*, desde que seja contextualizada com seções de Introdução, Discussão, Conclusão e Referências, não limitada a estas”. São 5 capítulos que fornecem uma base conceitual para o entendimento sistêmico. Os tópicos a seguir descrevem o conteúdo de cada um dos capítulos:

- O capítulo 1 apresenta esta Introdução, explicando as justificativas, juntamente com os problemas levantados e a suposição de pesquisa;

- O capítulo 2 apresenta parte do Mapeamento Sistemático aceito e publicado na *25th International Conference on Enterprise Information Systems*. Com o título: *An Exploratory Analysis of the Use of ETL in Transparency Portals;*
- O capítulo 3 apresenta parte de um artigo submetido à Revista Gestão e Tecnologia (*Journal of Management & Tecnology*), resumindo todo o trabalho efetuado nesta pesquisa. São descritos Planejamento, Operação e Resultados da Pesquisa-Ação;
- O capítulo 4 traz uma síntese narrativa do Mapeamento Sistemático, juntamente com uma discussão das questões de pesquisa;
- Finalmente, no capítulo 5, é apresentada uma compilação de conclusões, contribuições e sugestões de trabalhos futuros.

CAPÍTULO 2 - MAPEAMENTO SISTEMÁTICO

Neste capítulo, será apresentada parte do artigo intitulado: *An Exploratory Analysis of the Use of ETL in Transparency Portals*.

Abstract: Context: Government transparency portals are built based on ETL (Extract, Transform and Load) processes, which increase the quality and interoperability of data, making a critical subsystem for these applications, subject to evaluative research for improvements. **Objective:** To analyze publications on the use of ETL in transparency portals, in order to characterize them in relation to their scenarios, impacts, empirical methods and general bibliometric data. **Method:** Using the PICO strategy (Population, Intervention, Comparison and Outcome), a systematic mapping of the literature was performed. **Summary of Results:** In a total of 204 publications researched, 25 works were selected, of which 40% present, as the main impact for the portals, the availability of support for the construction of loads through a graphical interface, followed by the possibility of connectivity between bases of heterogeneous data (27%) and the ability to monitor loads (22%). Regarding the real automation of loads and their quality control, respectively, only 8% and 3% of the works discussed the impacts of these characteristics. **Conclusion**: The research showed that the use of ETLs in transparency portals still lacks comparative and feasibility studies. In this sense, an existing challenge is the lack of research that carries out replications to consolidate and validate the works already published, evidenced by the scarcity of controlled experiments in the area. Finally, analyzes on the quality control of loads was an important gap identified.

Keywords: Open Data, E-government, Public Transparency, ETL, Data Pipeline.

2.1 INTRODUCTION

Although Brazil has adopted an open data policy since 2011, in accordance with the provisions of Law No. 12,527 - Law on Access to Information (LAI) (Brazil, 2011), access to transparency portals in the 26 Brazilian states, and Federal, shows that theory and practice are at different stages. LAI establishes that every municipality with more than ten thousand inhabitants must make government data available on a transparency portal (Brazil, 2011). For greater effectiveness, data must be made available following

the principles of the Open Government Data (OGD) standard, proposed by the World Wide Web Consortium (W3C) (Transparência Internacional, 2020).

The predominance in transparency portals of Brazilian capitals and states is the data availability in formats that do not follow the OGD principles, resulting in a lack of interoperability (Eberhardt & Silveira, 2018; Oliveira & Silveira, 2018; Bachtiar, Suhardi & Muhamad, 2020; Cenci, Fillotrani & Ardenghi, 2017). There are also inadequacies regarding the completeness, primacy, opportunity, and accessibility of the data available on these portals (Oliveira & Silveira, 2018). This is a problem that deserves attention, as the reuse of this data, in initiatives that promote population engagement, becomes a major technological challenge (Bachtiar, Suhardi & Muhamad, 2020; Muller, Gil-Garcia & Tirelli, 2018; Dahbi, Lamharhar & Chiadmi, 2018). In December 2020, the Federal Government's Anti-Corruption Plan was published, bringing the implementation of 142 actions distributed among prevention, detection, and accountability mechanisms (Tian et al., 2021). These actions also aim to improve public transparency and promote population engagement. However, it is a necessary technological structure that makes possible their executions.

In this context, one of the critical points for these transparency portals creation is the data loading, in large institutions, is usually spread over several heterogeneous systems with the most varied types of databases (Pan, Zhang & Qin, 2018). In other words, the data need to be Extracted (E) from these source systems, Transformed (T), and Loaded (Load – L) in integrated databases that will be accessed by the portals (Sun & Lan, 2012). These three tasks gave names to the systems that perform them, they are the ETL Systems (Extract, Transform, and Load) (Sun & Lan, 2012).

Portals need to build and/or acquire these systems, considering impacts such as connectivity to all types of bases, ease and agility of use through a graphical interface, control and data quality assurance, automation, and monitoring of loads (Sreemathy, Infant Joseph, Nisha, Chaaru & Gokula, 2020). In addition, ETL systems are also directly involved with creating metadata, publishing data and metadata on the web, as well as creating appropriate catalog records (Saranya et al., 2021).

Given this need, in this article, we present a systematic mapping to characterize articles in relation to the performance of ETL tools in transparency portals, identifying impacts, approaches, scenarios, research methods, and general bibliometric data.

2.2 SYSTEMATIC MAPPING PLANNING

2.2.1 OBJECTIVE

This mapping had as its principal objective to identify and characterize the impacts caused by the ETL Systems in transparency portals use.

2.2.2 RESEARCH QUESTIONS

The research questions were developed to present an area overview, highlighting fundamental aspects of primary studies (Kitchenham, 2004; Petersen, Vakkalanka & Kuzniarz, 2015). They were prepared based on the PICO model (Bergin & Wraight, 2006; Santos, Pimenta & Nobre, 2007), to highlight the effects of an intervention in a given population and structure the research into four fundamental elements: Population, Intervention, Control, and "Outcomes" (Results). These elements, according to Santos, Pimenta & Nobre (2007), can be used to build research questions of different natures. Table 1 illustrates the PICO model used in this work.

Table 1: PICO model for research question compliance

Acronym	Category	Description
P	Population	Publications that directly address government transparency portals, with or without open data.
I	Intervention	Development and/or application of ETL tools to optimize the process of Extraction, Transformation, and Loading of portals.
C	Control	Articles on Transparency Portals that do not use tools developed specifically for ETL processes. **Articles that fit the intervention:** Linked Pipes *ETL* in use: Practical publication and consumption of Linked Data (Klímek & Skoda, 2017); A Content-Driven *ETL* Processes for Open Data (Berro & Teste, 2015). **Control article:** Analysis of the energy service in non-interconnected zones of Colombia using business intelligence (Colmenares-Quintero, 2021).
O	Result	Automation, scheduling, orchestration, and monitoring of ETL processes, prioritizing quality through clean and accurate data, as well as ease through graphical interfaces.

Thus, from the PICO model definition, research questions were elaborated, based on the guidelines of the Systematic Literature Mapping protocol observed in (Kitchenham, 2004; Petersen, Vakkalanka & Kuzniarz, 2015). They are: Q1: What methods are used in research on ETLs that deal with transparency portals?; Q2: What are

the approaches used for ETL?; Q3: What are the most used data extraction, transformation, and loading tools within the transparency portals scope?; Q4: What are the portal ownership scenarios?; Q5: What publication types or forums have addressed the ETL issue in the public transparency context?; Q6: Which countries have the most researchers who have published in this area?; Q7: Which years had the most publications in this area?; Q8: What are the impacts of using ETL on Transparency Portals?

2.2.3 SEARCH AND SELECTION STRATEGY

To search for articles, the databases responsible for publishing the main journals in the Computer Science field were selected, namely, ACM Digital Library (ACM), IEEE Xplore (IEEE), SCOPUS, and Science WEB. The searches were carried out using the filtering tools available in each database, considering the searches: title, abstract, and keywords. Access to the databases was performed through the CAPES journal portal (Capes, 2021) using an institutional subscription without any article restrictions.

For digital databases search, a search string was defined and composed of English terms and synonyms, associated with transparency portals and advantages of the applying ETL. The identified terms from the papers of the PICO model and the control articles are defined in Table 1, and later refined and adapted for better string use. Table 2 presents the terms before refinement.

Table 2: Terms before refinement

Category	**Description**
Population	Open government data portals, open government, e-government, public sector information, OGD, OGD portals, open government data, digital government, e-services, government data, transparency, government accountability, government transparency, open government data ecosystem, open data platforms, open data
Intervention	Extraction, transformation, and loading; *ETL*; Extract, transform, load; Load Procedures; Extraction Procedures; Load Software; Extraction Software; Load Program; Extraction Program.
Control	-
Result	Automation; Orchestration; Scheduling; Quality; Connectivity; Graphic Interface; Graphical User Interface; User Interface; Ease; Facilitate; Monitoring; Alerts; Broadcasting.

After refinement, the adjusted terms were used to build the search string, which is described in Table 3.

Table 3: Strings chosen after refinement

Search string terms		
Population	Intervention	Result
open data, open government, e-government, digital government, public transparency, government transparen, electronic governmeny, government accountability	Extraction, transformation and loading, ETL (extract, transform and load)	Orchestr*, gui, graphical user interface, eas*, facilit*, monitor*, alert*, warn*, connectivity, schedul*, quality

From the terms highlighted above, the following search string was created:

TITLE-ABS-KEY((*"open government" OR "open data" OR "e-government" OR "digital government" OR "public transparency" OR "government transparency" OR "electronic government" OR "government accountability") AND ("extraction, transformation and loading" OR "etl" OR "extract, transform and load") AND ("orchestr*" OR "gui" OR "graphical user interface" OR "eas*" OR "facilit*" OR "monitor*" OR "alert*" OR "warn*" OR "connectivity"OR "schedul*" OR "quality"))*

2.2.4 SOURCE SELECTION CRITERIA

Inclusion and exclusion criteria were established to filter articles relevant to systematic mapping. Only the studies selected for evaluation, which passed the inclusion and exclusion criteria, were counted.

The inclusion criteria are presented below:

1. Articles that are available online and in digital libraries;
2. Articles that contain the search String in the title, abstract, or keywords;
3. Articles that identify and characterize existing studies on the use of ETL tools in transparency portals;
4. The articles must have a publication date between the years 2011 and 2022. This period, significant and above 10 years, was chosen based on Law No. 12,527 - Law on Access to Brazilian Information (LAI), of 2011;
5. Articles written in English.

The exclusion criteria were:

1. Duplicate studies;
2. Surveys*;*
3. Systematic review;
4. Preliminary studies;
5. Short Papers.

2.2.5 INFORMATION EXTRACTION STRATEGY

To assess the work quality and answer the research questions set out in section 2.2, tables 4 and 5 were designed, which served as a guide for researchers to have no doubts about the data to be extracted for each article read in the whole. According to (Kitchenham, 2004), data extraction forms should be designed to collect all the information necessary to address the questions and quality criteria of the study.

Table 4: Impacts identified and extracted from the articles

Impacts	Definition
Connectivity	Providing connectivity to unstructured data and cloud data sources.
Ease of use through Graphical Interface	Resource savings with developers through an intuitive, code-free environment for extracting, transforming, and loading data.
Quality	Quality control, to determine the data consistency, accuracy, and control.
Automation	Automation capabilities with task scheduling and process orchestration.
Monitoring	Process metadata storage, such as, for example, the load time and the number of records loaded, issuing alerts and warnings about deviations and the failure or success of loads to stakeholders, as well as providing a monitoring panel.

Table 5: Classification of Extracted Approaches

Approach	Description
Market ETL	Ready and commercially available open-source or closed-source ETL tools.
Developed ETL	ETL tool developed especially for the project.
Not Specified	The article did not provide details about the Tool.

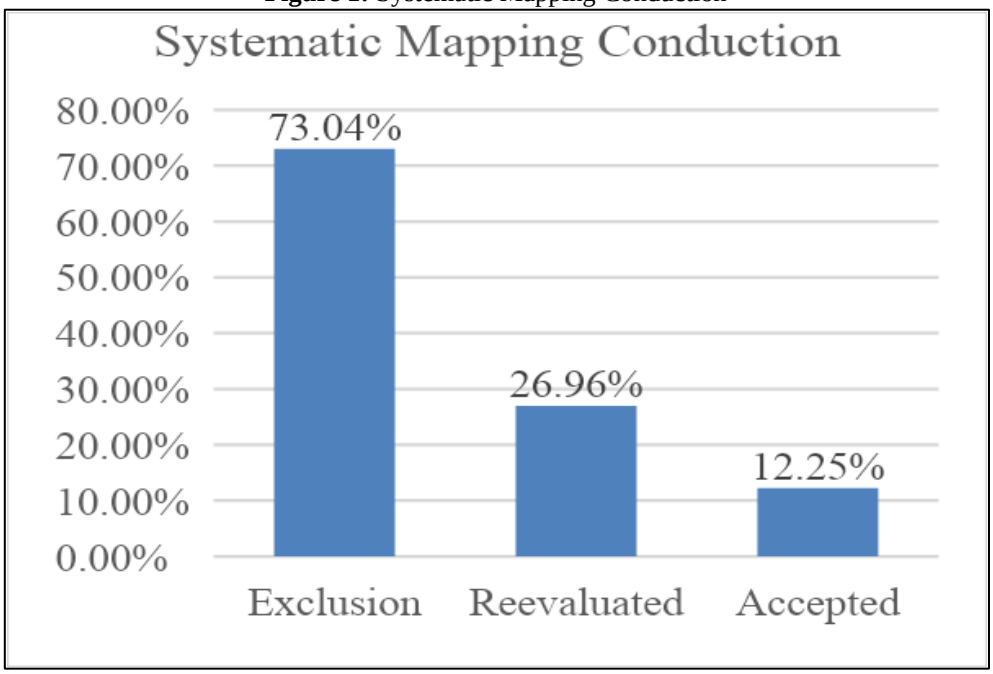
2.2.6 SYSTEMATIC MAPPING CONDUCTION

Responsible for publishing the major journals in the Computer Science field, the Scopus database (Yu et al., 2019) was chosen as the basis for defining and refining the

search string. This base also includes articles from distinct scientific databases, such as IEEE, ACM, and Web of Science. After being defined, refined, and judged adequate, the string was translated to the other search engines used in this work, which were: IEEE. ACM and Web of Science. In total, 204 works were returned, 24 (12%) from Scopus, 15 (7%) from Web of Science, 22 (11%) from IEEE, and 143 (70%) from ACM. The data are represented in Figure 1.

After searching the articles in the databases, the filtering process began based on the selection criteria defined in section 2.4. Each paper was classified as accepted or rejected. Of the 204 analyzed publications, 149 (73.04% of the total) were in accordance with the exclusion criteria. After removing these articles, the remaining works were read completely, 55 publications, 26.96% in relation to the total of publications. After being analyzed, following the inclusion criteria, 25 publications were selected, that is, 12.25% of the total will be analyzed. Figure 1 shows the summary of this step.

Figure 1: Systematic Mapping Conduction



Source: Prepared by the authors, 2023

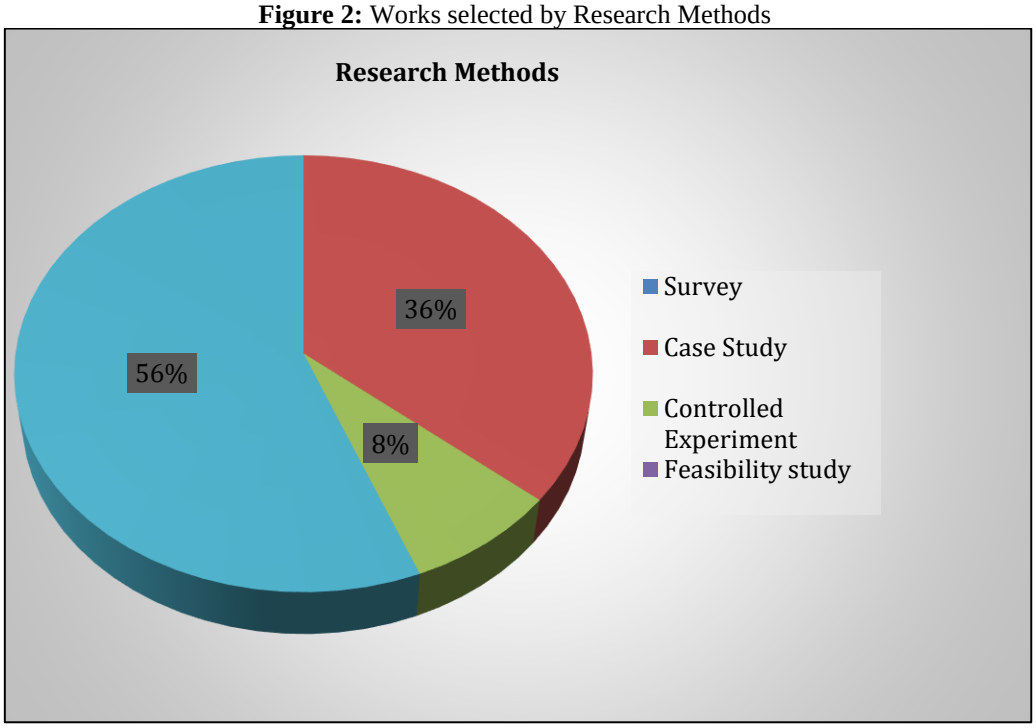
2.3 DATA SYNTHESIS AND RESULT PRESENTATION

In this section, the systematic mapping result will be presented according to the extracting process of the obtained articles and answering the research questions according to the extracted data.

What methods are used in research on ETLs dealing with transfer portals?

Figure 2 presents the methods used in research on ETLs dealing with transparency portals. The highlight goes to Exploratory Studies, with 56% of the works.

Figure 2: Works selected by Research Methods

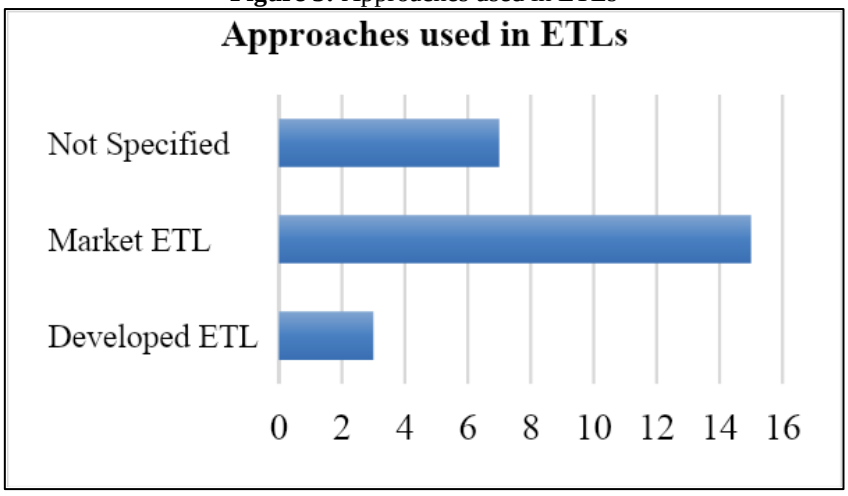


Source: Prepared by the authors, 2023

What are the approaches used for ETL?

Figure 3 presents the approaches used. We can see that 60% of the works use the ETL tools on the market.

Figure 3: Approaches used in ETLs



Source: Prepared by the authors, 2023

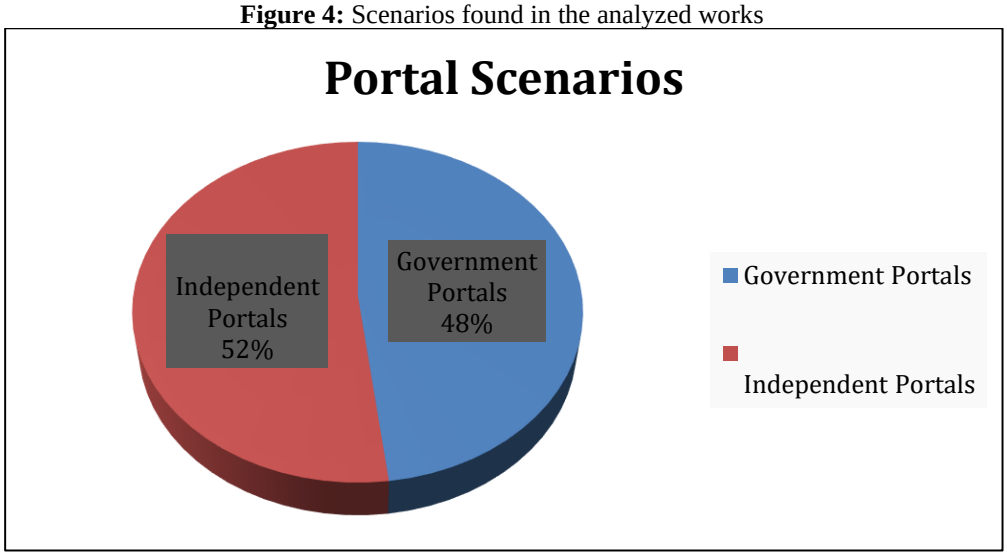
What are the most used data extraction, transformation, and loading tools within the scope of transparency portals?

Among the works analyzed, 28% did not specify the tool used. The instrument most cited in the works was Kettle, from the open-source software Pentaho, used in 32% of the analyzed papers. The second most quoted tool was Talend Open Studio, used in 16% of the analyzed works. The third most cited tool, with 12%, was Linked Pipes ETL. The other most representative were ODET (4%), Cylon (4%), and SpatialETL (4%).

What are the portal ownership scenarios?

Of the scenarios found in the works, 52% of them are independent portals, that is, they can be portals for universities and NGOs (Non-Governmental Organizations) but use government data. We present the scenarios found in the works in percentages in Figure 4.

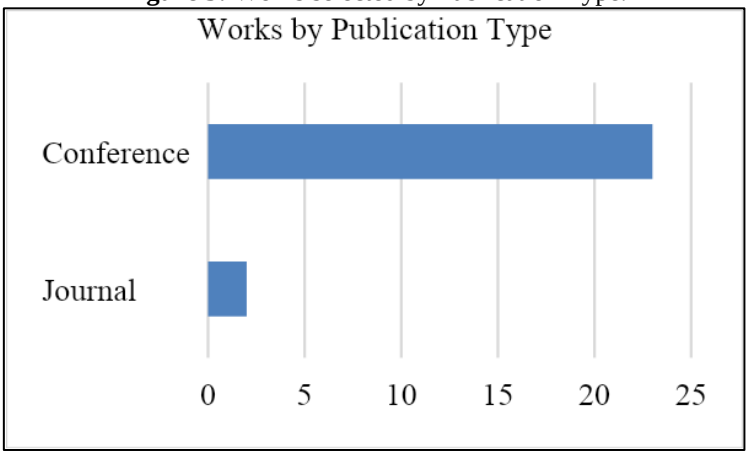
Figure 4: Scenarios found in the analyzed works



Source: Prepared by the authors, 2023

What publication types or forums have addressed the ETL issue in the context of public transparency?

Figure 5: Works selected by Publication Type.

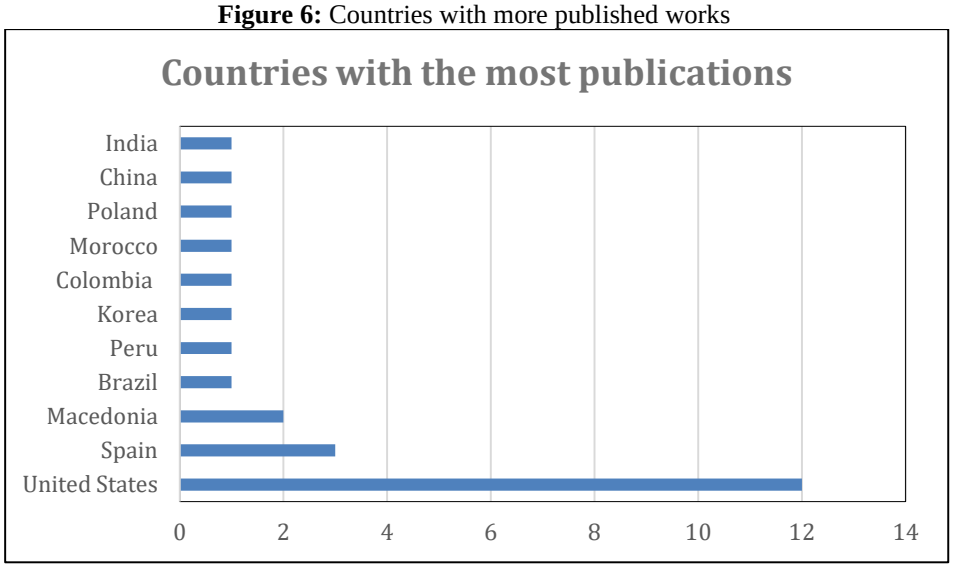


Source: Prepared by the authors, 2023

Which countries have published the most in this area?

Figure 6 shows the countries where the analyzed works were published. The United States appears with 48% of published works, being the country with the most publications in the area, followed by Spain, with 12%, and Macedonia, with 8%.

Figure 6: Countries with more published works

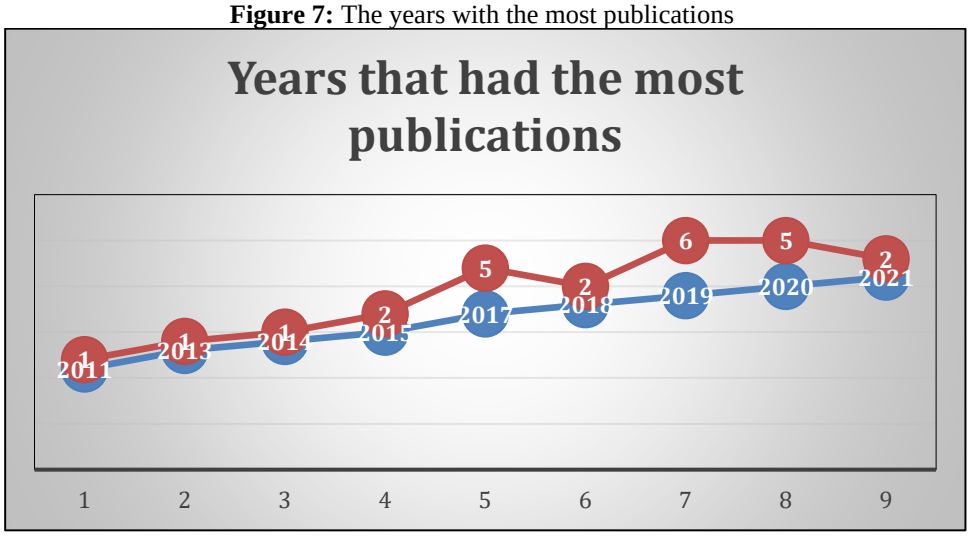


Source: Prepared by the author, 2023

Which years had the most publications in this area?

As shown in Figure 7, the year in which more works were published was 2019, with six publications.

Figure 7: The years with the most publications

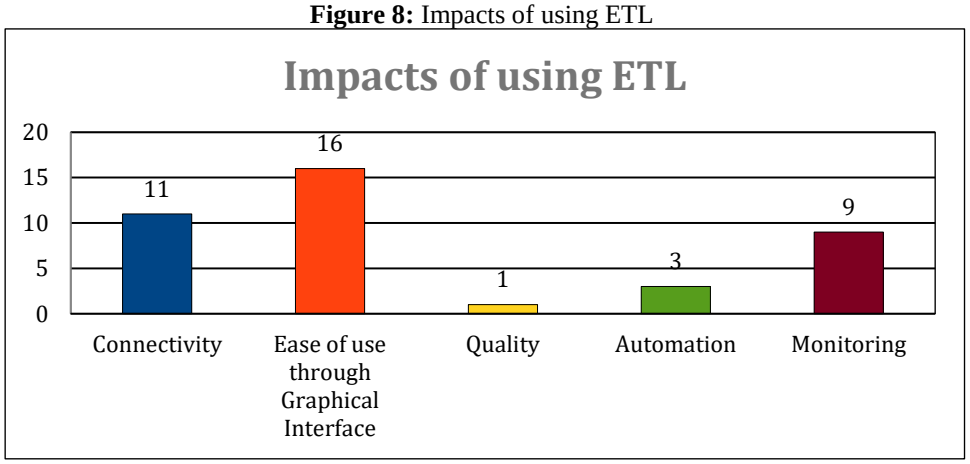


Source: Prepared by the author, 2023

What are the impacts of using ETL on Transparency Portals?

Figure 8 shows the most evident impacts in the works. Ease of use through a Graphical Interface was the most obvious impact, representing 40% of the analyzed papers.

Figure 8: Impacts of using ETL



Source: Prepared by the author, 2023

2.4 THREATS TO VALIDITY

Threats to validity may limit the ability to interpret and/or describe data results obtained in a study (Chapetta, 2006). Therefore, some threats should be taken into account:

- **Construct Validity:** The search string and elaborated research questions span the entire related studies area. To mitigate this threat, an attempt was made to develop

a string based on identified and refined terms with the help of control articles guided by the PICO model.

- **Internal Validity:** To mitigate data extraction and characterization problems, forms were designed to be filled in by the publication analysis and each analysis and extraction were reviewed by two researchers.

CAPÍTULO 3 - ANÁLISE EXPLORATÓRIA E COMPARATIVA DO USO DE *ETL* EM PORTAIS DE TRANSPARÊNCIA

Neste capítulo, será apresentado parte do artigo intitulado: Análise Exploratória e Comparativa do Uso de *ETL* em Portais de Transparência.

RESUMO: ***Contexto*****:** Os portais de dados abertos são construídos com base em processos *ETL* (*Extract, Transform and Load*), os quais aumentam a qualidade e interoperabilidade dos dados, perfazendo um subsistema crítico para estas aplicações, passível de pesquisas avaliativas para melhorias. ***Objetivo***: Este artigo teve como objetivo geral desenvolver e avaliar um módulo *ETL* para um portal de transparência, comparando-o qualitativamente com módulos desenvolvidos em duas ferramentas *ETL* amplamente usadas no mercado. Adicionalmente, foi feita uma análise das eficiências dos procedimentos de carga gerados pelos 3 tratamentos avaliados. ***Método***: Foi executada uma Pesquisa-Ação para construção de procedimentos *ETL* do Anuário Econômico de Sergipe. As ferramentas avaliadas durante o processo de desenvolvimento foram: (1) *Pentaho Data Integration - Kettle, Open Source,* e (2) *SQL Server Integration Services - SSIS, Closed Source,* contra (3) um código ETL construído na linguagem *Python*. ***Resultados:*** Foram encontradas evidências de destaque da ferramenta *Kettle,* do ponto de vista da usabilidade e eficiência de desenvolvimento por meio de interface gráfica, bem como do ponto de vista da curva de aprendizagem. Na sequência, vieram a linguagem de programação *Python* e a ferramenta *SSIS*. Em relação à eficiência, a mensuração do tempo de carga mostrou um melhor desempenho da linguagem *Python*, seguida do *Kettle* e do SSIS. ***Conclusão:*** Definidas as prioridades contextuais de portais de transparência, como por exemplo a eficiência das cargas ou a eficiência de desenvolvimento, a avaliação sistematizada de soluções disponíveis tal como a proposta neste artigo norteia situações de *trade-off* e seleção do melhor custo-benefício.

Palavras-Chave: Portais de Transparência, ETL, eficiência, usabilidade, qualidade.

3.1 TRABALHOS RELACIONADOS

Não foram encontrados artigos com relação direta com o trabalho aqui apresentado, todavia, a seguir, são destacas três publicações que abordam avaliações de ferramentas e processos *ETL* específicos.

O artigo "*Python-Based ETL Process for Data Integration*", de Islam e Rahman (2021) concentra-se especificamente na utilização da linguagem *Python* no processo de integração de dados, destacando os benefícios e vantagens dessa abordagem. Ele fornece informações sobre como a linguagem *Python* pode ser empregada de maneira eficaz na construção de cargas nos portais de dados abertos, contribuindo para a eficiência e qualidade dos processos.

O artigo "*Comparative Analysis of ETL Tools for Big Data Processing*", de Ansari e Ali (2021), realiza uma análise comparativa de ferramentas *ETL* para o processamento de big data. São avaliadas diferentes ferramentas, excetuando o *Kettle* e o *SSIS*, fornecendo uma visão abrangente das características, eficiência e desempenho dessas ferramentas. Esse estudo foi particularmente relevante para a pesquisa-ação pois, apesar de ter sido publicado após a realização da seleção das ferramentas utilizadas neste trabalho e não abordar especificamente portais de transparência, seus resultados serviram para nortear as análises comparativas apresentadas na seção de resultados.

Por sua vez, o artigo "*A Comparative Study of ETL Tools for Data Warehousing*", de Erol e Demirel (2021) apresenta uma pesquisa comparativa de ferramentas *ETL* para *Data Warehousing*. São analisadas diversas ferramentas, investigando aspectos como funcionalidades, desempenho e recursos oferecidos por essas ferramentas. Essa pesquisa também contribuiu para o comparativo das características e diferenças entre as ferramentas.

3.2 BASE CONCEITUAL

3.2.1 SISTEMAS ETL

Os sistemas *ETL* são utilizados para mover e transformar dados de fontes múltiplas, carregando-os em vários destinos. Um *ETL* deve ter escalabilidade para acompanhar o crescimento do volume de informações e, consequentemente, passará por manutenções. É um tipo de integração de dados em três etapas: Extração, Transformação

e Carregamento. Esses dados podem ser extraídos de diversos sistemas para inserção em bancos de dados históricos, comumente chamados de *Data Warehouses*, conforme descrito nos trabalhos de Senghane, Boly e Ndiaye (2021) e de Martínez e Galvis-Lista (2012).

Basicamente, a primeira etapa consiste em extrair os dados dos diversos sistemas de uma organização e conduzi-los para uma *staging area*, ou seja, uma área de transição na qual são efetuadas conversões e unificações de formato. Na segunda etapa de transformação, ocorre a limpeza, padronizações e correções de inconsistências. A última etapa envolve a transmissão e carga dos dados transformados para a base de dados históricos, a qual poderá ser acessada por diversos tipos de aplicações, tais como ferramentas de *Business Intelligence* (BI), Mineração de Dados, Inteligência Artificial e Sistemas de Apoio à Decisão. Tal descrição é encontrada nos trabalhos de Sun e Lan (2012) e de Adnan (2017).

3.2.2 CARACTERÍSTICAS E IMPACTOS DE UM SISTEMA ETL

A seguir, serão listadas as características essenciais e consequentes impactos causados pela adoção de ferramentas *ETL* para a manutenção de bases de dados históricas. Estes impactos foram extraídos de um mapeamento sistemático da literatura realizado nesta pesquisa, baseado nos estudos de Poletti, Colaço Júnior & Nascimento (2023) e podem servir para apontar os fatores de sucesso que podem ser generalizados para outros ambientes interdisciplinares que possuem cargas de dados.

3.2.3 CONECTIVIDADE

A conexão de uma ferramenta *ETL* com múltiplas fontes de dados, idealmente, deve ser transparente, ou seja, o acesso a bases estruturadas, desestruturadas e na nuvem não deve exigir codificação. Estas características são previamente implementadas e disponibilizadas nos pacotes de instalação, os quais permitem a conexão com dados oriundos de bases de dados diversas, tais como, por exemplo, oriundos de arquivos dos tipos json, xml e csv. Usando uma interface gráfica, característica melhor detalhada a seguir, o operador da ferramenta *ETL* precisará apenas selecionar parâmetros e definir os locais das bases, como explicam Souza, Abrantes e Lisboa-Filho (2021).

3.2.4 INTERFACE GRÁFICA

As interfaces gráficas utilizadas nas ferramentas *ETL* devem ser intuitivas e de fácil entendimento, separando componentes pré-definidos por funcionalidade e permitindo aos operadores o seu uso na construção do *ETL*.

As ferramentas também devem permitir tratamentos de exceção e construções de fluxos de execução, de forma visual, ou seja, por meio da seleção de componentes pré-definidos e ordenação destes em um painel gráfico. Este painel, o qual pode ser chamado de painel de execução, deve apresentar dados quantitativos sobre as etapas de carga em andamento e registros afetados. Quando essas características supracitadas são agregadas em um mesmo ambiente, levanta-se a hipótese de aumento de produtividade para construção de *ETL* (Pan, Zhang, & Qin, 2018).

3.2.5 AUTOMAÇÃO E MONITORAMENTO

As ferramentas *ETL* devem permitir a automação e monitoramento através da execução dos componentes definidos e visualizados no painel de execução. O *ETL* criado e a execução são monitoradas a cada sequência lógica, ou seja, no painel de execução mostra o monitoramento de cada registro por componentes no sequenciamento da execução, alertando de possíveis erros nos componentes em execução e criando logs, permitindo saber onde acontecem os erros. As ferramentas de mercado permitem o agendamento (*Schedule*). O sistema de agendamento deve ser capaz de suportar grande número de processos que são executados ao longo do dia em tempo real e a sua execução deve funcionar como um relógio, sem qualquer intervenção humana e sem falhas. Caso ocorra algum problema com os processos, as soluções de *ETL* devem ter a capacidade de comunicar diferentes grupos ou pessoas, dependendo do trabalho ou do tipo de falha. Essa comunicação, conforme Radhakrishna, Kiran, & Ravikiran (2012) pode ocorrer através de mecanismos de notificação ou *scripts* personalizados (Sreemathy et al, 2021)

3.2.6 GARANTIA E CONTROLE DE QUALIDADE

A garantia de qualidade das ferramentas *ETL* deve estar presente na etapa de transformação dos dados, limpando os dados redundantes, tratando as inconsistências e reportando dados inválidos. A forma de tratamento pode ser:

- Procura-se realizar as conversões dos dados corretamente;
- Nos casos de atributos nulos, substituí-los por valores pré-configurados;
- Validação e separação de valores aberrantes.

Em suma, essa etapa serve para atender aos requisitos com qualidade e controle, perfazendo os mapeamentos dos dados de entrada e provendo informações saneadas e acuráveis. Via de regra, um Sistema *ETL* eficaz tem componentes de entradas e saídas de dados que podem ser controlados na execução, ou seja, é possível saber como o dado entra e como o dado vai sair para o uso dos clientes, segundo Biplob, Sheraji, & Khan (2018).

3.3 METODOLOGIA

O estudo foi desenvolvido sob uma perspectiva metodológica de aplicação do conhecimento gerado, também chamada de pesquisa aplicada, a qual prioriza a utilização e as ações práticas, conforme orientações de Gil (2008).

Do ponto de vista do tipo de pesquisa, o estudo foi feito de modo exploratório e descritivo, tendo como objetivo comparar a eficiência e impactos principais das ferramentas *ETL* selecionadas, por meio da condução de uma pesquisa-ação. A pesquisa exploratória, segundo Lakatos e Marconi (2003), tem como finalidade a familiarização do problema por meio da análise de dados ou de observações empíricas. Em relação ao objetivo descritivo, sua conduta procura a caracterização e a determinação de fenômenos ou populações (Gil, 2008).

Neste sentido, a pesquisa-ação culminou com a realização de um estudo de caso em ambiente real, o qual avaliou as ferramentas. O ambiente foi o *Data Warehouse* do projeto intitulado Transparência Traduzida — Monitoramento Social da Economia e dos Atos Públicos em Sergipe, elaborado pelos Departamentos de Economia e de Sistemas de Informação da Universidade Federal de Sergipe (2023), o qual tem como principal objetivo fornecer informações a respeito de atos públicos, características da população

brasileira e da economia, sempre apoiadas por análises de especialistas, dotando o cidadão comum e agentes responsáveis por políticas públicas com conhecimento para poder controlar, fiscalizar, sugerir iniciativas e tomar decisões que venham a beneficiar a população em geral.

Os tratamentos analisados foram as ferramentas *ETL* (1) *Pentaho Data Integration - Kettle, Open Source,* (2) *SQL Server Integration Services - SSIS, Closed Source,* e (3) um código *ETL* construído na linguagem *Python*. Os comparativos foram abordados essencialmente de maneira qualitativa, no entanto, do ponto de vista da eficiência, os tempos de cargas das 3 abordagens foram medidos. A abordagem qualitativa, de acordo com Gerhartd e Silveira (2009), compreende aspectos da realidade que não podem ser quantificados, tais como valores sociais, motivações e aspirações, bem como a avaliação preliminar de um produto ou processo em desenvolvimento.

O *Kettle* foi selecionado pelo mapeamento sistemático. O *Talend*, segundo colocado no mapeamento, deu lugar ao SSIS, para que um dos objetivos específicos fosse cumprido: ferramentas amplamente utilizadas no mercado. Em tempo, o SSIS é a 4ª ferramenta com maior participação de mercado, à frente do *Talend* (Gartner, 2022).

Em relação aos procedimentos técnicos, este trabalho propôs a criação de módulos *ETL* como objetos de avaliação, no contexto do Anuário Econômico de Sergipe, subprojeto do portal Transparência Traduzida supracitado, disponível no site Transparência Traduzida (http://www.transparenciatraduzida.ufs.br/). A seguir, são descritos os passos utilizados na avaliação:

1 – Foi estabelecido o conjunto de requisitos *ETL* para atender o site Transparência Traduzida. Esses requisitos serviram para a criação de 3 processos de carga, selecionados pela complexidade e leitura de tipos de dados variados;

2 – Foram estabelecidos os aspectos lógicos (tipos de dados, e.g. data) das informações contidas nas bases de dados. Neste passo, foi feita a modelagem da informação, replicando os requisitos;

3 – As cargas foram construídas para avaliação, em suas três instâncias, uma para cada ferramenta avaliada, conforme o passo 1. Foram criados cenários para que a equipe do departamento de Sistemas de Informação da UFS pudesse avaliar as eficiências e dificuldades de manutenção dessas cargas. Além disso, também foram mensurados os tempos das cargas executadas e foi criado um questionário para avaliação da ferramenta por parte dos desenvolvedores.

A seguir, de forma autocontida, a metodologia da pesquisa-ação e sua execução será melhor detalhada.

3.4 DEFINIÇÃO E PLANEJAMENTO DA PESQUISA-AÇÃO

3.4.1 DEFINIÇÃO DO OBJETIVO

Este trabalho tem como objetivo geral a concepção e avaliação de três cenários de cargas de um módulo *ETL* do Anuário Econômico de Sergipe (UFS, 2023), desenvolvido pelos Departamentos de Economia e Sistemas de Informação da Universidade Federal de Sergipe (UFS), disponível no site Transparência Traduzida.

3.4.2 PLANEJAMENTO: QUESTÕES DE PESQUISA E FORMULAÇÃO DE QUESTÕES

O problema de pesquisa a ser averiguado é a eficiência das cargas e a facilidade de implementação ao utilizar ferramentas amplamente utilizadas para gerar e automatizar processos *ETL*, em comparação com a codificação específica em linguagem de programação. Nesse contexto, os seguintes questionamentos devem ser considerados:

a) RQ1: A implementação de processos *ETL* por meio de ferramentas semiautônomas aumenta a produtividade?

b) RQ2: As ferramentas *ETL* de mercado podem facilitar a implementação e reduzir o tempo necessário?

c) RQ3: Qual abordagem gera cargas mais eficientes em termos de desempenho e qualidade dos dados?

Essa pesquisa busca averiguar as vantagens e desafios associados ao uso de ferramentas de *ETL* em relação à codificação manual. Os resultados obtidos serão relevantes para auxiliar profissionais e organizações na escolha da abordagem mais adequada às suas necessidades e na melhoria dos processos de *ETL*.

3.4.3 SELEÇÃO DE OBJETOS E PARTICIPANTES

Os participantes das avaliações foram 3 desenvolvedores que já trabalham nas construções dos *ETL* que alimentam as informações do Anuário Econômico de Sergipe, ou seja, a amostra foi estratificada, com participantes que possuem as características da população de interesse. Os desenvolvedores possuem idades entre 24 e 41 anos, todos formados em Sistemas de Informação ou Ciências da Computação, com tempo de experiência de 2 a 8 anos na implementação de processos *ETL*.

Do ponto de vista dos procedimentos de carga avaliados com relação ao desenvolvimento, foram selecionados dois, com níveis de complexidade diferentes. Um considerado complexo pelos programadores, com diversas transformações de dados, e um simples, totalmente direto. No primeiro procedimento, a partir daqui, chamado de PC1, apresentamos informações referentes à Lavoura Permanente: Evolução da Razão Entre Área Colhida e Destinada à Colheita de Todos os Produtos, no período de 2010 até 2019, abrangendo os contextos do Brasil, Nordeste e Sergipe. Os dados resultantes da execução do processo *ETL* estão disponíveis para análise por meio do seguinte link: http://www.transparenciatraduzida.ufs.br/anuario/pages/2_agricultura/grafico_2_8.html.

O segundo procedimento, chamado de PC2, refere-se às informações sobre a Evolução do Comércio Ampliado, no período de 2010 até 2020, abrangendo os contextos do Brasil, Nordeste e Sergipe. Os dados resultantes da execução do processo *ETL* estão disponíveis para análise por meio do seguinte link: http://www.transparenciatraduzida.ufs.br/anuario/pages/8_comercio/grafico_8_2.html.

3.4.4 PROJETO DA PESQUISA-AÇÃO

A pesquisa-ação adotou o modelo processual como referência, aproveitando sua flexibilidade para não impor delimitadores rígidos entre as etapas. Essa abordagem, construída por Filippo (2011) permitiu que algumas etapas fossem executadas simultaneamente, agilizando o processo de construção e avaliação do módulo *ETL*. Foram realizados 3 ciclos de pesquisa-ação, detalhados a partir da seção 4.

3.4.5 INSTRUMENTAÇÃO

Foram adotadas práticas mistas entre geração de código manual e códigos gerados por meio de ferramentas específicas para construção de *ETL*. Foram utilizadas as seguintes ferramentas:

a) *Pentaho Data Integration*, versão 8.1, conhecida como *Kettle*, com a versão da plataforma Java *Standard Edition 8 Development kit* (JDK 8), versão 1.8.0_333;

b) Microsoft Visual Studio 2022, utilizando o projeto *Integration Services*, conhecido como *SSIS*;

c) *ETL* codificado manualmente com a linguagem *Python*, para construção do *ETL* via código;

d) Sistema Gerenciador de Banco de Dados (SGBD) *PostgreSQL, versão 11.0*.

Com relação ao hardware utilizado, trata-se da 8ª Geração de Processadores Intel® Core™ i7, com 16 GB de RAM. Além disso, foi utilizado o questionário de avaliação descrito na seção a seguir.

3.4.6 SURVEY

As avaliações foram feitas de forma individual para cada participante. Foi realizada uma apresentação de cada ferramenta, esclarecendo o seu objetivo e explicando as suas funcionalidades. Após as apresentações, cada processo *ETL* citado na seção 3.1.5 foi construído e executado para efetuar a carga para o *Data Warehouse*. Ato contínuo, cada desenvolvedor teve um prazo para responder ao *survey*.

O questionário possui 9 questões de pesquisa na escala likert-5, 1 pergunta descritiva e 6 perguntas na escala de 1 (para melhor ferramenta) a 3 (para pior ferramenta), totalizando 16 perguntas. As questões foram distribuídas em três contextos específicos: (1) aspecto técnico (9 questões), refere-se às tecnologias, design (interface gráfica), recursos tecnológicos e conectividade; (2) comparativo entre ferramentas (6 questões), está relacionado ao aprendizado e à eficiência da ferramenta; (3) ao final, o participante deveria responder à única questão descritiva, já mencionada, indagando sobre possíveis alterações e melhorias. As questões são listadas a seguir:

a) SQP1-Você considera o design (interface gráfica) da plataforma da ferramenta *kettle* adequado para o uso de *ETL*?

b) SQP2-Você considera o design (interface gráfica) da plataforma da ferramenta *SSIS* adequado para o uso de *ETL*?

c) SQP3-Você considera os *steps* da ferramenta *Kettle* eficientes para o uso de *ETL*?

Um step ou etapa é um componente de uma transformação em um processo ETL e pode oferecer uma ampla gama de funcionalidades. Essas funcionalidades podem variar desde a leitura de arquivos de texto até a implementação de técnicas de dimensões lentamente alteradas (slowly changing dimensions). Cada step representa uma ação específica que é executada durante o processo de transformação dos dados, permitindo a manipulação e preparação dos dados para a carga em um destino final, como um Data Warehouse ou Data Mart. Os steps podem incluir operações como filtragem, limpeza, agregação, junção de tabelas, cálculos, entre outros, dependendo das necessidades do processo ETL.

d) SQP4-Você considera os *steps* da ferramenta *SSIS* eficientes para o uso de *ETL*?

e) SQP5-Desconsiderando a curva de aprendizado, qual método teve maior eficiência de desenvolvimento? Ordene sua resposta usando 1, para o melhor, e 3 para o pior.

- *Kettle* [ ]
- *SSIS* [ ]
- Construção do Código em *Python* [ ]

f) SQP6-Qual método tem a menor curva de aprendizado? Ordene sua resposta usando 1, para o menor, e 3 para o maior.

- *Kettle* [ ]
- *SSIS* [ ]
- Construção do Código em *Python* [ ]

g) SQP7-Qual método gerou o processo de carga mais rápido? Ordene sua resposta usando 1, para o melhor, e 3 para o pior.

- *Kettle* [ ]
- *SSIS* [ ]

Construção do Código em *Python* [ ]

h) SQP8-Você concorda que a ferramenta *Kettle* tem infraestrutura de conectividade que permite a interface com ambientes heterogêneos de dados abertos?

i) SQP9-Você concorda que a ferramenta *SSIS* tem infraestrutura de conectividade que permite a interface com ambientes heterogêneos de dados abertos?

j) SQP10-Em qual método houve a menor necessidade de uso de lógica de programação? Ordene sua resposta usando 1, para o menor, e 3 para o maior.

- *Kettle* [ ]
- *SSIS* [ ]
- Construção do Código em *Python* [ ]

k) SQP11-Em qual método houve menor uso de *SQL*? Ordene sua resposta usando 1, para o menor, e 3-para o maior.

- *Kettle* [ ]
- *SSIS* [ ]
- Construção do Código em *Python* [ ]

l) SQP12-As ferramentas *ETL* facilitam a construção de um DW (*Data Warehouse*) ou DM (*Data Mart*)?

m) SQP13-Você considera que a plataforma *kettle* possui todos os recursos disponíveis para atender à demanda de desenvolvimento de um *ETL*?

n) SQP14-Você considera que a plataforma *SSIS* possui todos os recursos disponíveis para atender à demanda de desenvolvimento de um *ETL*?

o) SQP15-Em linhas gerais, qual foi o melhor método? Ordene sua resposta usando 1, para o melhor, e 3, para o pior.

- *Kettle* [ ]
- *SSIS* [ ]
- Construção do Código em *Python* [ ]

p) SQP16-Descreva abaixo melhorias ou novas funcionalidades que gostaria de ver nas ferramentas, especificando a ferramenta que possui a lacuna (opcional).

3.5 RESULTADOS

3.5.1 PRIMEIRO CICLO

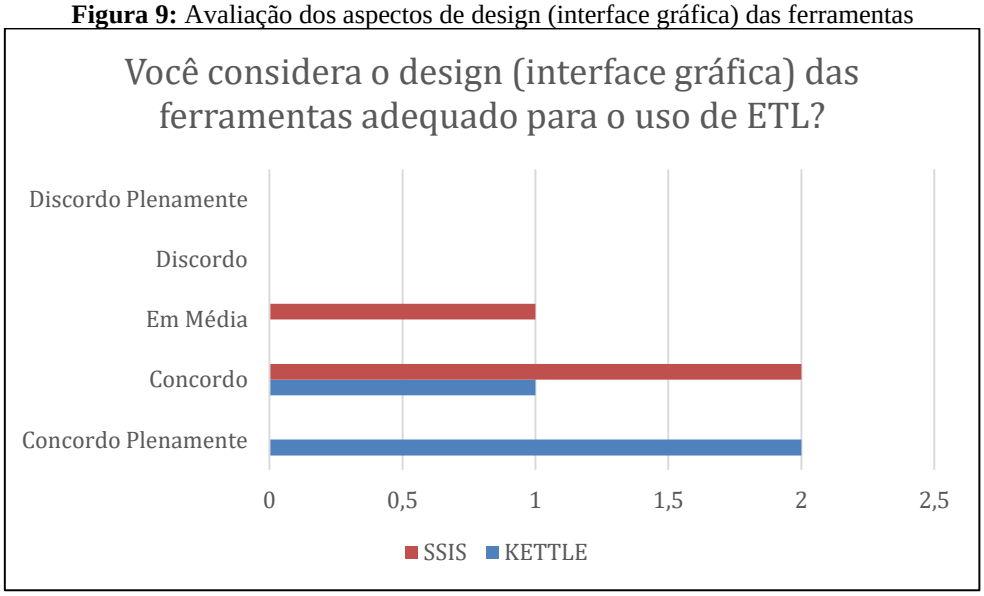
Foi realizada a construção dos procedimentos de carga na linguagem de programação *Python*, atendendo aos requisitos do Departamento de Economia da Universidade Federal de Sergipe. O processo de pesquisa-ação foi iniciado a partir desses procedimentos construídos, os quais também foram desenvolvidos no *Kettle* e no *SSIS*.

Durante a construção das cargas, foram realizadas reuniões, em que foram apresentados os requisitos técnicos, tais como: conectividade com a base de dados do *Data Warehouse*, construção de consultas *SQL (Structured Query Language*) para validar as cargas e transformações que os dados iriam sofrer para atender às inserções no *Data Warehouse*.

3.5.2 SEGUNDO CICLO

No segundo ciclo da pesquisa-ação, o diagnóstico foi conduzido pelos três desenvolvedores responsáveis pelas cargas, que participaram do *survey* e forneceram suas respostas. A seguir, apresentaremos uma explicação dos resultados obtidos com base nessas respostas.

Figura 9: Avaliação dos aspectos de design (interface gráfica) das ferramentas

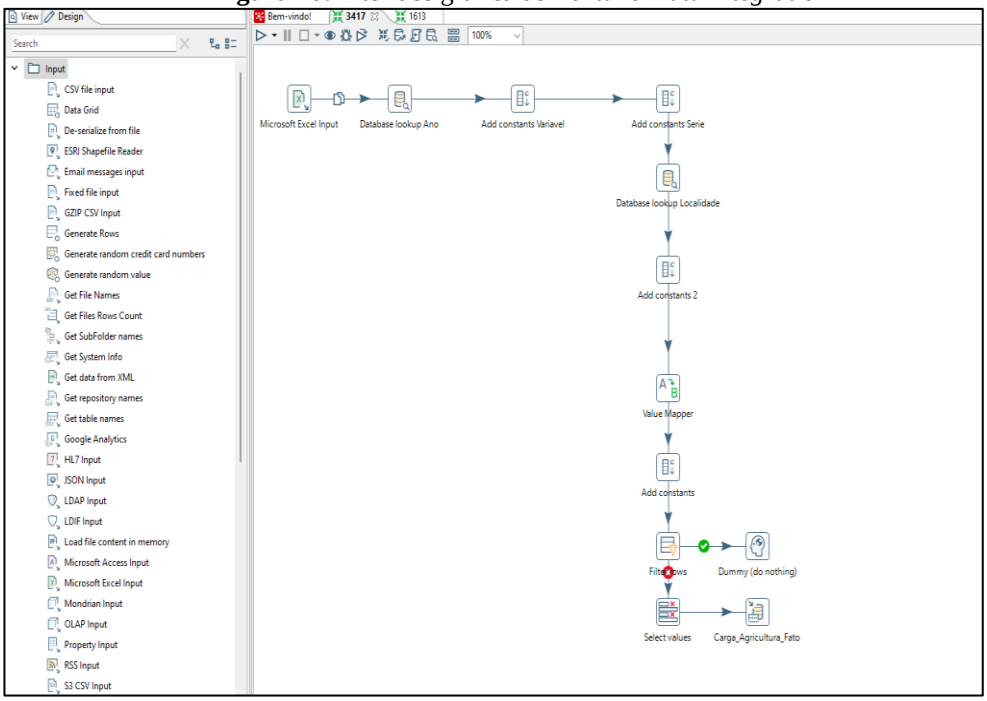


Fonte: Elaborado pelo autor, 2023

As questões SQP1 e SQP2 são comparativos entre as ferramentas, sob o aspecto de design (interface gráfica), e tiveram valores mensurados entre Concordo Plenamente, Concordo, Em Média, Discordo e Discordo Plenamente.

No caso do *Kettle*, vide Figura 9, duas respostas indicaram um nível alto de concordância com a afirmação de que sua interface gráfica é satisfatória (Concordo Plenamente), enquanto uma resposta mostrou um nível moderado de concordância (Concordo). Esses resultados sugerem que os desenvolvedores consideram o design do *Kettle* como positivo e eficiente para o desenvolvimento de *ETL* (vide Figura 10).

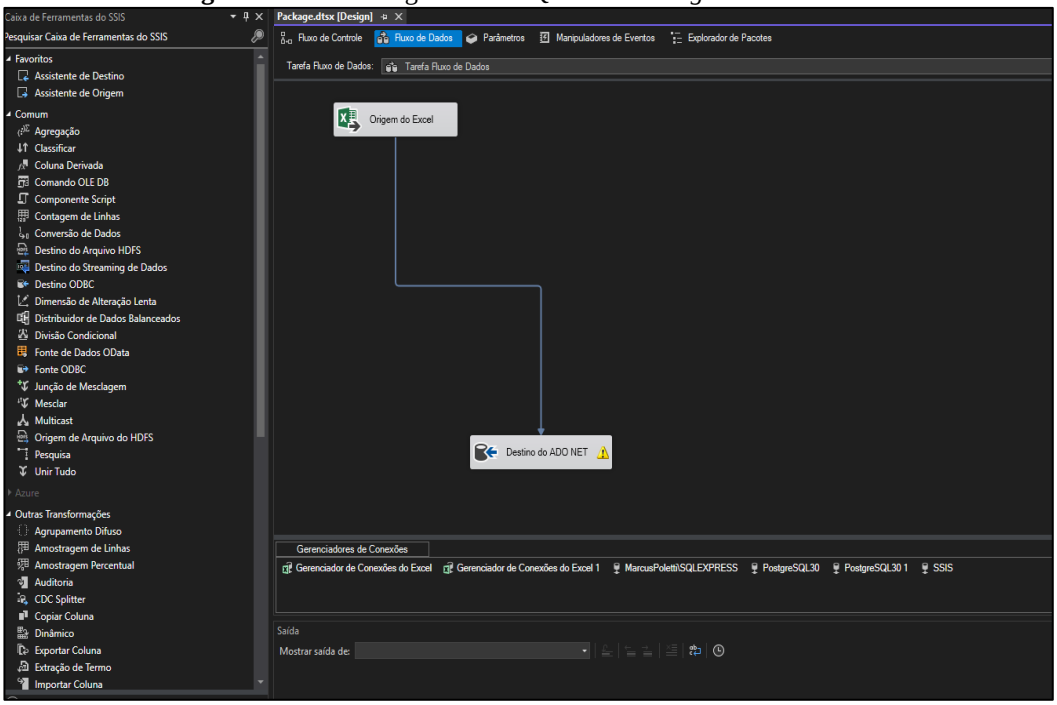
Figura 10: Interface gráfica do *Pentaho Data Integration*



Fonte: (*Pentaho Data Integration*, *version* 8.1, 2023)

Por outro lado, em relação ao *SSIS*, duas respostas indicaram concordância (Concordo) com a afirmação sobre sua interface gráfica (vide Figura 11), e uma resposta foi classificada como Em Média. Esses resultados sugerem que, embora o *SSIS* também tenha sido bem avaliado, a preferência dos desenvolvedores foi mais direcionada ao *Kettle*. Neste caso e daqui em diante, é importante considerar as experiências dos desenvolvedores, os quais desconheciam as ferramentas e possuíam experiência em Python. Além disso, o contexto específico e os requisitos exigidos influenciam a escolha da ferramenta mais adequada.

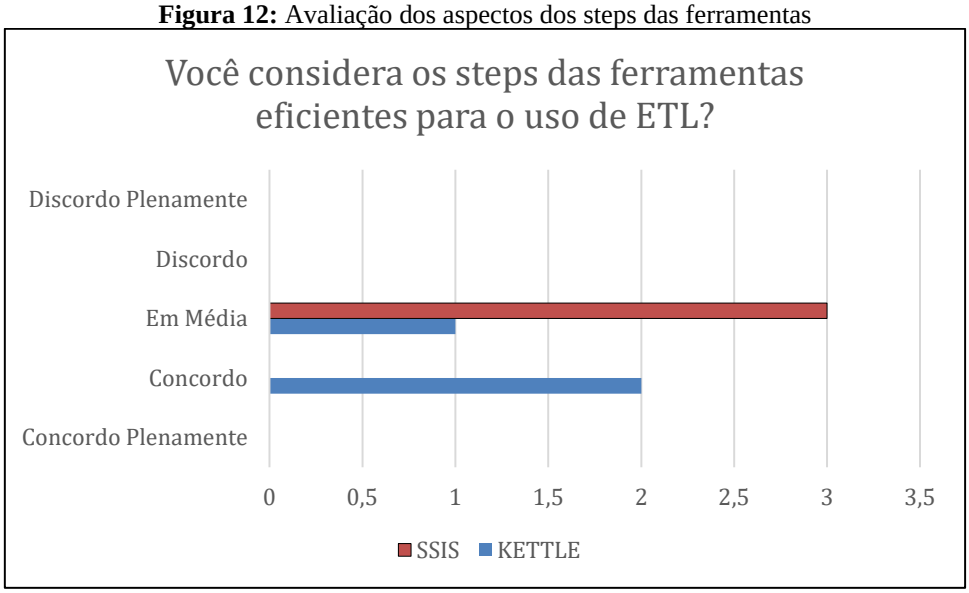
Figura 11: Interface gráfica do *SQL Server Integration Services*



Fonte: (*Microsoft Visual Studio Tools for Applications, version* 17.6.4, 2022)

Na Figura 12, são apresentados os resultados das perguntas SQP3 e SQP4, sobre avaliação dos *Steps* das ferramentas de *ETL*. A ferramenta *Kettle* recebeu duas classificações de "Concordo" e uma de "Em Média", enquanto a ferramenta *SSIS* obteve três classificações de "Em Média". Mais uma vez, a preferência foi pelo *Kettle.*

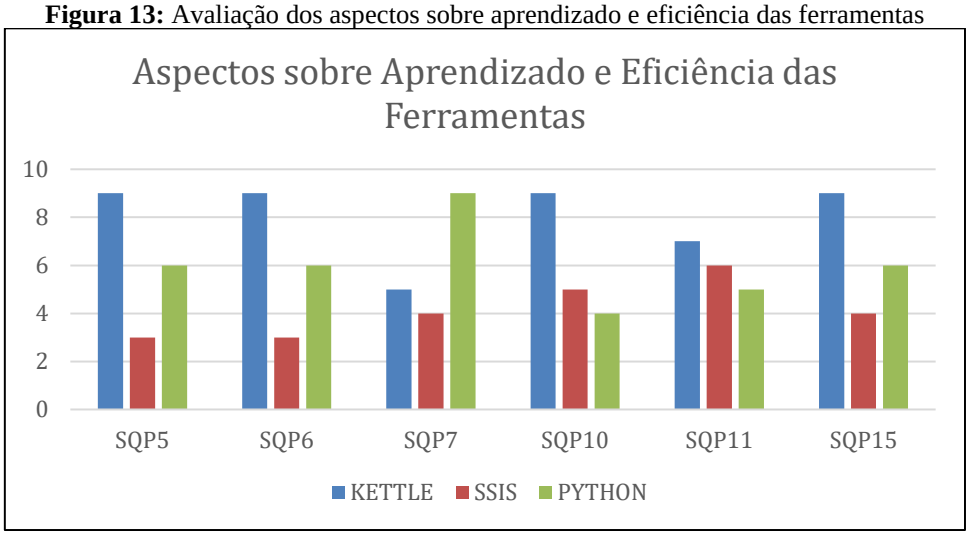
Figura 12: Avaliação dos aspectos dos steps das ferramentas



Fonte: Elaborado pelo autor, 2023

A Figura 13 apresenta os resultados das perguntas SPQ5, SPQ6, SPQ7, SPQ10, SPQ11 e SPQ15. Para a construção do gráfico, os valores de ordem foram invertidos, ou seja, a barra mais proeminente representa a abordagem que obteve a melhor avaliação.

Figura 13: Avaliação dos aspectos sobre aprendizado e eficiência das ferramentas



Fonte: Elaborado pelos autores, 2023

Na pergunta SPQ5, a eficiência no desenvolvimento foi analisada, e o *Kettle* obteve a melhor classificação, seguido pelo *Python* e pelo *SSIS*. Isso evidencia que os programadores se adaptaram melhor à interface *Kettle* e corrobora as respostas dadas nas questões anteriores.

A menor curva de aprendizagem foi avaliada na pergunta SQP6, na qual o *Kettle* também obteve a melhor classificação, seguido pelo *Python* e pelo *SSIS*. O *SSIS* recebeu uma classificação inferior devido à mudança de perspectiva nos resultados de execução do *ETL*, o que, segundo os avaliados, torna mais difícil a interpretação dos *Steps* executados. Mudar de perspectiva, no SSIS, implica em deixar o ambiente de desenvolvimento e abrir uma nova janela dedicada, chamada Visualizador de Execução, no SSIS Designer. Essa ferramenta oferece informações em tempo real sobre o progresso da execução de um pacote. Com o Visualizador de Execução, é possível acompanhar o status das tarefas e transformações, registros de erros e informações de desempenho, enquanto o pacote está sendo executado.

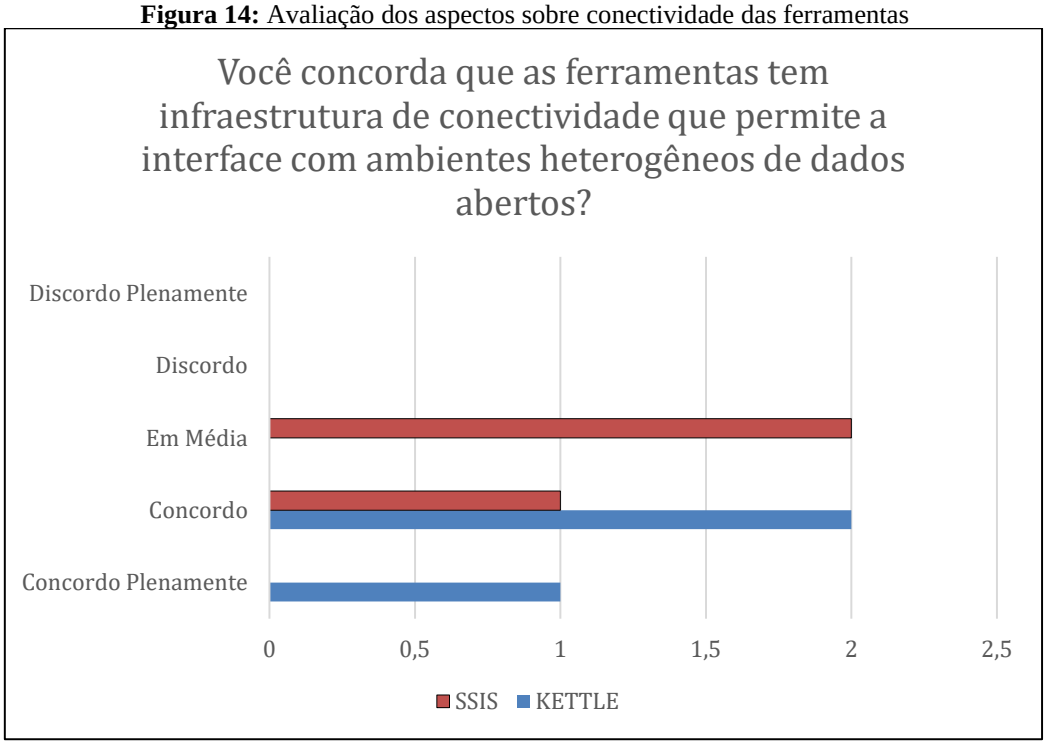
A eficiência das cargas foi o foco da pergunta SPQ7, e o *Python* recebeu a melhor classificação, seguido pelo *Kettle* e pelo *SSIS*. Nesse contexto, a eficiência das cargas refere-se ao processo de povoamento de um Data Warehouse, mensurando o tempo de carga de cada ferramenta. Todas as cargas foram executadas no mesmo hardware,

permitindo uma comparação direta entre as ferramentas. No terceiro ciclo, apresentamos todos os dados mensurados de cada ferramenta, os quais corroboraram as repostas aqui apresentadas.

A pergunta SPQ10 teve como objetivo avaliar as ferramentas que exigem menos lógica de programação. De acordo com o gráfico apresentado, o *Kettle* recebeu a melhor classificação, seguido pelo *SSIS* e pelo *Python*. Essa avaliação considerou que o *Kettle* e o *SSIS* se baseiam, principalmente, em desenvolvimento por *Steps*, enquanto o *Python* requer o desenvolvimento de lógica de programação em todo o processo. Neste mesmo sentido, o resultado se repetiu para a pergunta SPQ11, que analisou o menor uso de *SQL*.

Por fim, na pergunta SPQ15, os desenvolvedores foram solicitados a identificar a melhor ferramenta em geral. Segundo suas avaliações, o *Kettle* foi considerado a melhor, seguido pelo *Python* e pelo *SSIS*. A seguir, apresentamos as respostas das perguntas SPQ8 e SPQ9.

Figura 14: Avaliação dos aspectos sobre conectividade das ferramentas



Fonte: Elaborado pelos autores, 2023.

As perguntas SQP8 e SQP9 avaliaram a infraestrutura de conectividade das ferramentas com ambientes heterogêneos.

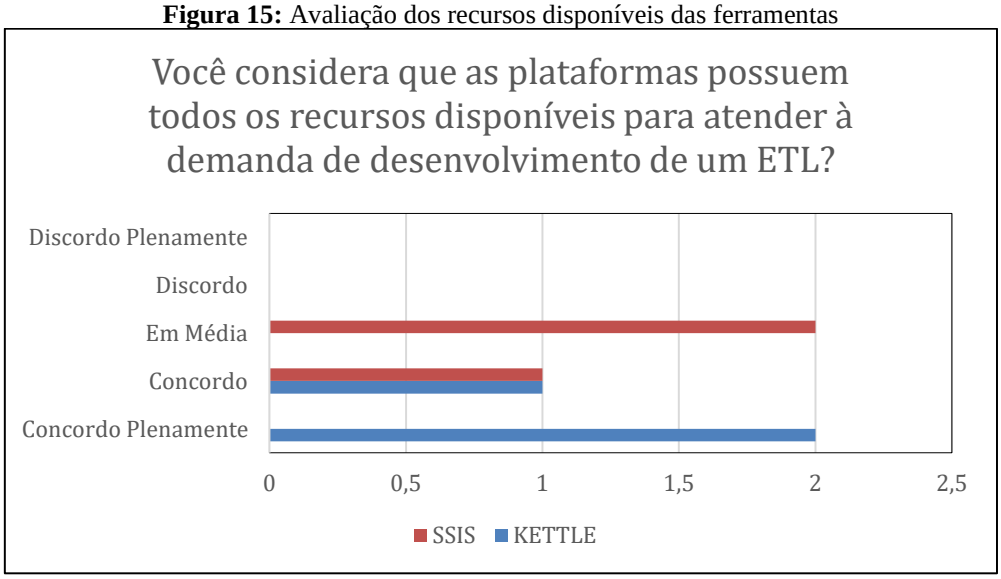
Na avaliação, a ferramenta *Kettle* recebeu uma classificação de "Concordo Plenamente" em uma resposta e "Concordo" em outras duas respostas. Por outro lado, a ferramenta *SSIS* foi classificada como "Concordo" em uma resposta e "Em Média" em

duas respostas. Segundo os programadores, esse resultado se deu pelo fato do *Kettle* suportar mais drivers pré-definidos para a conexão com tipos diferentes de bases de dados.

A pergunta SPQ12 avaliou se as ferramentas *ETL* em geral facilitam a construção de um *Data Warehouse* (*DW*) ou *Data Mart* (*DM*). Apesar da experiência prévia em Python, um desenvolvedor concordou plenamente e dois concordaram. Além disso, foi destacado que essas ferramentas permitem um acompanhamento em tempo real da execução do processo *ETL*, possibilitando a verificação dos dados passo a passo (*step by step*), bem como o monitoramento do tempo de execução em cada etapa. Isso facilita a identificação de problemas e o aprimoramento do processo como um todo.

Na Figura 15, são apresentados os resultados das perguntas SPQ13 e SPQ14. Dois respondentes "Concordam Plenamente" com o *Kettle* e um respondeu "Concorda". Em relação à ferramenta *SSIS*, apenas um respondeu "Concorda" e dois responderam "Em Média".

Figura 15: Avaliação dos recursos disponíveis das ferramentas



Fonte: Elaborado pelos autores, 2023

Na pergunta SPQ16, não tivemos sugestão de melhorias ou de novas funcionalidades para as ferramentas *ETL*.

3.5.3 TERCEIRO CICLO

No terceiro ciclo de pesquisa-ação, além da percepção inicial dos programadores para a execução, foram mensuradas e coletadas as eficiências (tempo de execução) das cargas e calculada a média do tempo de execução de 10 cargas para cada ferramenta, as quais foram executadas na máquina especificada na seção 3.1.5. A execução dos procedimentos, para o povoamento do Data Warehouse, contempla todas as fases de um ETL, desde a extração dos dados de origem, passando pelas transformações dos dados, até as cargas das tabelas de dimensões e fatos.

Para a carga PC1, a linguagem de programação *Python* foi mais eficiente, com a média de tempo de 9,9800 segundos, seguida do *Kettle,* com a média de tempo de 29,4700 segundos, e do *SSIS,* que ficou com a média de tempo de carga mais alta, perfazendo 41,9674 segundos. A Tabela 6 abaixo resume esses dados.

Tabela 6: Média de tempo de execução das cargas (PC1)

Média de Tempo de Execução das Cargas em Segundos	
Ferramentas	**Média de Tempo em Segundos**
Linguagem Python	9,9800
Kettle	29,4700
SSIS	41,9674

Fonte: Elaborado pelos autores, 2023

Na carga PC2, a linguagem de programação *Python* foi mais eficiente novamente, com a média tempo de carga de 1,4781 segundo. Em seguida, o *Kettle* teve uma média de tempo de carga de 2,2652 segundos, enquanto o *SSIS* apresentou a média de tempo de carga mais alto, com 4,6892 segundos. Esses dados estão ilustrados na Tabela 7. O terceiro ciclo direcionou a equipe para o desenvolvimento de novas cargas em *Kettle,* desde que o tempo de carga não se transforme em um impedimento para a disponibilização de informações em tempo satisfatório para os clientes. Em outras palavras, em casos de prevenção de danos de relacionamento com o cliente, as cargas continuarão sendo implementadas em Python.

Tabela 7: Média de tempo de execução das cargas (PC2)

Média de Tempo de Execução das Cargas em Segundos	
Ferramentas	**Tempo em Segundos**
Linguagem Python	1,4781
Kettle	2,2652
SSIS	4,6892

Fonte: Elaborado pelos autores, 2023

3.5.4 AMEAÇA À VALIDADE

3.5.4.1 AMEAÇAS À VALIDADE EXTERNA

A fim de mitigar os problemas relacionados à validade externa, a pesquisa foi conduzida com o público-alvo específico: os desenvolvedores do site Transparência Traduzida. Essa abordagem direcionada ao público relevante aumentou a aplicabilidade dos resultados obtidos, todavia, restrita às características dos programadores e cargas avaliados.

3.5.4.2 AMEAÇAS À VALIDADE DE CONSTRUÇÃO

Para mitigar as ameaças de construção, o projeto foi desenvolvido sob o modelo de pesquisa-ação, seguindo o ciclo de fases estabelecido pela literatura (Filippo, 2011).

CAPÍTULO 4 - DISCUSSÃO

Neste capítulo, será apresentada uma discussão dos resultados obtidos após a realização do Mapeamento Sistemático e das respostas às questões principais de pesquisa.

No mapeamento sistemático, a maioria dos artigos publicados sobre o tema ocorreu em 2019, o que indica que a área de pesquisa ainda é jovem. Os resultados também mostram que publicações relacionadas ao assunto estão presentes em vários países, o que indica que a busca por soluções para uso de *ETL* em portais de transparência está em expansão.

As evidências sobre controle e garantia de qualidade de *ETL* foram abordadas em apenas 3% dos estudos. Isso denota a necessidade de pesquisas sobre como são realizadas as transformações dos dados, sobre a quantidade de dados rejeitados e não validados nos processos de aprovação, os códigos-fonte gerados e como são realizadas validações dos valores transformados, principalmente antes do início da produção do portal.

As publicações também evidenciaram a ausência de avaliações e experimentos com ferramentas de código fechado, tais como, por exemplo, *Microsoft Integration Services* e *Oracle Cloud Infrastructure Data Integration*. Tanto comparativos entre ferramentas de código fechado quanto contra ferramentas de código aberto.

Por fim, os Experimentos Controlados representam apenas 8% dos trabalhos, além disso, os métodos de comparação (*comparison*) e estudo de viabilidade (*feasibility study*) não foram identificados nos trabalhos. Isto demonstra uma carência de pesquisas que realizem replicações para consolidação e validação de trabalhos, bem como a escassez de estudos experimentais com protocolos mais rigorosos que permitam estas replicações. Uma base de conhecimento sobre *ETLs* e sobre portais de transparência eficaz dependerá de um direcionamento das pesquisas para aplicação efetiva do método científico supracitado.

4.1 RESPOSTAS ÀS QUESTÕES DE PESQUISA

Para responder à primeira questão (RQ1), o Mapeamento Sistemático da Literatura, sintetizado na seção anterior, definiu uma lacuna de avaliações entre ferramentas em geral para o contexto da produtividade. Ato contínuo, foram selecionadas

ferramentas com grande participação no mercado para serem avaliadas, todavia, não confrontadas.

A avaliação geral das ferramentas E|TL sobre produtividade, nos artigos mapeados e também no *survey* respondido pelos Cientistas de Dados, evidenciou uma melhoria ou aumento de eficiência, devido ao uso de *steps* pré-configurados e parametrizáveis, interfaces intuitivas, conectividade com diversas fontes de dados, transformações pré-definidas e monitoramento centralizado.

No que diz respeito à segunda questão (RQ2), as ferramentas *ETL* de mercado também se mostraram mais eficientes para a implementação dos processos *ETL*, resultando em uma redução do tempo de implementação. Com recursos gráficos e pré-configurados, essas ferramentas podem agilizar a criação de fluxos de trabalho e automatização de tarefas.

Os resultados das respostas às questões SPQ1 e SPQ2 revelaram uma aceitação do design das ferramentas, incluindo suas interfaces gráficas, que foram consideradas de fácil entendimento e utilização (Figura 9). Nas questões SPQ3 e SPQ4, ficou evidente a utilidade dos *steps* pré-configurados nas ferramentas, simplificando a construção das cargas e contribuindo para a redução do tempo necessário para sua implementação. Essas características receberam uma avaliação positiva, sem nenhuma resposta em desacordo (Figura 10).

Após analisar as respostas às questões SPQ5, SPQ6, SPQ10, SPQ11 e SPQ15 do *survey*, que abordaram o aprendizado das ferramentas *ETL* avaliadas e uso em geral, foi observado que a ferramenta *Kettle* recebeu avaliações mais positivas. A linguagem de programação *Python* não foi bem avaliada nas questões SPQ10 e SPQ11, pois, por se tratar de uma linguagem de programação, não possui *steps* com funcionalidades pré-definidas que agilizam o processo *ETL* (Figura 11).

No que diz respeito à conectividade, as questões SPQ8 e SPQ9 confirmaram a facilidade de conexão com diversas bases de dados, graças à presença de interfaces de conexões pré-configurados nas ferramentas. Mais uma vez, não houve discordâncias nas respostas, indicando uma avaliação positiva por parte dos usuários (Figura 12).

Esses resultados também reforçam a eficácia das ferramentas *ETL* especificamente analisadas, com todas as respostas alinhadas com sua utilidade. As questões SPQ13 e SPQ14 demonstraram que essas ferramentas possuem todos os recursos necessários para atender às demandas de construção de cargas (Figura 13).

Para a terceira questão (RQ3), foi conduzida uma análise do tempo de carga nas ferramentas *ETL*. Os resultados revelaram que a linguagem de programação *Python* obteve o melhor desempenho em termos de tempo de carga, seguida pela ferramenta *Kettle* e *SSIS*. A questão SPQ7 corroborou esses tempos. Vale ressaltar que o tempo de carga foi medido em todas as ferramentas utilizando o mesmo hardware, o que possibilitou uma comparação direta. Os resultados dessa análise estão detalhados no capítulo da Análise Exploratória e Comparativa do Uso de *ETL* em Portais de Transparência.

CAPÍTULO 5 - CONCLUSÃO

O desenvolvimento de processos ETL em múltiplas plataformas é uma atividade complexa que exige recursos tecnológicos substanciais, um entendimento profundo de bases de dados e proficiência no uso das ferramentas apropriadas. Com o propósito de investigar e ressaltar as pesquisas mais relevantes nesse âmbito, um dos objetivos desta dissertação foi caracterizar a aplicação de ETLs em portais de transparência.

Dos 204 estudos recuperados das bases científicas, 25 foram selecionados conforme os critérios de inclusão, sendo notável que 24% deles foram publicados somente no ano de 2019. Isso evidencia uma tendência crescente na área, refletindo um aumento recente do interesse por parte de pesquisadores nesse tópico. No tocante às vias de disseminação, as conferências se destacaram, abrigando 23 trabalhos (92%), enquanto os periódicos compreenderam apenas 2 artigos (8%).

Identificaram-se lacunas consideráveis, destacando-se a necessidade de mais estudos relacionados à qualidade dos processos ETL, bem como a ausência de benchmarks que englobem ferramentas de código fechado amplamente utilizadas no mercado. Do ponto de vista metodológico, um desafio premente é a carência de pesquisas que conduzam replicações visando consolidar e validar os resultados previamente publicados, além da insuficiência de experimentos bem estruturados que possibilitem tais replicações. Por outro lado, os resultados obtidos nesta pesquisa demonstram que a aplicação de ETLs em portais de transparência é objeto de estudo em diversos países, permitindo adaptações das soluções aos variados contextos de governança e transparência.

Além disso, o mapeamento realizado nesta pesquisa traz contribuições relevantes para a academia, estabelecendo-se como uma fonte de referência que identifica as principais lacunas e tendências no uso de ETLs em portais de transparência.

Após as lacunas identificadas no mapeamento, empreendeu-se uma iniciativa com o objetivo inequívoco de aprimorar o processo de desenvolvimento e validação das ferramentas ETL. Para alcançar essa meta, a escolha recaiu sobre a adoção de uma abordagem de pesquisa-ação, na qual os próprios desenvolvedores que utilizam ferramentas ETL em suas rotinas diárias tiveram um papel central e proeminente. Em outras palavras, apesar da amostra de programadores ter sido pequena, esta foi do tipo estratificada, com participantes que possuem as características da população de interesse

e permitem inferências para o nicho aqui analisado. Além disso, também houve uma avaliação de desempenho das cargas desenvolvidas.

Em cada ciclo do processo, os desenvolvedores tiveram a oportunidade de avaliar as ferramentas e expressar suas opiniões por meio de um *survey*. Isso permitiu a mensuração da eficácia das ferramentas e confirmou a validade do modelo de pesquisa-ação. Os resultados obtidos evidenciaram a efetividade das ferramentas e sua relevância na prática profissional. Na perspectiva dos desenvolvedores envolvidos, a pesquisa-ação atendeu às demandas e confirmou os resultados observados no contexto prático.

Durante a avaliação, a ferramenta *Kettle* obteve um destaque geral, com curva de aprendizado menor que a da linguagem de programação *Python* e a da ferramenta *SSIS*. Essas conclusões foram confirmadas pelas respostas às questões SPQ1, SPQ2, SPQ3, SPQ4, SPQ5, SPQ6, SPQ8, SPQ9, SPQ13, SPQ14 e SPQ15.

Em termos de eficiência, a conclusão do questionamento SPQ7 é que a ferramenta desenvolvida em *Python* obteve o menor tempo de execução, seguida pelo Kettle e pelo *SSIS*. A coleta dos tempos em ambiente controlado corroborou as respostas dos entrevistados.

Finalmente, nos trabalhos futuros, pretende-se incluir mais ferramentas *ETL* nas pesquisas, explorando diferentes ambientes de *Business Intelligence* (*BI*) e auxiliando novos desenvolvedores na seleção das ferramentas mais adequadas para suas tarefas. Essas contribuições têm como objetivo fornecer informações relevantes para o estudo de *ETL*, tanto no meio acadêmico quanto na prática profissional.

REFERÊNCIAS

ADNAN, A. A. I. S. U. (2017). Performance analysis of extract, transform, load (ETL) in apache Hadoop atop NAS storage using ISCSI. **4th International Conference on Computer Applications and Information Processing Technology** (CAIPT).

ANSARI, M. A.; ALI, A. (2021). Comparative Analysis of ETL Tools for Big Data Processing. **International Journal of Advanced Science and Technology**, 30(8), 3323-3332. [Link: https://sersc.org/journals/index.php/IJAST/article/view/27130].

BACHTIAR, A.; MUHAMAD. Literature Review of Open Goverment Data. **International Conference on Information Tecnology Systems and Inovation (ICITSI).** Bandung – Pandang, October 19 – 23, 2020. ISBN: 978-1-7281-8196-7.

BERGIN, S.; WRAIGHT, P. Silver based wound dressings and topical agents for treating diabetic foot ulcers (Review). **Cochrane Database of Systematic Reviews** (Online) 2006 (01 2006), CD005082. https://doi.org/10.1002/14651858.

BIPLOB, M. B.; SHERAJI, G. A.; KHAN, S. I. (2018). Comparison of Different Extraction Transformation and Loading Tools for Data Warehousing. **Conferência Internacional sobre Inovações em Ciência, Engenharia e Tecnologia (ICISET)**.

BRASIL. **Lei Federal n. 12.527 de 18 de novembro de 2011.** Regula o acesso a informações previsto no inciso XXXIII do art. 5º no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Recuperado de: < https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>.

CAPES (2021). **Portal Periódicos CAPES/MEC** [Journal Portal CAPES/MEC]. Disponível em: <https://ww.periodicos.capes.gov.br>. Acesso em: 01 dez. 2021.

CENCI, K.; FILLOTRANI, P.; ARDENGHI, J. Government Data Interoperability: a Case Study from Academia. In: **Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance (ICEGOV'17).** Association for Computing Machinery, New York, NY, USA, p. 625-628, 2017. DOI: https://doi.org/10.1145/3047273.3047382.

CHAPETTA, W.A. Uma Infraestrutura para Planejamento, Execução e Empacotamento de Estudos Experimentais em Engenharia de Software. Ph.D. Dissertation. **Programa de Engenharia de Sistemas e Computação.** COPPE/UFRJ. Universidade Federal do Rio de Janeiro: Rio de Janeiro, 2006.

COMITÊ INTERMINISTERIAL DE COMBATE À CORRUPÇÃO. **Plano Anticorrupção:** Diagnóstico e Ações do Governo Federal. Disponível em: <https://www.gov.br/cgu/pt-br/anticorrupcao/plano-anticorrupcao.pdf>

DAHBI, K.Y.; LAMHARHAR, H.; CHIADMI, D. Exploring dimensions influencing the usage of Open Government Data Portals. In **Proceedings of the 12th International**

Conference on Intelligent Systems: Theories and applications, Rabat Morocco, October, 2018.

DIOUF, P.S.; BOLY, A.; NDIAYE, A. (2018) Variety of data in the ETL processes in the cloud: State of the art. **IEEE International Conference on Innovative Research and Development (ICIRD)**.

EBERHARDT, A.; SILVEIRA, M. S. Show me the Data! A Systematic Mapping on Open Government Data Visualization. In: **Proceedings of the 19th Annual International Conference on Digital Government Research**, May 30-June 1, 2018, Delft, Netherlands. ACM, New York, NY, USA, 10 pages. DOI: https://doi.org/10.1145/3209281.3209337.

EROL, S.; DEMIREL, B. (2021). A Comparative Study of ETL Tools for Data Warehousing. In **Proceedings of the 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies** (pp. 1-4). IEEE. [Link: https://ieeexplore.ieee.org/document/9535737].

FILIPPO, D. Pesquisa-ação em sistemas colaborativos. In: **Sistemas Colaborativos**: Livro-Texto da SBC. Rio de Janeiro: Elsevier Editora Ltda, 2011.

GARTNER, (2022). **Microsoft nomeada líder no Gartner® Magic Quadrant™ de 2022 em ferramentas de integração de dados.** Disponível em: https://azure.microsoft.com/en-us/blog/microsoft-named-a-leader-in-2022-gartner-magic-quadrant-for-data-integration-tools/.

GERHARTD, T. E.; SILVEIRA, D. T. Métodos de Pesquisa. Porto Alegre: Editora da UFRGS, 2009.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. São Paulo: Atlas, 2008.

INTERNATIONAL TRANSPARENCY. **Página inicial**. Disponível em: <https://www.transparenciainternacional.org.br/ipc/>. Acesso em: 03 de junho de 2021a.

INTERNATIONAL TRANSPARENCY. **Índice de Percepção da Corrupção 2020**. Disponível em: <https://comunidade.transparenciainternacional.org.br/ipc-indice-de-percepcao-da-corrupcao-2020>. Acesso em: 03 junho de 2021b.

INTERNATIONAL TRANSPARENCY. **Retrospectiva Brasil 2020**. Disponível em: < https://www.transparenciainternacional.org.br/ipc/>. Acesso em: 03 de junho de 2021c.

KITCHENHAM, B. **Procedures for Perfoming Systematic Reviews.** Keele, UK, Keele Univ. 33 (08 2004).

LAKATOS, E. M.; MARCONI, M. de A. **Fundamentos da Metodologia Científica.** 5. ed. São Paulo: Atlas S.A., 2003.

MARTÍNEZ, A.B.; GALVIS-LISTA, E. A. (2012). Modeling techniques for extraction transformation and load processes: A critical review. **Euro American Conference on Telematics and Information Systems (EATIS)**.

MULLER, P; GIL-GARCIA, J.; TIRELLI, C. The Impact of Political, Technological and Social Variables on the Development of Local Egoverment: Lessons from Brazil. In

Proccedings of the 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland, April-2018.

OLIVEIRA, E. F. De; SILVEIRA, M. S. Open Government Data in Brazil: A Systematic Review of its Uses and Issues. In **Procedings of 19th Annual International Conference on Digital Government Research,** May 30-June 1, 2018.

PAN, B.; ZHANG, G.; QIN, X. (2018). Design and realization of an ETL method in business intelligence Project. **IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)**.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and Software Technology** 64 (08 2015).

POLETTI, M.; COLAÇO JUNIOR, M.; NASCIMENTO, A. (2023). An Exploratory Analysis of the Use of ETL in Transparency Portals. In: **Proceedings of the 25th International Conference on Enterprise Information Systems** - Volume 1, ISBN 978-989-758-648-4, ISSN 2184-4992, pages 91-98.

RADHAKRISHNA, V;. SRAVANKIRAN, V.; RAVIKIRAN, K. (2012). Automating ETL process with scripting technology. **Conferência Internacional de Engenharia da Universidade Nirma (NuiCONE)**.

SANTOS, C. M. da C.; PIMENTA, C. A. de M.; NOBRE, M. R. C. A estratégia PICO para a construção da pergunta de pesquisa e busca de evidências. **Revista Latino-Americana de Enfermagem,** 15, 3 (jun 2007), p. 508-511. https://doi.org/10.1590/s0104-11692007000300023.

SARANYA, N.; BRINDHA, R.; AISHWARIYA, N.; KOKILA, R.; MATHESWARAN, P; POONGAVI, P. (2021). Data Migration using ETL Workflow. **7th International Conference on Advanced Computing & Communication Systems (ICACCS)**.

SOUZA, E. da S.; ABRANTES, L. A.; LISBOA-FILHO, J. (2021). ETL Process in a Federal Educational Institution: Obtaining Functional Information and Geolocation of Retired Server. **16ª Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI).**

SREEMATHY, J.; BRINDHA, R.; SELVA NAGALAKSHMI, M.; SUVEKHA, N.; KARTHICK RAGUL, N.; PRAVEENNANDHA, M. (2021). Overview of ETL Tools and Talend-Data Integration. **7th International Conference on Advanced Computing and Communication Systems (ICACCS).**

SUN, K.; LAN, Y. (2012). Projeto de Engenharia e Informatização da Manufatura. **3ª Conferência Internacional de Ciência de Sistemas.**

SUN, K.; LAN, Y. (2012). SETL: A Scalable and High Performance ETL System. **3rd International Conference on System Science, Engineering Design and Manufacturing Informatization.**

TIAN, Q.; HAN, Z.; YU, P. Application of openEHR archetypes to automate data quality rules for electronic health records: a case study. BMC Med Inform Decis Mak 21, 113, 2021.

UNIVERSIDADE FEDERAL DE SERGIPE. (S.D.). **Anuário Socioeconômico de Sergipe.** Disponível em: <http://www.transparenciatraduzida.ufs.br/anuario/pages/index.html >.

YU, J.; QIAO, Y.; SHU, N.; SUN, K.; ZHOU, S.; YANG, J. Neural Network Based Transaction Classification System for Chinese Transaction Behavior Analysis. In **2019 IEEE International Congress on Big Data (BigDataCongress),** p. 64-71, 2019. DOI: https://doi.org/10.1109/BigDataCongress.2019.0021.