UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

JACKSON CARDOSO

ENGENHARIA DE DADOS NO APOIO À ANÁLISE DOS EGRESSOS DE UM CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

CURITIBA

2023

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

ENGENHARIA DE DADOS NO APOIO À ANÁLISE DOS EGRESSOS DE UM CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Data engineering in supporting the analysis of graduates of a bachelor's degree program in information systems

Dissertação de mestrado profissional apresentada como requisito para obtenção do título de Mestre no Programa de Pós-graduação em Computação Aplicada da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Alexandre Reis Graeml.

CURITIBA

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba

JACKSON CARDOSO

ENGENHARIA DE DADOS NO APOIO À ANÁLISE DOS EGRESSOS DE UM CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Computação Aplicada da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia De Sistemas Computacionais.

Data de aprovação: 19 de Outubro de 2023

Dr. Alexandre Reis Graeml, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Nadia Puchalski Kozievitch, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Renata Mendes De Araujo, Doutorado - Universidade Presbiteriana Mackenzie (Mackenzie)

Dedico este trabalho à Deus,
minha esposa e família.

"Toda a decisão certa é proveniente de experiência. E toda a experiência é proveniente de uma decisão errada."
(Albert Einstein)

"In God we trust; all others must bring data."
(W. Edwards Deming)

RESUMO

Considerando a 'fragmentação/dispersão de dados' e a necessidade de suporte para analisar o desempenho e a empregabilidade dos egressos, bem como o desejo de estabelecer a engenharia de dados, este trabalho tem como objetivo explorar o papel da engenharia de dados no apoio à tomada de decisões durante a reestruturação do curso de bacharelado em Sistemas de Informação em uma universidade pública brasileira. Utilizando técnicas de processamento analítico de dados, integrando informações de diversas fontes (sistema da universidade, LinkedIn e formulários web), e disponibilizando indicadores-chave de desempenho (KPIs) para monitoramento pelos decisores (coordenador e colegiado do curso), buscou-se criar uma ferramenta de suporte para o planejamento estratégico do curso. Essa ferramenta apresenta recursos visuais intuitivos e interativos baseados em nuvem, proporcionando uma análise abrangente e facilitando a interpretação dos dados. A intenção é impulsionar a inteligência de dados da organização, possibilitando a tomada de decisões fundamentadas em dados de maneira colaborativa. A pesquisa exploratória abordou conceitos essenciais e contribuições sobre o tema, seguindo um processo de engenharia de dados aplicado à análise de egressos como metodologia para a tecnologia e geração de conhecimento. O artefato resultante foi demonstrado ao grupo de usuários em um ambiente de TI corporativo, obtendo feedback sobre sua eficácia para subsidiar a tomada de decisões.

Palavras-chave: engenharia de dados; inteligência coletiva; *analytics*; *big data*; tomada de decisão.

ABSTRACT

Considering the 'fragmentation/scattering of data' and the need for support in analyzing the outcomes and employability of graduates, along with the motivation to outline data engineering, this work aims to explore the use of data engineering in supporting the decision-making process during the restructuring of the bachelor's degree program in Information Systems at a Brazilian public university. Utilizing analytical data processing techniques, integrating information from various sources (university system, LinkedIn, and web forms), and providing key performance indicators (KPIs) for monitoring by decision-makers (coordinator and course board), a support tool was created for the strategic planning of the program. This tool features intuitive, cloud-based visual resources, enabling comprehensive data exploration and analysis. The goal is to foster the organization's data intelligence, enabling collectively informed decision-making. The exploratory research addressed essential concepts and contributions on the subject, following a data engineering process applied to the analysis of graduates as a methodology for technology and knowledge generation. The resulting artifact was demonstrated to the user group in a corporate IT environment, obtaining feedback on its effectiveness in supporting decision-making.

Keywords: data engineering; collective intelligence; analytics; big data; decision-making.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ACID	*Atomicity, Consistency, Isolation, Durability*
AWS	*Amazon Web Services*
BI	*Business Intelligence* (inteligência de negócio)
BSI	Bacharelado em Sistemas de Informação
BSC	*Balanced Score Cards*
CEO	*Chief Executive Officer*
DBA	*Database Administrator*
DBaaS	*Database as a Service*
DBMS	*Database Management System*
DDDM	*Data-Driven Decision Making*
DSS	*Decision Support System*
DLH	*Data Lakehouse*
DW	*Data Warehouse*
ED	Engenharia de Dados
ETL	*Extract, Transform and Load*
HD	*Hard disk* (disco rígido)
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
KPI	*Key Performance Indicator*
ML	*Machine Learning* (aprendizado de máquina)
NDE	Núcleo Docente Estruturante
OLAP	*On-line Analytical Processing*
OLTP	*On-line Transaction Processing*
SI	Sistema de Informação
SQL	*Structured Query Language*
UOM	*Unit Of Measure*
UTFPR	Universidade Tecnológica Federal do Paraná

SUMÁRIO

1 INTRODUÇÃO

No dia a dia empresarial, decisões acabam sendo tomadas quase a todo momento, exigindo que os gestores sejam cada vez mais competentes na arte de tomar decisões. Yu *et al.* (2017, p. 228) lembram que

> decidir é um dos atos humanos que realizamos tantas vezes ao dia, muitas vezes sem nos darmos conta de que estamos decidindo - desde decisões simples e com consequências triviais até casos que, pela complexidade e grande responsabilidade, traduzem momentos marcantes na vida e interferem no sono e na qualidade de vida.

De acordo com Baltzan (2016), embora cada departamento de uma empresa tenha um foco e os seus próprios dados, ninguém pode trabalhar de forma independente em uma organização, quando o intuito é obter a coerência do todo. É fácil ver como uma decisão tomada por um departamento pode afetar os outros. Decisões individuais nem sempre são bem-vindas, pois o saber coletivo pode trazer muito mais vantagens ao corpo empresarial do que uma tomada de decisão centralizada em um indivíduo. Em função disso, as organizações precisam desenvolver flexibilidade e inovação para permitirem a ocorrência de decisões descentralizadas, mas coordenadas.

Mesmo que nas empresas exista a figura de um presidente ou diretor chefe, grupos formados para suporte nas decisões organizacionais contribuem para a compreensão da empresa como um todo e o êxito nos propósitos corporativos, bem como para a obtenção de diferencial ante a concorrência. Afinal, as organizações que cultivam o engajamento coletivo geram desempenho superior nos serviços e criam vantagem competitiva nos negócios (ELDOR, 2020).

Broadbeck *et al.* (2007) sugerem que os grupos de tomada de decisão em um ambiente organizacional distribuído falham em usar efetivamente as informações que são distribuídas entre seus membros, mas, podem superar os tomadores de decisões individuais no caso de conseguirem compartilhar essas informações.

Enquanto a centralização em estruturas organizacionais pode restringir a perspectiva de uma empresa a alguns poucos tomadores de decisão e limitar perspectivas alternativas, a descentralização tende a promover a tomada de decisão participativa e a empurrar a autoridade da tomada de decisão para outros níveis da empresa (Martin *et al.*, 2016).

Mas, para que uma organização distribuída seja capaz de tomar decisões mais rapidamente do que uma com estrutura centralizada, pessoas com diferentes funções e responsabilidades devem trabalhar juntas na tomada de decisões coordenadas, ainda que estejam geograficamente distribuídas (Bearman *et al.*, 2010).

A informação é útil para comunicar a uma organização o desempenho das suas operações atuais, estimar operações futuras e possibilitar a definição de estratégias em relação ao seu funcionamento. Abrem-se novas perspectivas, quando as pessoas têm a informação certa e sabem como usá-la. Diante do cenário empresarial atual, e das inúmeras fontes de informações existentes, tomar decisões se torna uma tarefa difícil, quando não existem ferramentas que possam respaldá-las (Baltzan, 2016).

Para Kumar e Takai (2007), vários tomadores de decisão locais devem interagir uns com os outros em um ambiente de tomada de decisão distribuído para se chegar a uma decisão global. Mas, para que isso aconteça, os funcionários devem conseguir acessar e analisar todas as informações relevantes (Batlzan, 2016). O problema é que as informações podem estar dispersas e disponíveis de forma desorganizada, em diferentes partes da organização, aparecendo em diferentes níveis, formatos e granularidades.

De que forma um tomador de decisão pode encontrar informação válida e relevante para sua empresa, no meio de infindáveis fontes de dados no ambiente de trabalho?

Algumas empresas, mesmo com inúmeros sistemas de informação sofisticados, ainda possuem controles baseados em planilhas eletrônicas ou arquivos não compartilhados, ocasionando problemas de confiabilidade das informações no momento de tomada de decisão. Isto acarreta muita demora para reunir a informação relevante para um determinado estudo da direção da organização. Conforme Gil (2017), é importante integrar a informação processada pelo SIG (sistema de informação gerencial) ao conhecimento necessário ao exercício da tomada de decisão, para que a decisão seja precisa, direcionada e correta para a solução de problemas da organização.

Redundância das informações é um outro problema recorrente nas organizações. A duplicação dos dados, guardados em vários lugares, torna difícil determinar quais valores são os mais atuais ou os mais precisos, além de acarretar problemas de integridade dos dados nas organizações (Baltzan, 2016).

Sendo assim, Inmon e Linstedt (2015) mencionam que uma corporação recebe dados integrados e fundamentais nos quais se pode confiar, por meio da implementação de um sistema de *data warehouse*, possibilitando uma visão unificada de dados relevantes e fidedignos para apoiar as decisões corporativas. Baltzan (2016) enfatiza que a criação de um acervo lógico de informações, reunidas a partir de diversos bancos de dados diferentes, auxilia em atividades de análise de negócios e em tomadas de decisão.

Segundo o *Data Warehousing Institute,* em 2002, empresas norte-americanas sofriam com dados de baixa qualidade, o que lhes custava 600 bilhões de dólares em gastos por ano,

gerando um enorme impacto negativo na capacidade de tomar as decisões certas (Eckerson, 2009).

Para aumentar a qualidade da informação organizacional e, consequentemente, a eficiência das decisões, as empresas devem criar estratégias para manter a informação depurada. A limpeza da informação em um *data warehouse* ocorre no processo de extração, transformação e carregamento dos dados, mas também quando a informação já está no repositório (Baltzan, 2016).

A maioria dos CEOs já experimentou o fenômeno "duelo de planilhas", no qual executivos sêniores utilizam planilhas criadas por analistas de negócios de suas unidades e passam reuniões inteiras discutindo sobre quais dados estão corretos, ao invés de tomar decisões com base em dados disponíveis para todos e não contestados por ninguém (Eckerson, 2009). Isso ocorre porque cada analista, departamento ou divisão vê o mundo através de suas próprias lentes, definindo e modelando de maneira diferente os dados. Como há discordância a respeito de definições simples no ambiente de negócio, a respeito do cliente, venda, produto, receita ou despesa, esta discordância acaba ocorrendo também com relação às análises realizadas.

Nesse ambiente, construir um sistema que objetiva apoiar as decisões, tendo uma interface que possibilite ao usuário analisar a empresa a partir de diversas perspectivas, mas com base nos mesmos dados, pode contribuir com a qualidade das decisões tomadas em reuniões de executivos, além de estimular a coletividade no processo decisório, promovendo a tomada de decisão coletiva na organização.

Os gerentes sabem que dados de baixa qualidade são problemáticos. Dados ruins desperdiçam tempo, aumentam custos, enfraquecem a tomada de decisão, irritam os clientes e dificultam a execução de qualquer tipo de estratégia. Apesar de haver um problema de credibilidade dos dados, em muitos casos, poucos gerentes realizam uma apreciação do impacto de dados ruins sobre o desempenho de suas equipes e departamentos (Nagle *et al.*, 2017).

Luellen (2018) explica que esse problema de credibilidade dos dados assume muitas formas: dados sujos, identificados incorretamente ou simplesmente errados, dados ausentes, ou talvez dados que parecem estar corretos, mas que estão sendo definidos de distintas maneiras. Embora existam estratégias para lidar com dados ausentes, como excluir registros, inserir a média ou mediana para os dados restantes, remover valores discrepantes, isso é demorado e não representa sempre a melhor solução. Para Van Der Lans (2012), em um mundo ideal, os próprios sistemas de produção devem ser feitos de um jeito que seja quase impossível os usuários introduzirem dados incorretos.

É fundamental saber como e quando usar cada estratégia, o que exige que se envolva alguém de TI com experiência em *pipelines* de extração, transformação e carregamento de dados em sistemas, que tenha o conhecimento profundo de dicionário de dados e das definições em uso em cada ambiente organizacional, e que domine os fundamentos e técnicas de modelagem para sistemas OLAP (processamento analítico). Essas habilidades são atributos do engenheiro de dados nas empresas, que transforma os dados brutos, administra armazéns de dados e entrega informações em formas visuais de relatórios, *dashboards*, *scorecards e* para as atividades de *data mining* (Inmon; Linstedt, 2015; Ponniah, 2011).

Nesse cenário, o profissional de engenharia de dados, em um projeto de *software* que visa a apoiar as decisões empresariais, acaba se envolvendo desde a fase de definição de requisitos, de concepção de indicadores de desempenho, até a implementação dos armazéns de dados e painéis de controle a serem disponibilizados na organização (Sharda *et al.,* 2019), uma vez que a engenharia de dados não trata apenas de manter um repositório para grandes volumes de dados, mas de criar possibilidades para todos, de desenvolvedores a cientistas de dados e executivos, obterem acesso aos dados de que precisam para suas decisões (Tamir *et al.*, 2015).

Em 1970, os *Decision Support Systems* (DSS) já eram sistemas computadorizados interativos presentes no apoio de resolução de problemas semiestruturados e desestruturados da classe executiva (Mann; Watson, 1984). Nos anos 1980, os *Executive Information Systems* emergem ampliando o suporte computadorizado para gestores e executivos do primeiro escalão. Eram DSSs visualmente atraentes, projetados como painéis gráficos e concentrados em indicadores básicos de desempenho. Na década de 1990, surgem os DSSs baseados em DW. *Dashboards* e *scorecards* desenvolvidos obtinham dados desses repositórios para o suporte decisório em empresas de grande e médio porte (Sharda *et al.*, 2019).

Nessa época a Gartner cunhou o termo *business intelligence* (BI), a fim de reunir todos os tipos de sistemas de apoio à decisão e recursos de bancos de dados analíticos para a área executiva. Na virada do milênio, os DSSs baseados em DW começaram a ser chamados de sistemas de BI. “Conforme aumentava a quantidade de dados longitudinais acumulados nos DWs, o mesmo ocorria com as capacidades de *hardware* e *software* para acompanhar a rápida evolução dos tomadores de decisões” (Sharda *et al.*, 2019, p. 13).

> Os DSS orientados a dados são projetados para sintetizar grandes quantidades de dados em um formato que seja útil e facilmente interpretável para ajudar os gerentes a avaliarem o desempenho dos processos de negócios e tomar decisões para abordar e resolver quaisquer problemas críticos encontrados. Seu valor depende do tipo e da qualidade dos dados gerados por meio de instrumentos sintéticos denominados *Business Intelligence.* O termo inteligência é usado com o significado de investigar

> para descobrir algo interessante, como no Serviço de Inteligência (Albano*, 2015, p. 4)*.

Nos processos decisórios, decisões tendem a ser mais estruturadas no nível operacional, essencialmente semiestruturadas no nível tático e, tipicamente, não-estruturadas no nível estratégico (Albano, 2015) em que o conselho de diretores e comitê executivo do presidente e principais executivos desenvolvem as metas globais e monitoram o desempenho estratégico da organização e sua direção geral no ambiente político, econômico e competitivo dos negócios (O'Brien, 2001).

Diante disso, entende-se que os principais usuários de um sistema de apoio à decisão são os executivos, que atuam como gestores na empresa, ou seja, todos aqueles que tomam decisões relevantes ao andamento dos negócios da corporação, considerados os principais *stakeholders* de uma ferramenta de apoio à tomada de decisão (Sharda *et al.*, 2019).

Neste estudo, serão captados e tratados dados de egressos da UTFPR, do curso de bacharelado em sistemas de informação. Pretende-se reunir dados oriundos da rede social *LinkedIn*, do sistema de informação acadêmica do *campus* Curitiba da universidade e dados alimentados pelos próprios egressos por meio da plataforma *Google Forms*, a fim de desenvolver uma ferramenta de apoio à decisão, que ajude a compreender o papel da engenharia de dados no processo decisório, proporcionando informações que contribuam com o planejamento estratégico do curso de bacharelado em Sistemas de Informação e a definição de seus rumos futuros.

Este trabalho tem como objetivo principal explorar o uso da engenharia de dados para apoiar a tomada de decisão, sua importância e conceitos fundamentais.

Para ajudar na consecução deste objetivo geral, foram definidos os seguintes objetivos específicos:

- Implementar um projeto de engenharia de dados que forneça suporte ao processo de tomada de decisão;
- Implementar um artefato que estimule a tomada de decisões de maneira coletiva (pelo colegiado de curso);
- Compreender o papel da engenharia de dados na estruturação do artefato pretendido;
- Avaliar a capacidade da engenharia de dados contribuir para a melhoria do processo decisório, de forma coletiva;

- Desenvolver um processo de engenharia de dados aplicado para análise de dados de egressos;
- Explorar a engenharia de dados, possibilitando sua definição, delineamento, informações e mais aprofundamento sobre o papel no universo de dados.

2 FUNDAMENTAÇÃO

Desde que acordam pela manhã, as pessoas podem interagir com tecnologias da informação, utilizando despertadores programados em seus *smartphones*, verificando suas agendas armazenadas em nuvem ou ouvindo músicas e notícias por meio de plataformas de *streaming* no trajeto ao trabalho.

O uso cada vez mais intenso da tecnologia da informação nesta era digital faz com que a humanidade contribua para a geração de dados em uma escala jamais imaginada, fenômeno a que se tem chamado de "*Big Data*".

Macey (2021, p. 453) considera que a quarta revolução industrial surgiu e a chama de "Era de *Analytics*", em que há uma "corrida maluca" para estabelecer a supremacia em um mundo imensamente competitivo e baseado em dados. Começando com análises preditivas e entrando no domínio do aprendizado de máquina (ML) e inteligência artificial (IA).

Conforme Ramzan *et al.* (2019), a engenharia de *software* enfrenta novos desafios relacionados à *cloud computing* acumular grande quantidade de dados diariamente. Esses desafios incluem captura, consulta, compartilhamento, armazenamento, análise e visualização de grande quantidade de dados. *Facebook, Yahoo, Google* e *Amazon* foram os primeiros *sites* a perceber a necessidade de lidar com dados na faixa de *petabytes* a *exabytes*. Essa enorme exigência de espaço de armazenamento de dados fez com que sistemas de gerenciamento de dados relacionais trabalhassem a alto custo visto que "*big data*" exige escalabilidade horizontal para dados massivos, estruturados, semiestruturados e não-estruturados, fazendo com que desenvolvedores investigassem outras alternativas de armazenamento, os bancos de dados *NoSQL* (Zaki, 2014).

Laudon (2007, p. 9) lembra que:

> informação quer dizer dados apresentados em uma forma significativa e útil para os seres humanos. Dados, ao contrário, são sequências de fatos brutos que representam eventos que ocorrem nas organizações ou no ambiente físico, antes de terem sido organizados e arranjados de uma forma que as pessoas possam analisá-los ou usá-los.

Neste universo tecnológico da atualidade, as fontes de dados são inúmeras, pois até as coisas passaram a gerar dados. Espera-se que as "coisas inteligentes" sejam participantes ativas em negócios e processos sociais, onde possam interagir e se comunicar entre si e com o ambiente, trocando dados e informações "sentidas" sobre o ambiente, enquanto reagem de forma autônoma aos eventos do mundo físico real, influenciando-o, executando processos, criando serviços com ou sem intervenção humana direta (Vermesam *et al.*, 2009).

Os objetos passaram a ser conectados à rede mundial de computadores, tornando-se gradativamente mais sofisticados para executarem as atividades de monitoramento, medição e captura de dados por meio da *IoT (Internet of Things)*, ou seja, a "internet das coisas". Para Amaral (2016, p. 13),

> em um futuro próximo, bilhões de dispositivos estarão conectados à internet: veículos, sistemas de compras, automação residencial e industrial, eletrodomésticos, controle logístico e de tráfego. Estes dispositivos serão capazes de trocar dados sem que seja preciso ligar cabos, criar conexões e digitar senhas ou passar cartões. Este cenário está intimamente relacionado com ciência de dados: dispositivos conectados por toda a parte serão grandes produtores de dados.

No âmbito corporativo, os dados podem vir de diversas fontes, dentre elas, diferentes sistemas de informação e tecnologias. Segundo Medeiros (2013, p. 27), "com o advento da *internet*, os sistemas de informação romperam a barreira das redes locais e internas das empresas para disponibilizarem informações de forma global na *web*".

Para Laudon (2007), um sistema de informação (SI) pode ser definido tecnicamente como um conjunto de componentes inter-relacionados que coletam, processam, armazenam e distribuem informações destinadas a apoiar uma organização.

Os sistemas de informação são facilitadores nas rotinas empresariais. Nas organizações é comum que haja diversos deles, dentre os quais: CRM (*Customer Relationship Management*), ERP (*Enterprise Resource Planning*) e sistema de *e-commerce*. Esses sistemas são classificados como OLTP (*On-Line Transaction Processing*).

Para a engenharia de dados, sistemas OLTP representam fontes fidedignas de dados, capazes de alimentar sistemas estratégicos, que se apropriam de informações oriundas de sistemas gerenciais, viabilizando a tomada de decisões estratégicas, agregando dados internos e externos à organização (Pinheiro, 2017).

> No processo decisório, a informação assume capital relevância, na medida em que, se adequada, diminui a incerteza provocada pelo ambiente. [...] a tecnologia da informação é um componente imprescindível para armazenamento de dados, obtenção de informações e geração do conhecimento, itens necessários para o administrador agir de forma segura e consistente (Primak, 2008, p. 15).

Com o crescimento contínuo do volume de informação e da complexidade da era digital, pode haver dificuldade na percepção daquilo que realmente importa para que os negócios sejam, de fato, mais inteligentes. Nesse sentido, Cai e Zhu (2015) alertam para o fato que a era de *big data* faz com que os dados em diversas indústrias e campos apresentem um

crescimento explosivo. Porém, a má qualidade dos dados pode levar a uma baixa eficiência da sua utilização, conduzindo a sérios erros de tomada de decisão.

2.1 Engenharia de dados

A origem da engenharia de dados é antiga. Ela já se fazia presente desde os primórdios dos sistemas de apoio à decisão e inteligência de negócios, sendo evidenciada pela implantação de *data warehouses* (DWs) e *data marts* (DMs) como bancos de dados especialistas para a tomada de decisão nas organizações, sobre os quais o profissional de engenharia de dados atua (Atwal, 2020; White; Olavsrud, 2022).

A engenharia de dados, como um papel distinto no ambiente organizacional. é relativamente nova, mas as responsabilidades a ela relacionadas existem há décadas. De modo geral, um engenheiro de dados disponibiliza dados para uso em *analytics*, *machine learning*, *business intelligence* etc. (Macey, 2021).

Assim, a engenharia de dados pode ser definida como uma disciplina da computação, inspirada pela engenharia de *software*, destinada a transformar dados brutos em modelos específicos para análise e descoberta de conhecimento, além de oferecer subsídio à ciência de dados, conforme Beaucheminm (2017, p. 1):

> O campo da engenharia de dados pode ser pensado como um superconjunto de *business intelligence* e *data warehousing* que traz mais elementos da engenharia de *software*. Essa disciplina também integra a especialização em torno da operação dos chamados sistemas distribuídos de "*big data*", juntamente com conceitos em torno do ecossistema *Hadoop* estendido, processamento de fluxo e em computação em escala.

Pode-se fazer uma analogia entre a engenharia de dados e a engenharia civil, imaginando a engenharia de dados como uma obra em que "o dado é um tijolo, a informação é uma parede construída por vários tijolos e o conhecimento é um cômodo construído a partir da organização do correto relacionamento das várias paredes" (Côrtes, 2008, p. 41).

A atuação da engenharia de dados pode contribuir para encontrar riqueza em grande e variada massa de dados, fornecendo informações que poderão se tornar pilares marcantes para se obter a inteligência empresarial por meio de técnicas de *dashboarding,* permitindo às empresas monitorarem e mensurarem seus negócios*,* melhorando suas relações com parceiros de negócios, com base em dados organizados em uma visão única da verdade (Eckerson, 2010).

Com base nisso, organizar os dados com qualidade para se ter informações relevantes sobre os negócios pode fazer com que as empresas tenham vantagens competitivas no mercado,

devido à possibilidade de promover ações baseadas no comportamento de seus clientes, no seu perfil e geolocalização, fomentando negócios inteligentes. No entanto, conforme Cai e Zhu (2015), mesmo que a análise e mineração de dados possibilitem descobrir a existência de informação ou conhecimento valioso, como descobertas científicas e tratamentos de doenças, isso depende da qualidade dos dados.

Conforme Tamir *et al.* (2015), embora os educadores trabalhem rapidamente na criação de programas acadêmicos para formação em ciência de dados, muito poucos estão focados em engenharia de dados. Apesar disto, as vagas de emprego de engenheiro de dados superam as vagas de cientista de dados, segundo pesquisa realizada por esses autores no *LinkedIn*, onde encontraram, em 2015, mais de 52 mil vagas para engenheiros de dados contra 25 mil vagas para cientistas de dados, aproximadamente.

Tamir *et al.* (2015) também afirmam que a engenharia de dados é um campo próprio e importante, que está recebendo pouca atenção se comparado à ciência de dados, pois aqueles que estão na vanguarda não estão apenas analisando dados, mas estão implementando cada vez mais soluções que mudam a forma como o negócio é executado, combinando dados de várias fontes, suportando muitos usuários simultâneos, com segurança, enquanto protegem a privacidade individual. Essas não são as habilidades do modelador de ciência de dados, mas do engenheiro de dados.

Reis e Housley (2022) também acreditam que a engenharia de dados é separada da ciência de dados e da análise. Elas se complementam, mas são distintamente diferentes. Esses autores definem que a que engenharia de dados é o desenvolvimento, implementação e manutenção de sistemas e processos que recebem dados brutos e produzem informações consistentes e de alta qualidade, que suportam casos de uso *downstream,* como análise e ML. Ela também é a interseção de segurança, gerenciamento de dados, *DataOps*, arquitetura de dados, orquestração e engenharia de software.

Alguns consideram os engenheiros de dados como manipuladores de dados que os limpam e preparam para a análise. Essa é uma parte fundamental do trabalho, mas é apenas uma das suas atribuições. O trabalho desses profissionais inclui uma ampla gama de conhecimentos e habilidades, dentre as quais: extrair, limpar e integrar os dados; *data wrangling;* e fazer ponte entre os modelos de *data science* e sistemas de produção. Isto exige profundo entendimento de bancos de dados relacionais e *NoSQL*, gráficos, computação distribuída, escalabilidade e segurança, além de preocupação com a proteção da privacidade e o anonimato do cliente (Tamir *et al.*, 2015).

Para ter credibilidade nos dados, existem várias estratégias para lidar com o tratamento de dados, sendo fundamental saber como e quando usar cada uma. O ideal é ter alguém com experiência em extração, transformação e carregamento de dados em sistemas e com conhecimento profundo do dicionário de dados e das definições que foram usadas em cada ambiente para realizar alinhamento entre elas, considerando que cada entidade tem seu próprio sistema de dados e definições. Esse alinhamento é um pré-requisito para os projetos de ciência de dados corporativos (Luellen, 2018).

Conforme Provost e Fawcett (2013), a engenharia de dados é essencial para dar suporte às atividades de ciência de dados, apoiando a tomada de decisão orientada por dados (DDDM), que se refere à prática de basear as decisões na análise de dados e não puramente na intuição.

Os *datasets* considerados grandes demais para os sistemas tradicionais de processamento de dados exigiram novas tecnologias, a *Hadoop*, *Hbase* e *CouchDB*, sendo também conhecidas como tecnologias de *big data,* usadas para muitas tarefas, incluindo a engenharia de dados (Provost; Fawcett, 2013). Essas tecnologias de processamento de dados massivos são usadas para o processamento de dados em suporte às técnicas de *data mining* e outras atividades de ciência de dados.

Com base no exposto por Luellen (2018), Provost e Fawcett (2013) e Tamir *et al.* (2015), compreende-se que a engenharia de dados tem campo próprio, constituindo-se na disciplina que organiza e prepara os dados para o suporte à decisão, subsidiando a ciência de dados e promovendo a tomada de decisão baseada em dados nas organizações.

Como já mencionado, sistemas de informação da classe DSS (*Decision Support System*) já existiam em meados da década de 1990, utilizando a arquitetura de *data warehousing* como componente, na qual o papel da engenharia de dados já tinha grande relevância. “Os metadados estão no centro do processamento bem-sucedido de DSS”, conforme lembram Inmon *et al.* (2010, p. 706).

Tanto a engenharia de dados como a ciência de dados são inspiradas e herdeiras da engenharia de *software*, como uma disciplina já madura e estabilizada na computação, pois conforme Saltz e Yilmazel (2016, p. 1),

> o papel de um engenheiro de dados parece semelhante ao de um engenheiro de *software*. Em ambas as funções, a pessoa está focada em escrever *software* que será usado por outros, ou *software* que irá gerar resultados que outros irão alavancar. Além disso, tanto para projetos de dados quanto para projetos de *software* mais tradicionais, as metodologias e o ciclo de vida do projeto são semelhantes em termos da necessidade de projetar o sistema, escrever o código e depois executar os *scripts* de teste.

Porém, isso não significa que são iguais ou que a há garantia de converter um *software engineer* em *data engineer*. Eles se parecem no sentido que escrevem documentos de acordo com uma especificação, projetam a construção, traduzem requisitos de negócio para requisitos de sistema. Sobre essa relação, concluiu-se que há o princípio de "herança" da programação orientada a objetos, em que um engenheiro de dados é um engenheiro de *software*, mas o inverso não necessariamente ocorre (Saltz; Yilmazel, 2016).

Pode-se dizer que a engenharia de dados é, ainda, muito ligada à engenharia de *software*, uma vez que, no desenvolvimento de uma aplicação baseada em dados, utilizam-se fases sequenciais até atingir o objetivo designado, apresentando a visualização dos dados por meio de *interfaces*, como *dashboards*, relatórios e outras formas para apreciação (Sharda *et al.*, 2019; Macey, 2021).

Por sua vez, a ciência de dados tem como alicerce a engenharia de dados, pois em cima dos repositórios de dados construídos pelos engenheiros, os cientistas podem aplicar algoritmos matemáticos e estatísticos para realizar outras atividades, além, do uso de *machine learning* e inteligência artificial (Provost; Fawcett, 2013).

2.2 O papel do engenheiro de dados

O acúmulo constante de informação distribuída em diferentes plataformas e mecanismos de tecnologia provocou a necessidade de se formar recursos humanos capazes de trabalhar com esta quantidade massiva de dados. Nesse aspecto, Macey (2021) afirma que a introdução de tecnologias de *big data*, *data science*, computação distribuída e computação em nuvem contribuíram para tornar o trabalho do engenheiro de dados mais necessário e mais complexo.

Segundo Saltz e Grady (2017), *big data* representa uma mudança significativa nas técnicas e tecnologias usadas para computação intensiva de dados. Embora o termo "ciência de dados" tenha se tornado "onipresente" para descrever qualquer atividade que se relacione a dados, nas equipes de projetos, para criar e implantar sistemas de análise de *big data*, há profissionais que se concentram em análises, geralmente chamados de cientistas de dados, e outros que se concentram na coleta/limpeza de dados, conhecidos como engenheiros de dados.

Um profissional de engenharia de dados se envolve mais no *pipeline* de dados e na ingestão de dados de bancos de dados convencionais ou não. Em contraste, o papel do cientista de dados é mais focado em habilidades como ML e análise de dados. A *Gartner* considera o

papel dos engenheiros como tornar os dados acessíveis e disponíveis para o cientista de dados (Saltz; Grady, 2017).

Esta função é vital para a análise dos dados corporativos, tornando-os mais significantes para as necessidades da empresa, exigindo várias técnicas e habilidades, incluindo o projeto e desenvolvimento de banco de dados SQL (Macey, 2021; Beach, 2022). Sendo assim, entende-se que o engenheiro de dados é aquele que implementa, por meio de tecnologia da informação, a transformação de dados brutos em informações que subsidiarão a tomada de decisão impulsionada por dados e pela ciência de dados.

Segundo Beach (2022), a maioria das estruturas populares de engenharia de dados, como *Spark*, suportam SQL, que continua sendo uma tecnologia importante para a engenharia de dados. Engenheiros de dados não devem ser confundidos com DBAs, pois não estão configurando *clusters* de *failover* automático, mas modelando dados, projetando índices e ajustando consultas, o que exige domínio de SQL e de bancos de dados (Beach, 2022).

Reis e Housley (2022) mencionam que, mesmo com o advento do *MapReduce* e tecnologias de *big data* que relegaram SQL, um engenheiro de dados deve ser altamente competente em SQL, inclusive para reconhecer quando SQL não é a ferramenta certa para um trabalho e codificar de forma alternativa, desenvolvendo em linguagens de programação secundárias, como *Julia*, R, *Go*, *Python* ou C# e *PowerShell*.

Mas, para construir repositórios de dados que permitam melhor gestão dos dados oriundos de diversas fontes existentes, convertidos em informações úteis e relevantes para os negócios, é necessário antes compreender o que a organização espera de impacto nas suas decisões.

Conhecer bem os negócios da empresa e os objetivos esperados com a análise de dados faz parte da rotina dos engenheiros de dados, que também ficam responsáveis pela recuperação de dados e elaboração de *dashboards*, relatórios ou outras visualizações dos dados para as partes interessadas na organização (Sharda *et al.*, 2019).

Segundo Atwal (2020) e White e Olavsrud (2022), existem três funções principais exercidas por engenheiros de dados:

a) Generalistas: são normalmente encontrados em pequenas equipes ou em pequenas empresas e responsáveis pelas diversas etapas do processo de dados, desde o seu gerenciamento até sua análise;
b) Centrados em *pipeline*: geralmente encontrados em empresas de médio porte, trabalham junto com cientistas de dados para ajudar a usar os dados coletados,

dispondo de conhecimento profundo em sistemas distribuídos e ciência da computação;

c) Centrados em bancos de dados: em organizações maiores, onde o gerenciamento do fluxo de dados é um trabalho em tempo integral, concentrando-se em bancos de dados analíticos (OLAP), trabalhando em vários sistemas de gerenciamento de bancos de dados (DBMS) e assumindo a responsabilidade pelo desenvolvimento e manutenção dos esquemas de tabelas.

Segundo Turban *et al.* (2004, p. 402), "o objetivo do *data warehouse* é criar um repositório de dados que dê acesso a dados operacionais sob formas facilmente aceitáveis para as atividades de processamento analítico, como por exemplo apoio à decisão". Um engenheiro de dados em grandes organizações é focado na configuração e preenchimento destes bancos de dados analíticos, envolvendo grandes esforços com ETL (*Extract, Transform and Load*), coleta de dados dos bancos de dados operacionais, externos e de *big data*.

Um engenheiro de dados gerencia o ciclo de vida da engenharia de dados, começando com a obtenção de dados dos sistemas de origem e terminando com a veiculação de dados para casos de uso, como análise ou aprendizado de máquina (Reis; Housley, 2022).

Para White e Olavsrud (2022), no rol de responsabilidades mais comuns de um engenheiro de dados estão:

a) desenvolver, construir, testar e manter arquiteturas;
b) alinhar a arquitetura com os requisitos de negócios;
c) realizar a aquisição de dados;
d) desenvolver processos de conjunto de dados;
e) utilizar linguagens de programação e ferramentas;
f) identificar formas de melhorar a confiabilidade, eficiência e qualidade dos dados;
g) realizar pesquisas para questões industriais e de negócios;
h) usar grandes conjuntos de dados para resolver problemas de negócios;
i) implantar programas analíticos sofisticados, de aprendizado de máquina e métodos estatísticos;
j) preparar dados para modelagem preditiva e prescritiva;
k) encontrar padrões ocultos nos dados;
l) usar dados para descobrir tarefas que podem ser automatizadas;
m) entregar atualizações para os *stakeholders* com base em *Analytics*.

Na atualidade, é esperado que um engenheiro de dados possua habilidades concretas para criar *pipelines* confiáveis, combinar fontes de dados, arquitetar *data warehousing* e *data lakes,* saber atuar com limpeza, preparação, processamento e ingestão de dados, metadados, segurança e qualidade dos dados (Macey, 2021; Beach, 2022), além de colaborar em projetos de *data science*, visto que grande parte do tempo desses projetos é gasto na coleta/seleção dos dados, preparação de dados e análise exploratória (Macey, 2021).

Assim, o cientista de dados pode concentrar seus esforços no aumento da precisão de seus modelos de *machine learning*, utilizando dados oriundos de muitas fontes já manipulados e agregados pelo engenheiro. Ou seja, os engenheiros de dados fornecem as entradas usadas pelos cientistas de dados, que convertem essas entradas em algo útil (Reis; Housley, 2022).

Uma equipe de engenharia de dados é responsável por criar produtos de dados e a arquitetura para criar produtos de dados. Esses produtos de dados são a força vital do resto da organização. O restante da organização consome esses produtos de dados, obtendo *insights* que impulsionam o planejamento ou criam produtos de dados derivados para uso posterior. Cabe aos engenheiros de dados a tarefa de criar *pipelines* de dados que se referem ao processo de pegar dados brutos e transformá-los, para que se tornem utilizáveis por toda a organização, sabendo escolher as tecnologias certas para os dados e os casos de uso (Anderson, 2020).

Em um mundo ideal, os cientistas de dados devem passar mais de 90% do seu tempo focados nas camadas superiores de análise, experimentação, IA e ML, enquanto os engenheiros de dados se concentram nas camadas de coleta, limpeza e processamento, criando uma base sólida para que os cientistas de dados tenham sucesso (Reis; Housley, 2022).

2.3 Computação em nuvem no apoio à engenharia de dados

Não há como se falar de engenharia de dados sem se falar de *cloud computing*, visto que ela se tornou parte das rotinas profissionais deste processo, oferecendo serviços e estruturas especialistas para executar as tarefas. A nuvem é um avanço que oferece uma solução econômica e escalável para armazenar *big data* (Sandhu, 2021). Mas, afinal, do que trata a *cloud computing*, ou computação em nuvem?

Segundo a *Microsoft* (2022), a computação em nuvem é um paradigma computacional que revoluciona os processos empresariais e o cotidiano da humanidade, tendo como suas principais características, o autosserviço por demanda, o amplo acesso à rede, a elasticidade, a medição de serviços e o agrupamento de recursos.

A possibilidade de armazenamento de arquivos e informações, sem ocupar discos rígidos (HDs) localmente, permitindo o acesso de qualquer lugar físico e compartilhando com pessoas de qualquer lugar no planeta, atribui flexibilidade às tarefas diárias das pessoas e rotinas das organizações. Nesse sentido, Susanto *et al.* (2020) dizem que os ecossistemas digitais conectam pessoas em todo o mundo, que podem manter contato e compartilhar suas experiências. As diversas plataformas de sistemas inteligentes em nuvem possuem diferentes funcionalidades disponíveis para uso de seus assinantes. Esses sistemas inteligentes em nuvem também facilitam os negócios, porque abrem o mercado global e barateiam a publicidade.

A computação em nuvem vem fazendo parte do cotidiano das pessoas e empresas (Sunyaev, 2020). Quem atualmente não utiliza contas de correio eletrônico ofertadas pelas empresas *Google* e *Microsoft* ou serviços de armazenamento em nuvem como *Google Drive*, *Microsoft OneDrive* ou *Dropbox*?

É cada vez mais comum dados serem disponibilizados na nuvem. Desde o âmbito corporativo até a vida pessoal, os próprios objetos passam a interagir com bancos de dados em nuvem, a partir da internet das coisas. Conforme Kurose e Ross (2013, p. 1):

> A Internet de hoje é provavelmente o maior sistema de engenharia já criado pela humanidade, com centenas de milhões de computadores conectados, enlaces de comunicação e comutadores; bilhões de usuários que se conectam por meio de *laptops*, *tablets* e *smartphones*; e com uma série de dispositivos como sensores, *webcams*, consoles para jogos, quadro de imagens, e até mesmo máquinas de lavar sendo conectadas.

Nas empresas, a computação em nuvem vem trazendo a redução da necessidade de investimento com infraestrutura interna de *data centers* e espaço em disco rígido nas estações de trabalho, além dos custos com o licenciamento de *software* (Young, 2019).

A razão disso é que os serviços de computação em nuvem são oferecidos por *data centers* acessados pela *Internet*, não sendo necessários muitos recursos computacionais dos microcomputadores e *notebooks* conectados. Também deixa de ser necessário implantar centros de processamento e armazenamento de dados nas organizações para tal finalidade, pois a nuvem terceiriza os encargos e custos de um *data center* corporativo (Hoffman, 2020).

Enfim, a computação em nuvem é o fornecimento de serviços de computação, servidores, armazenamento, bancos de dados, rede, *software*, análise e muito mais, tudo pela *Internet*. Segundo a *Gartner*, as empresas *Amazon*, *Google*, IBM, *Microsoft*, *Oracle* são as principais fornecedoras de nuvem. Uma tendência para o futuro é a *cloud* (Neff; Woodyer, 2020).

Diante de tantos provedores de serviços na nuvem pode haver até “guerras [de preços] de nuvens”. Empresas que não utilizem computação em nuvem se tornarão tão raras quanto as que hoje não utilizam a *Internet* (Kansal *et al.*, 2014).

Segundo a *Microsoft* (2022), empresas que oferecem esses serviços de computação são denominadas provedoras de nuvem e costumam cobrar pelo serviço com base no uso, da mesma forma que empresas de serviços públicos cobram pela água ou energia. Dentre os principais motivos para se utilizar a computação em nuvem estão:

a) Custo: eliminando gastos de capital com compra de *hardware*, *software*, instalação e execução de *data centers* locais (*racks* de servidores, eletricidade com disponibilidade permanente para energia e resfriamento e especialistas de TI para o gerenciamento da infraestrutura);
b) Velocidade: grandes quantidades de recursos computacionais são provisionadas rapidamente, garantindo às empresas grande flexibilidade no planejamento de capacidade computacional, por razão de a maioria dos serviços ser realizada com base no autosserviço e sob demanda;
c) Escala global: capacidade de dimensionamento elástico, que significa prover a quantidade adequada de recursos de tecnologia da informação;
d) Produtividade: equipes de tecnologia da informação conseguem obter metas de negócios em tempo acelerado.
e) Desempenho: os serviços são executados em uma rede mundial de *datacenters* seguros, atualizados regularmente com *hardware* de computação rápida e eficiente;
f) Confiabilidade: redução dos custos de *backup* de dados, recuperação de desastres e continuidade dos negócios, já que os dados podem ser espelhados em diversos *sites* redundantes na rede do provedor de nuvem.

Diante deste cenário, surgiu o *Database as a Service* (DBaaS), ou banco de dados como serviço, que se trata de uma abordagem baseada em computação em nuvem para armazenamento e gerenciamento de dados estruturados, sendo também referido como banco de dados em nuvem (Astrova *et al.*, 2018). Esse serviço oferece funcionalidade semelhante à de um banco de dados, porém baseado na nuvem, fornecendo uma plataforma com flexibilidade, escalabilidade e demais características de *cloud computing*, sendo encontrado nas grandes provedoras de serviços de *cloud computing* (Nirsberger, 2020).

As mudanças tecnológicas exigem flexibilidade e aprendizado contínuo. Exemplos disso são as várias tecnologias existentes para repositórios de dados para processamento

analítico, *data lake*, DW e agora o *data lakehouse* (DLH), que são desafiadoras, mas também representam fonte de oportunidades (Inmon *et al.*, 2021). Nas principais provedoras de serviços de nuvem, é comum encontrar soluções computacionais especialistas em plataformas de dados como serviços nativos de nuvem para *data* e *analytics* a fim de ajudar empresas de qualquer porte a resolver seus problemas computacionais (Sarkar, 2015).

Lakehouses representam um novo padrão de arquitetura de plataforma para dados, que combina os principais benefícios de *data lakes* e *data warehouses* em nuvem (Inmon *et al.*, 2021). *Delta lake* é um exemplo maduro, com recursos ACID e aplicação de esquema que oferece uma confiabilidade que *data lakes* tradicionais não possuem, os metadados são escaláveis e salvos em armazenamento de objetos em nuvem. Um DLH, ao combinar uma camada de metadados otimizada com dados validados e armazenados em nuvem, permite que cientistas de dados criem modelos a partir dos mesmos dados que geram relatórios de BI, visto que muitas organizações usam *data lakes* para ciência de dados e ML, mas não para relatórios de BI devido à sua natureza não validada (Databricks, 2022).

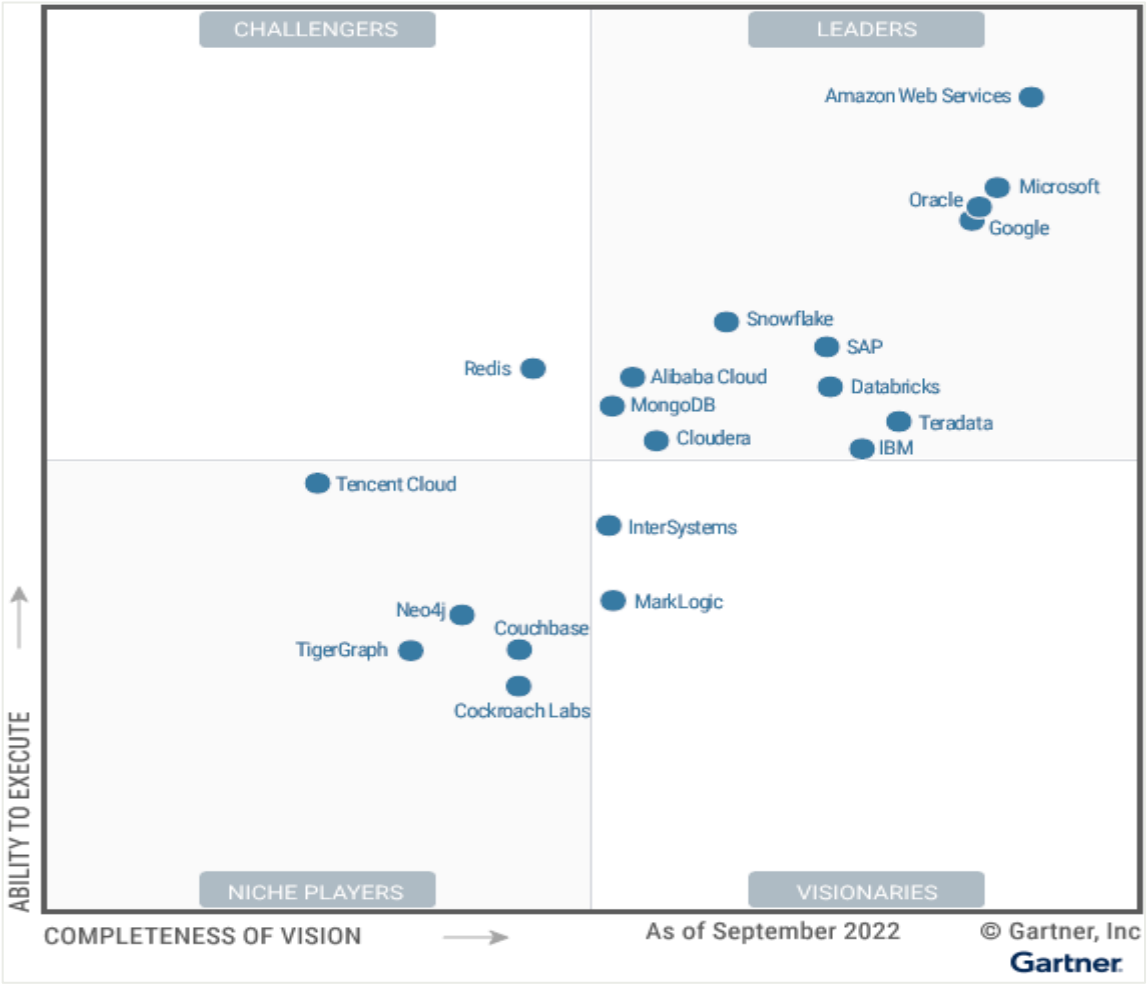
Embora a computação em nuvem traga muitos benefícios, como a escalabilidade e a elasticidade, não se pode esquecer da necessidade de se dispor de uma conexão permanente com a *Internet*. Afinal, qualquer indisponibilidade faz com que as aplicações fiquem inacessíveis ou tenham seu desempenho prejudicado (Microsoft, 2022).

Além disso, algumas empresas poderão ser avessas ao uso da computação em nuvem, principalmente as grandes, devido à questão de segurança e privacidade dos dados corporativos, cabendo aos profissionais de TI envolvidos promover a disseminação das vantagens e dos recursos de segurança em *cloud computing* para as diretorias, informando sobre as formas de implantação de nuvem, a fim de extinguir a insegurança em se usar tecnologia de nuvem (Kavis, 2014).

A computação em nuvem oferece aos profissionais de engenharia de dados, uma "caixa de ferramentas" especialista para atuar neste ecossistema de dados e análises, serviços de nuvem para *data lakes* e *warehouses* corporativos, *big data analytics*, integração de dados, suporte a metadados e recursos de BI.

A Gartner (2022) classifica a *Amazon* como a principal provedora de DBMS em nuvem, seguida da *Microsoft*, *Oracle*, *Google*, *Snowflake*, *Cloudera*, etc. Nessa gama de serviços *cloud*, para os casos de uso analítico, destacam-se os produtos: *Amazon Athena*, *Redshift e Elastic MapReduce*; *Databricks Lakehouse Platform*; *Snowflake Data Cloud*; e *Cloudera Data Platform* (Cook *et al.*, 2022).

Figura 1 – *Magic quadrant* - sistemas de gerenciamento de banco de dados em nuvem



Fonte: Gartner (2022)

Em 2021, a AWS lançou um serviço de *business intelligence* em escala de nuvem sem servidor, com recursos de *data storytelling*, análises *ad hoc*, *insights* de dados automatizados, consulta de linguagem natural e tecnologia de *machine learning*.

Conforme a Gartner (2023), a *Amazon* vem subindo ao topo da classificação, se aproximando de líderes desse nicho de mercado, se juntando com empresas já maduras nesse ramo de ferramentas analíticas de negócios, como *MicroStrategy*, *Salesforce (Tableau)*, *Qlik*, SAP, dentre outras (Schlegel *et al.*, 2023), conforme mostra a Figura 2.

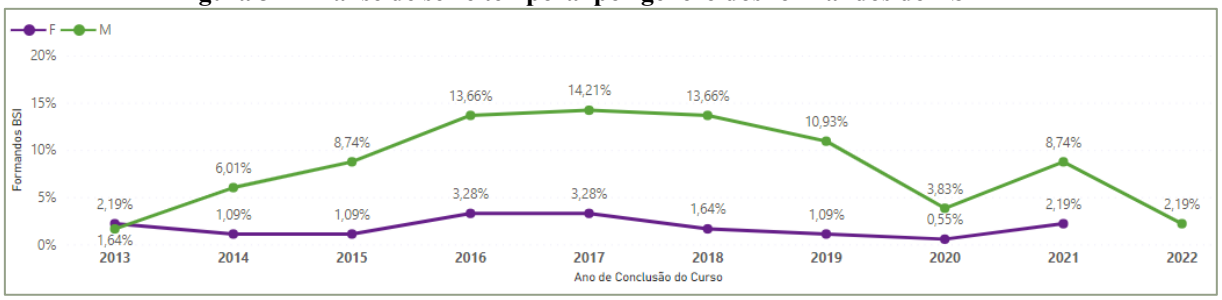
Figura 2 - *Magic quadrant* para plataformas de *analytics* e *business intelligence*

Fonte: Gartner (2023)

3 METODOLOGIA

Nossa primeira fonte de dados para o trabalho foi o Sistema de Gestão Acadêmica da universidade, a partir do qual foram obtidos dados de perfil dos formados no curso de bacharelado em sistemas de informação do *campus* Curitiba. Esse *dataset* continha os seguintes dados: código do aluno, *e-mail*, data de conclusão, data de colação de grau e nome. Com base no nome, foi criada a coluna para gênero, que permitiu a categorização dos formandos conforme mostra a Figura 3 na análise exploratória dos dados. Já com base nos tipos de cota presentes no mesmo *dataset*, foram feitos agrupamentos dos tipos de cotas, verificando os padrões de definições do MEC (Ministério da Educação) e da UTFPR para serem utilizados na análise dos egressos visando melhor explorar este tema de cotas no estudo, como será apresentado na Figura 19 em seção posterior.

Figura 3 – Análise de série temporal por gênero dos formandos do BSI



Fonte: elaborada pelo autor

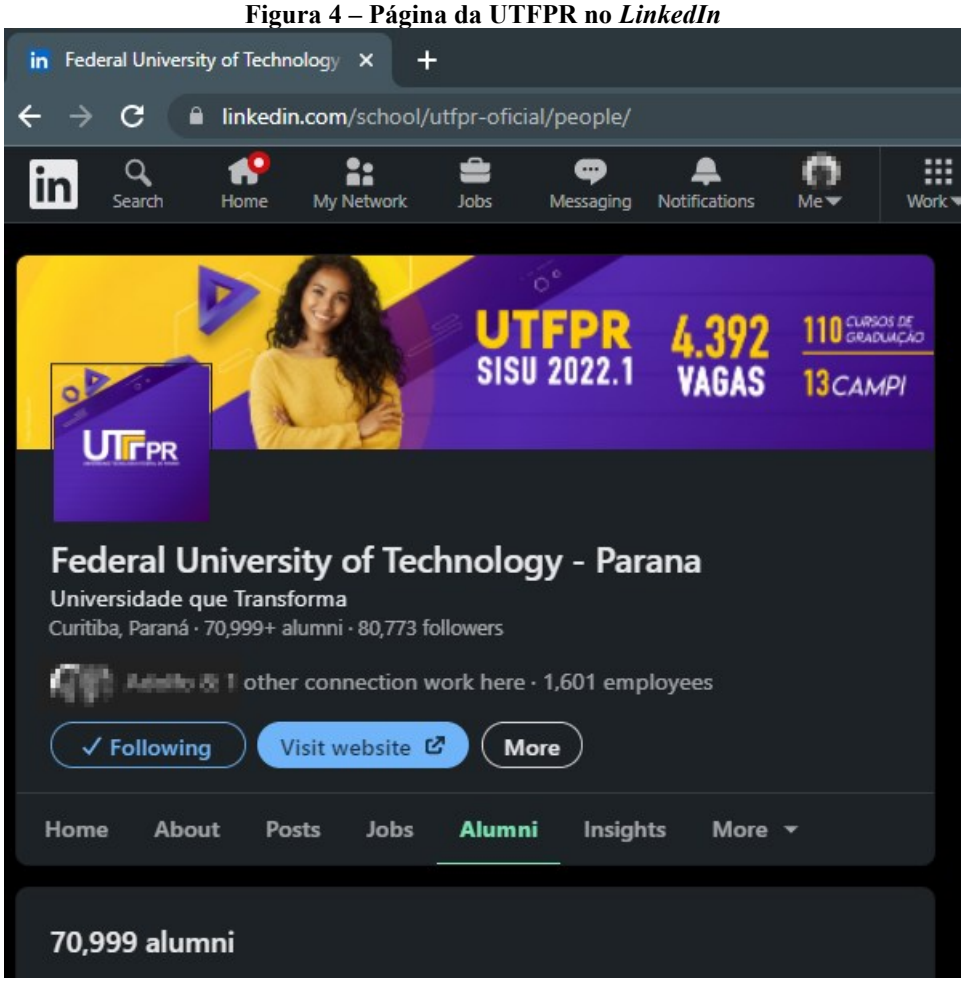
Nosso segundo *dataset* foi formado a partir de dados resultantes do uso da técnica *web scraping* da rede social *LinkedIn*. Foram coletados dados de perfis de profissionais que estavam vinculados à página da instituição acadêmica UTFPR, com intuito de obter informações sobre o que estão fazendo atualmente os egressos do curso de Bacharelado de Sistemas de informação, do *campus* Curitiba.

No processo, houve algumas limitações impostas pela própria plataforma *LinkedIn*, que permite apenas um máximo de mil resultados por pesquisa ou cem páginas de resultados de pesquisa. Diante disso, foi adotada, como estratégia para realização do trabalho de coleta de dados, a contratação de um plano superior ao básico na plataforma, optando pelo *premium business*, que foi escolhido por permitir a visão ilimitada de perfis (Linkedin, 2022).

Dispondo desta conta *premium*, acessou-se o endereço (https://www.linkedin.com/school/utfpr-oficial/people/) para coletar dados dos perfis de ex-alunos da UTFPR, ou seja, todos aqueles que cadastraram sua formação no perfil na rede social, mencionando que estudaram na UTFPR.

Conforme a Figura 4, nas estatísticas da página oficial da instituição na rede social *LinkedIn* existem mais de setenta mil ex-alunos catalogados na plataforma *LinkedIn*. Extrair dados deste elevado número de ex-alunos se tornou uma tarefa pouco produtiva, visto que o foco era apenas nos egressos do curso de Bacharelado em Sistemas de Informação.

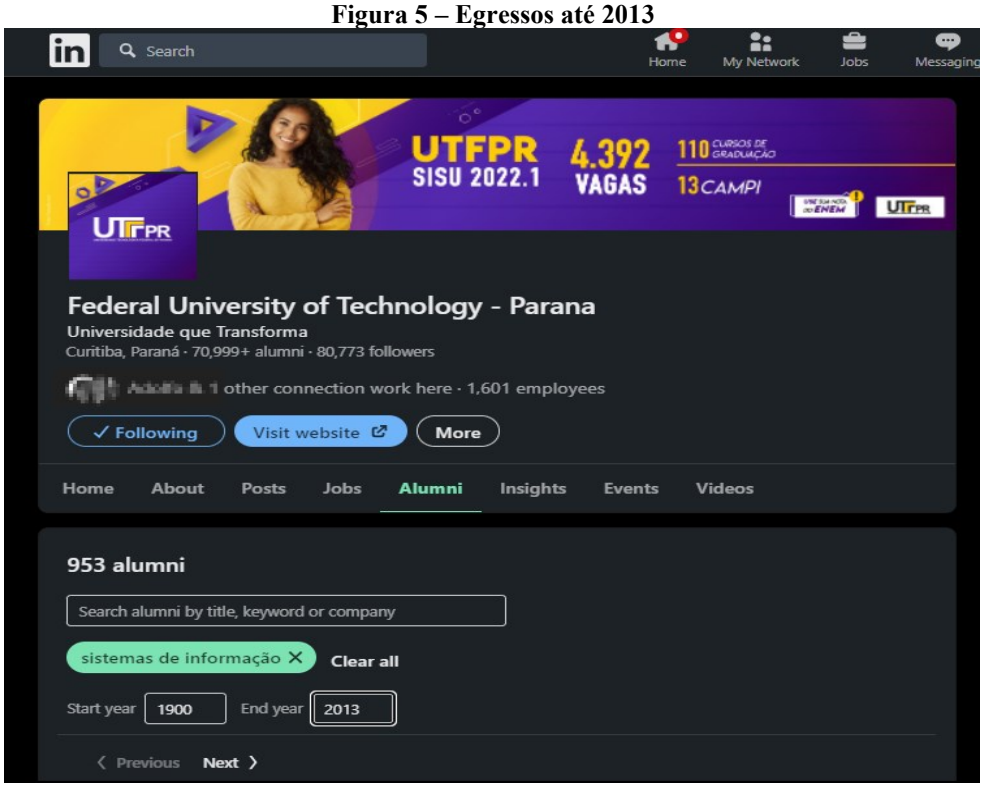
Figura 4 – Página da UTFPR no *LinkedIn*



Fonte: tela capturada do LinkedIn em maio/22

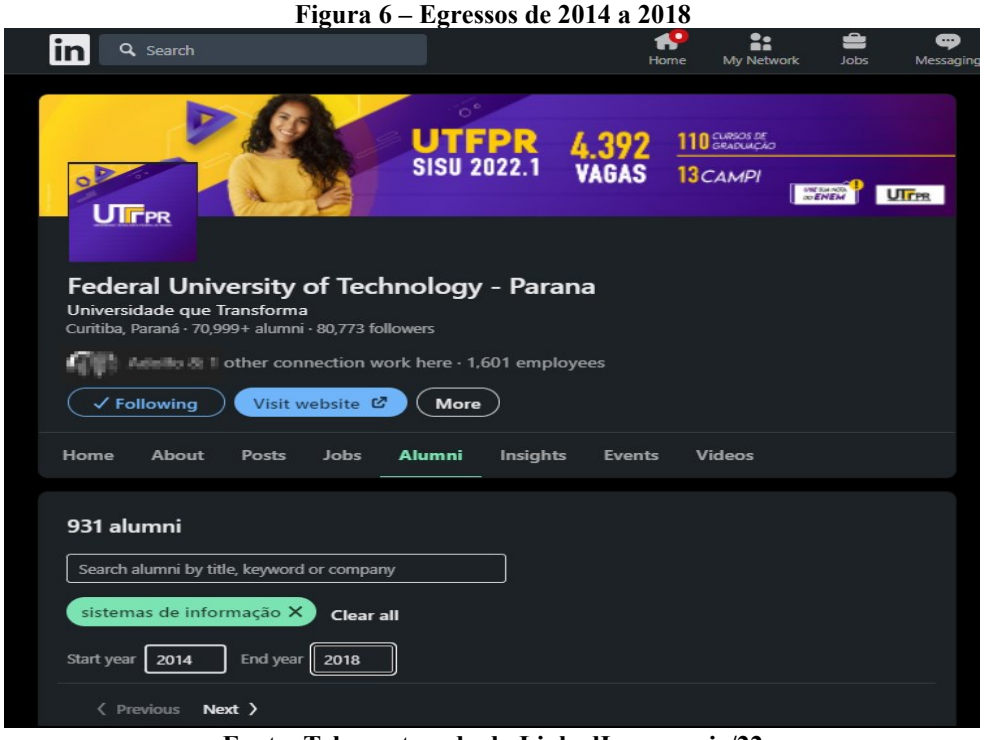
Para isso, inicialmente, utilizando-se os recursos de filtros disponíveis no plano *premium*, foram criados três grupos de resultados de pesquisa (perfis vinculados à página da universidade com a palavra-chave “sistemas de informação”), para aqueles que se formaram até 2013 (953 *alumni*), de 2014 a 2018 (931 *alumni*) e de 2019 até hoje (969 *alumni*), conforme mostrado nas Figuras 5, 6 e 7, a seguir.

Figura 5 – Egressos até 2013



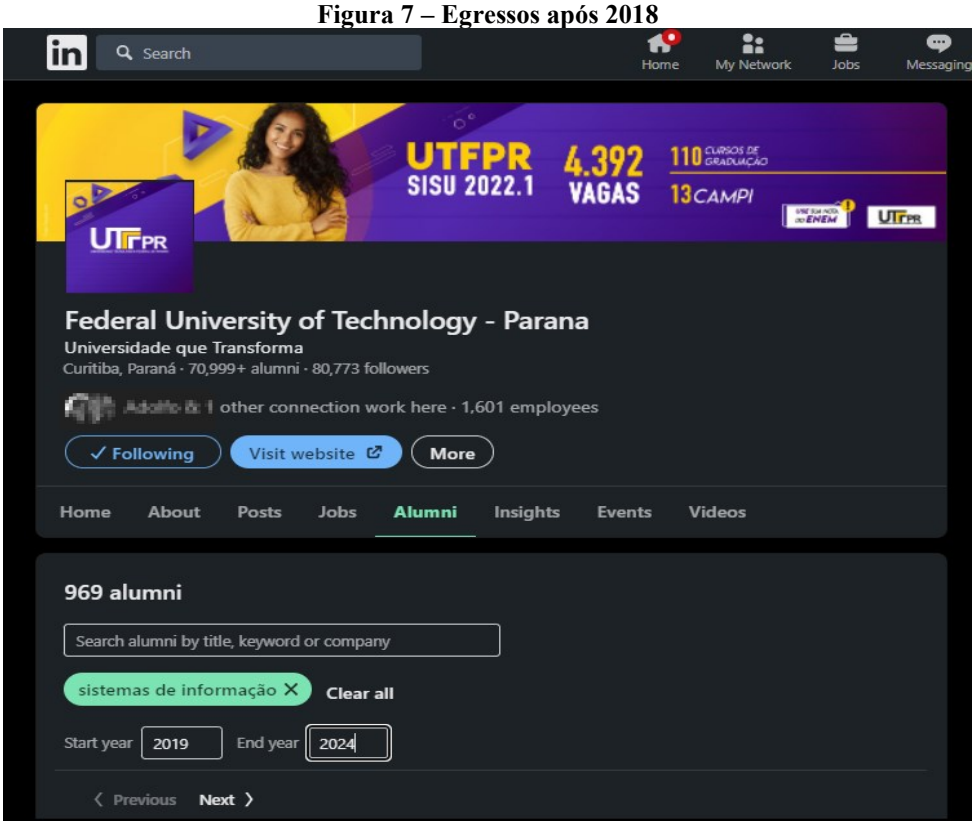
Fonte: Tela capturada do LinkedIn em maio/22

Figura 6 – Egressos de 2014 a 2018



Fonte: Tela capturada do LinkedIn em maio/22

Figura 7 – Egressos após 2018



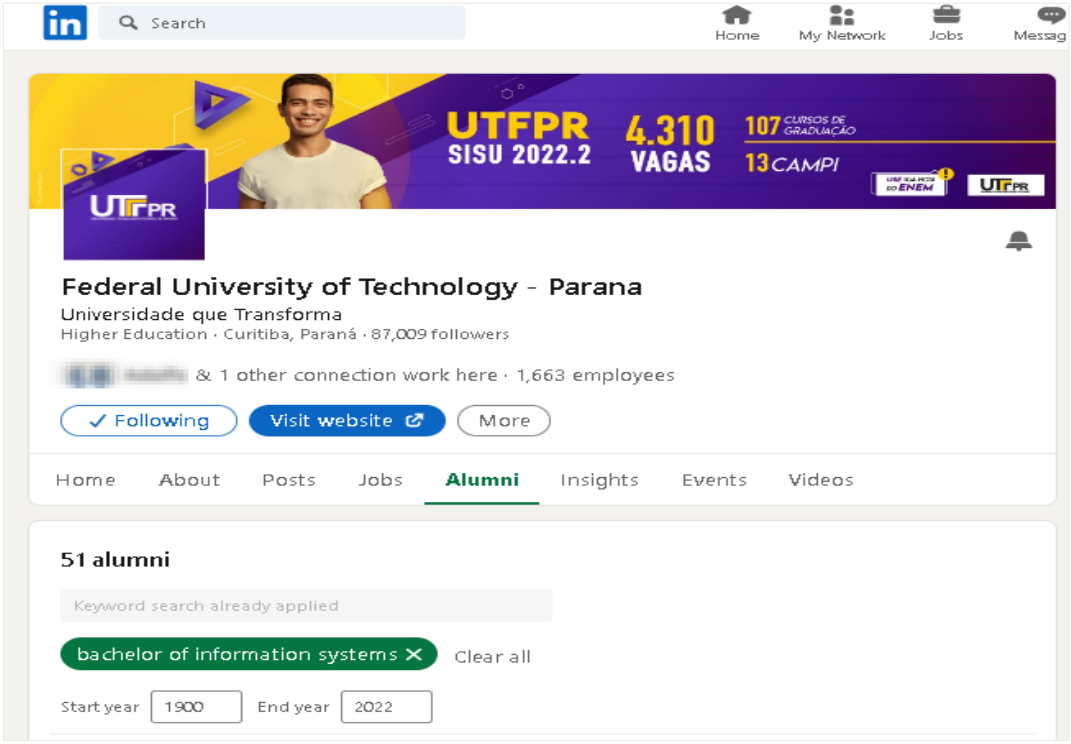
Fonte: Tela capturada do LinkedIn em maio/22

Entretanto, a busca se demonstrou pouco eficaz para os fins do estudo, pois retornou um número acima de 2000 mil perfis, considerando-se que, ao analisar os dados da primeira fonte, dados oriundos do sistema acadêmico fornecidos pela instituição acadêmica em junho/22, foi percebido que havia apenas 183 alunos formados no curso de BSI, os quais concluíram seus estudos entre 2013 e 2022.

A fim de melhorar a qualidade do resultado de filtragem e encontrar os perfis dos egressos do BSI na plataforma *LinkedIn*, foram incluídos os seguintes termos como palavras-chave: "bacharel em sistemas de informação" e "*bachelor in information systems*", conforme as Figuras 8 e 9.

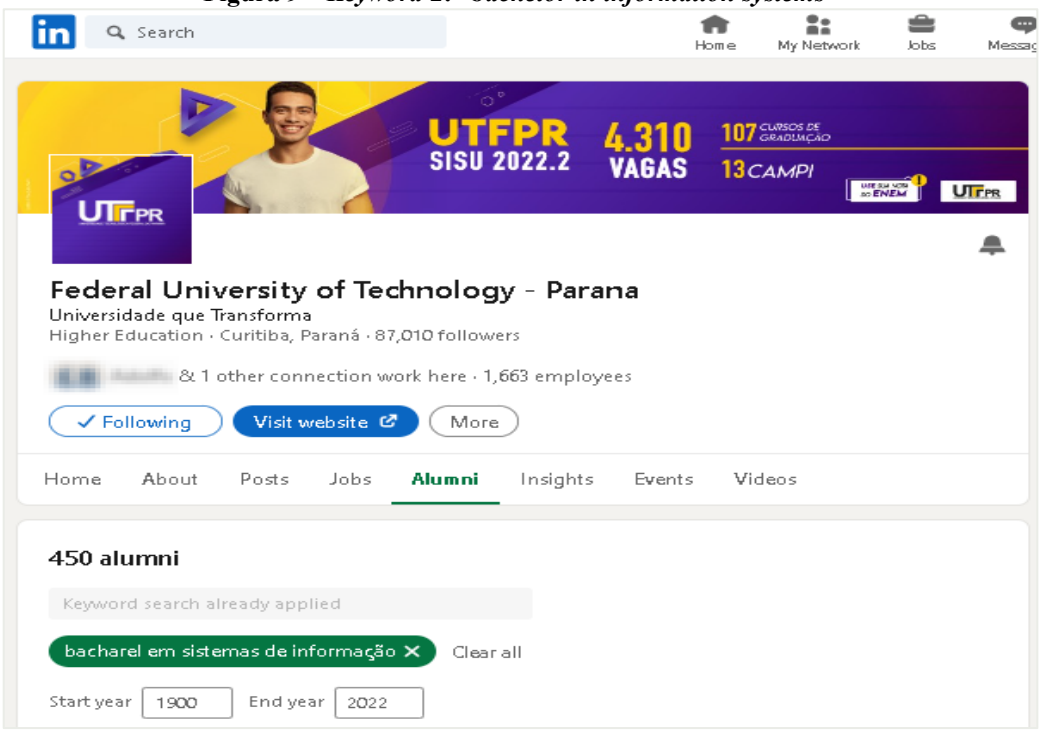
Isso, resultou em um número menor de profissionais comparando com a estratégia anterior. Entretanto, não significou garantia de realmente encontrar os perfis dos formados no curso de BSI, conforme a primeira fonte de dados, constatando-se que uma parte dos egressos não havia informado as palavras chaves no perfil na rede social *LinkedIn*, usou outros padrões de descrição de sua formação ou não atualizou seu cadastro na plataforma.

Figura 8 – *Keyword* 1: "bacharel em sistemas de informação"



Fonte: Tela capturada do LinkedIn em out/22

Figura 9 – *Keyword* 2: "*bachelor in information systems*"



Fonte: Tela capturada do LinkedIn em out/22

Nossa terceira fonte de dados foi resultante de um formulário elaborado na plataforma *Google Forms*, que foi disponibilizado no período de agosto a outubro/2022 e respondido pelos ex-alunos do BSI. O formulário incluiu as questões apresentadas no Quadro 1, abaixo:

Quadro 1 – Formulário ao grupo de egressos

ID	Pergunta	Tipo de Resposta
1	E-mail	Endereço de e-mail em formato válido
2	Qual seu nome completo?	Questão aberta
3	Em que ano se formou no BSI?	Questão aberta
4	Quanto tempo demorou para conseguir o primeiro emprego em uma função profissional relacionada ao BSI?	• Já atuava profissionalmente na área antes de se formar • Até 1 ano depois de formado • De 1 a 2 anos depois de formado • Mais de 2 anos depois de formado • Nunca atuou em profissão relacionada ao curso
5	Você realizou estágio(s) não obrigatório(s) durante o curso? Em caso afirmativo, conte-nos um pouco sobre o aproveitamento que considera ter tido desse(s) estágio(s).	Questão aberta
6	A partir de que período você considera que um aluno do BSI deveria realizar estágio(s) em empresas?	Questão aberta
7	Qual a faixa salarial do seu primeiro emprego?	• Sem Rendimento • Menos de 1 salário mínimo • De 1 a 2 salários mínimos • De 2 a 5 salários mínimos • De 5 a 10 salários mínimos • De 10 a 20 salários mínimos • Acima de 20 salários mínimos
8	Atualmente, qual sua faixa salarial?	• Sem rendimento • Menos de 1 salário mínimo • De 1 a 2 salários mínimos • De 2 a 5 salários mínimos • De 5 a 10 salários mínimos • De 10 a 20 salários mínimos • Acima de 20 salários mínimos
9	Demorou quanto tempo para conseguir o primeiro emprego em uma função profissional da formação acadêmica?	• Ainda não ingressou • Até 1 ano • De 1 a 2 anos • Acima de 2 anos • Já atuava antes de se formar
10	Indique o seu grau de satisfação com relação: • à qualidade geral do curso. • aos professores do curso. • à infraestrutura do curso. • ao seu atual emprego. • às atividades desempenhadas no emprego • à sua profissão • como bacharel em SI.	• Muito insatisfeito • Insatisfeito • Neutro • Satisfeito • Muito satisfeito
11	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional: • formar-se em um curso de nível superior. • ter se formado no BSI. • falar inglês.	• Não importante • Pouco importante • Neutro • Importante • Muito importante

	• ter realizado estágio(s) não obrigatório(s). • ter realizado iniciação científica. • ter realizado cursos e obtido certificações adicionais. • ter participado de um projeto de extensão.	
12	Há alguma competência específica que você gostaria que tivesse sido desenvolvida durante a sua graduação, ou alguma disciplina que considere que seria útil se acrescentada à graduação para formar um profissional mais completo para o mercado? Explique.	Questão aberta
13	Você recomendaria o curso BSI para outras pessoas? Por favor, explique os principais motivos.	Questão aberta
14	Como soube desta pesquisa?	• Por e-mail do professor • Pelo contato realizado pelo *LinkedIn* • Por indicação de um colega de turma • Por convite em grupo do Facebook • Outro. Qual?
15	Qual sua URL de perfil no *LinkedIn*? Complete por favor abaixo: https://www.linkedin.com/in/	Exemplo de URL completa: https://www.linkedin.com/in/SeuNome/

Fonte: elaborado pelo autor

Um pouco menos de um terço dos egressos (55 de 183) atenderam ao convite de participar respondendo a *survey* no *Google Forms*. Destes, cerca de 80% informaram suas respectivas URLs de perfil público na rede social *LinkedIn*, que foram então utilizadas em um cruzamento destes dados com os dados que haviam sido anteriormente obtidos pelo *scraping* do *LinkedIn*.

Com base na lista de URLs dos respondentes, foi reprocessada a extração de dados dessas páginas *web* para garantir a obtenção de dados de todas as URLs informadas na pesquisa.

A partir dos dados obtidos das três fontes, conforme relatado acima, foi desenvolvida a ferramenta, com *interface* amigável e baseada em nuvem, oferecendo os principais indicadores de desempenho resultantes dos dados de sistemas OLAP, permitindo que os usuários realizem análises de dados de forma interativa e intuitiva.

Por fim, depois de os usuários terem acesso à ferramenta, foram coletados depoimentos sobre a contribuição no processo decisório de abordagem de tomada de decisão orientada a dados de forma coletiva.

“Um aspecto importante da ciência artificial é a ideia da contextualização do artefato, seu ambiente de uso, caracterizando quem são seus usuários finais ou praticantes que o usarão" (Sordi, 2021, p. 431).

Segundo Sordi (2021), os artefatos precisam ser válidos e úteis para um contexto específico, caracterizado pelo espaço problemático e pelo grupo de profissionais (praticantes)

delineados pelo pesquisador. Não basta que sejam ou pareçam ser úteis a partir do julgamento do próprio pesquisador. Precisam atender aos profissionais que os utilizarão.

Seguindo assim, após a disponibilização do artefato ao grupo de decisores, composto por membros do Núcleo Docente Estruturante e Colegiado do curso de BSI da UTFPR em Curitiba, buscou-se capturar as seguintes avaliações: se o artefato satisfazia aos requisitos de negócio, se colaborava para a tomada de decisão, se fomentava decisões tomadas em grupo e novos *insights* e se os dados eram válidos.

Isso foi feito por meio das perguntas elencadas no Quadro 2 abaixo, que foram dispostas em um *link* do *Microsoft Forms* a fim de coletar as impressões dos membros do núcleo docente estruturante e do colegiado de curso em relação ao potencial de uso da ferramenta para a tomada de decisões sobre aprimoramentos do curso de bacharelado.

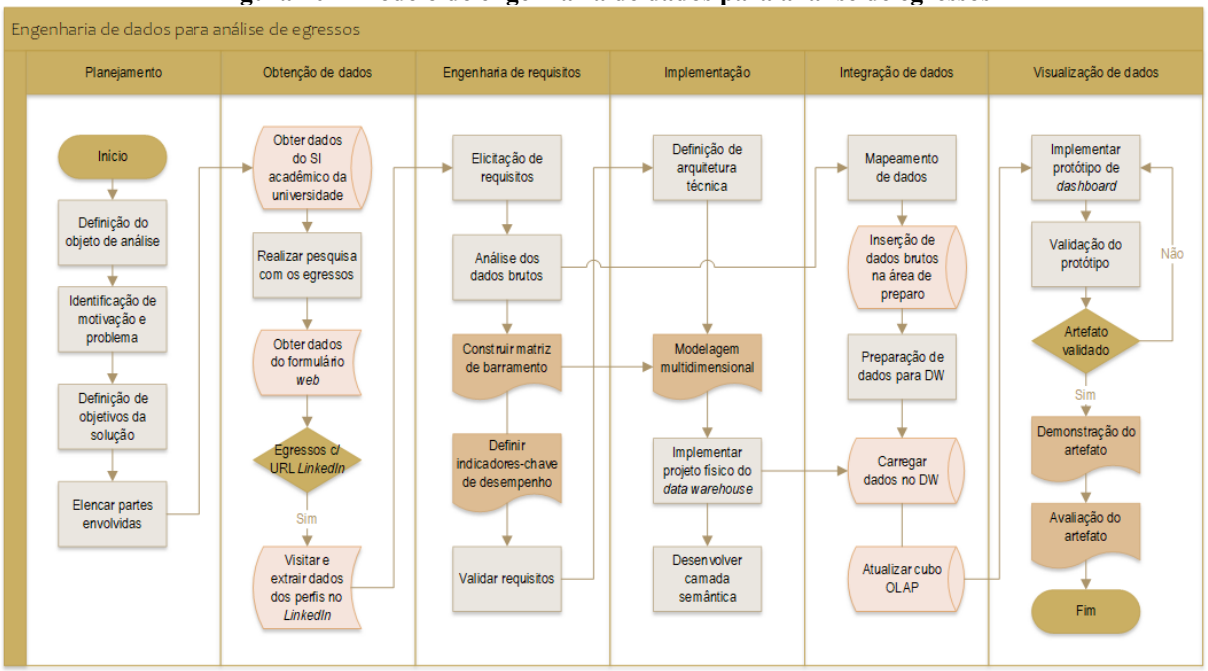
Quadro 2 – Formulário ao grupo de decisores

ID	Perguntas
1	O *dashboard* de indicadores BSI ajuda a compreender a situação dos egressos do curso BSI?
2	Quais indicadores chamaram mais atenção no *dashboard*? por quê?
3	Quais dos dados de empregabilidade te chamaram mais atenção e te fizeram rever alguma ideia anterior sobre o curso?
4	Os indicadores sobre empregabilidade ajudam a compreender a situação dos egressos de que forma?
5	O *dashboard* ajuda na tomada de decisão do colegiado/NDE do BSI?
6	O *dashboard* facilita as tomadas de decisões de forma coletiva sobre o curso BSI?
7	O *dashboard* estar disponível na nuvem facilita a tomada de decisão coletiva?
8	Os dados da forma que foram estruturados permitem analisar/explorar os temas em várias perspectivas?
9	Ao usar o *dashboard*, quais palavras vêm à mente?
10	Qual sua relação com o curso?
11	Quantos anos de experiência como professor?
12	Qual seu gênero?
13	Em que parte da Cidade você mora (Regional)?
14	Você gostaria de fazer mais algum comentário?

Fonte: elaborado pelo autor

Neste trabalho, foram realizadas as etapas apresentadas na Figura 10, que representa a implementação de um processo de engenharia de dados destinado a apoiar à análise de egressos, baseado nos dados empíricos do curso de bacharelado em sistemas de informação da UTFPR, que foi o objeto de análise empírica do estudo, sem deixar de levar em conta a replicabilidade por outros implementadores ou buscadores de conhecimento em projetos de sistemas de informação que objetivam apoiar a tomada de decisões em ambientes de educação superior ou de outra ordem.

Figura 10 – Modelo de engenharia de dados para análise de egressos



Fonte: elaborada pelo autor (2023)

O presente trabalho, objetiva desenvolver conhecimentos para que profissionais possam usar em projetos de soluções para seus problemas de campo, alcançar o conhecimento e a compreensão de um domínio de problema por meio da construção e aplicação de um artefato projetado (Hevner *et al*., 2004). Este objeto será detalhado nas seções conseguintes.

A pesquisa pode ser classificada como exploratória, quanto aos fins da pesquisa, pois teve como finalidade proporcionar mais informações e conhecimentos sobre a temática, possibilitando sua definição e delineamento, além de permitir mais aprofundamento (Gil, 2019).

Ela é quanti-qualitativa, quanto à forma de abordagem do problema, pois articula as dimensões quantitativa e qualitativa, empregando métodos estatísticos e matemáticos na coleta dos dados e no tratamento deles, que envolve a análise, descrição e o desenvolvimento de melhor compreensão do fenômeno.

Como instrumentos de obtenção de dados para a pesquisa foram utilizados: a observação participativa, visando compreender o comportamento do objeto de estudo; entrevistas estruturadas diretivas, com questões delineadas pelo entrevistador aos participantes por meio de ferramentas de formulário *web*; e pesquisa bibliográfica.

As fontes dos dados da pesquisa foram: o próprio SI acadêmico da universidade, fornecendo dados dos 183 egressos que concluíram o curso de BSI entre 2013 e 2022; os dados das respostas da pesquisa realizada com os egressos por meio do *Google Forms*, que contou

com a participação de 55 respondentes; e os dados recuperados do *LinkedIn* inerentes aos 44 perfis dos egressos que participaram da pesquisa e informaram suas respectivas URLs da plataforma. Os principais dados primários obtidos estão listados no Quadro 3 a seguir.

Quadro 3 – Dados brutos de pesquisa

Fonte	Campo	Tipo
SI Acadêmico	Código do aluno	inteiro
SI Acadêmico	Nome do aluno	texto
SI Acadêmico	Data de início do curso	data
SI Acadêmico	Data de conclusão do curso	data
SI Acadêmico	Data de colação de grau	data
SI Acadêmico	*E-mail*	texto
SI Acadêmico	Cotista	texto
SI Acadêmico	Telefone residencial	texto
SI Acadêmico	Telefone comercial	texto
SI Acadêmico	Telefone contato	texto
Google Forms	Carimbo de data/hora	data/hora
Google Forms	Endereço de *e-mail*	texto
Google Forms	Qual o seu nome completo?	texto
Google Forms	Em que ano se formou no BSI?	inteiro
Google Forms	Quanto tempo demorou para conseguir o primeiro emprego em uma função profissional relacionada ao BSI?	texto
Google Forms	Você realizou estágio(s) não obrigatório(s) durante o curso?	texto
Google Forms	A partir de que período você considera que um aluno do BSI deveria realizar estágio(s) em empresas?	inteiro
Google Forms	Qual a faixa salarial do seu primeiro estágio ou emprego?	texto
Google Forms	Atualmente, qual sua faixa salarial?	texto
Google Forms	Indique o seu grau de satisfação com relação... [à qualidade geral do curso.]	texto
Google Forms	Indique o seu grau de satisfação com relação... [aos professores do curso.]	texto
Google Forms	Indique o seu grau de satisfação com relação... [à infraestrutura do curso.]	texto
Google Forms	Indique o seu grau de satisfação com relação... [ao seu atual emprego.]	texto
Google Forms	Indique o seu grau de satisfação com relação... [às atividades desempenhadas no emprego.]	texto
Google Forms	Indique o seu grau de satisfação com relação... [à sua profissão como bacharel em SI.]	texto
Google Forms	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional. [formar-se em um curso de nível superior.]	texto
Google Forms	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional. [ter se formado no BSI.]	texto
Google Forms	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional. [falar inglês.]	texto
Google Forms	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional. [ter realizado estágio(s) não obrigatório(s).]	texto
Google Forms	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional. [ter realizado iniciação científica.]	texto
Google Forms	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional. [ter realizado cursos e obtido certificações adicionais.]	texto
Google Forms	Indique o grau de importância que você atribui a cada um dos itens a seguir, para o avanço na carreira profissional. [ter participado de um projeto de extensão]	texto
Google Forms	Há alguma competência específica que você gostaria que tivesse sido desenvolvida durante a sua graduação, ou alguma disciplina que considere que seria útil se acrescentada à graduação para formar um profissional mais completo para o mercado? Explique.	texto
Google Forms	Você recomendaria o curso BSI para outras pessoas?	texto
Google Forms	Por favor, explique os principais motivos para a resposta anterior.	texto

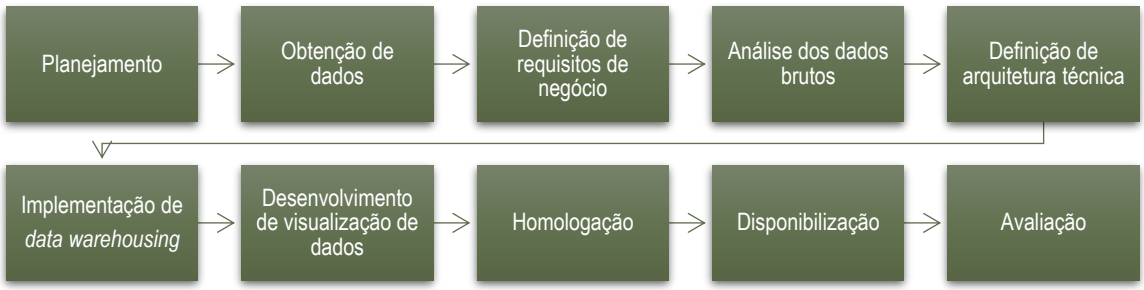
Google Forms	Por favor, nos conte um pouquinho sobre o que anda fazendo da vida, quais são seus objetivos e sonhos para os próximos anos...	texto
Google Forms	Como soube desta pesquisa?	texto
Google Forms	Qual a URL do seu perfil no *LinkedIn*? Complete por favor abaixo.	texto
LinkedIn	ID público	texto
LinkedIn	ID de membro	texto
LinkedIn	URL do perfil	texto
LinkedIn	Nome completo	texto
LinkedIn	Primeiro nome	texto
LinkedIn	Último nome	texto
LinkedIn	Nome da localização	texto
LinkedIn	Endereço	texto
LinkedIn	Aniversário	inteiro
LinkedIn	Título na organização	texto
LinkedIn	Início na organização	inteiro
LinkedIn	Fim na organização	inteiro
LinkedIn	Localização da organização	texto
LinkedIn	Título na organização 2	texto
LinkedIn	Início na organização 2	inteiro
LinkedIn	Fim na organização 2	inteiro
LinkedIn	Localização da organização 2	texto
LinkedIn	Título na organização 3	texto
LinkedIn	Início na organização 3	inteiro
LinkedIn	Fim na organização 3	inteiro
LinkedIn	Localização da organização 3	texto
LinkedIn	Idioma 1	texto
LinkedIn	Proficiência linguística 1	texto
LinkedIn	Idioma 2	texto
LinkedIn	Proficiência linguística 2	texto
LinkedIn	Idioma 3	texto
LinkedIn	Proficiência linguística 3	texto
LinkedIn	Idioma 4	texto
LinkedIn	Proficiência linguística 4	texto
LinkedIn	Idiomas	texto
LinkedIn	Contagem de seguidores	inteiro
LinkedIn	Contagem de conexões	inteiro

Fonte: elaborado pelo autor

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Neste capítulo, é apresentado o desenvolvimento de um sistema de apoio à decisão baseado em dados, a fim de demonstrar o papel da engenharia de dados no fomento à tomada de decisão orientada por dados (DDDM) pelos integrantes do colegiado de curso do bacharelado em Sistemas de Informação da UTFPR, buscando fazer uso de fatos, métricas e dados para orientar as decisões estratégicas alinhadas com as metas, iniciativas e objetivos da organização acadêmica. Em síntese, o ciclo de vida do desenvolvimento do projeto de engenharia de dados aplicado neste trabalho está representado na Figura 11 e explanado ao decorrer das próximas seções.

Figura 11 – Ciclo de vida do projeto de engenharia de dados



Fonte: elaborada pelo autor

No planejamento realiza-se a definição do objeto de análise, estabelecem-se quais são as perguntas a serem respondidas com o projeto e, de forma clara, quais os objetivos mensuráveis e relevantes para o negócio.

A definição de requisitos de negócio ocorre na engenharia de requisitos, após o que é importante identificar, mapear e averiguar se os dados primários atendem ao processo de ED e às necessidades do suporte à tomada de decisões do escopo pretendido.

4.1 Engenharia de requisitos

Na fase de definição de requisitos do negócio, foi elaborado um artefato do projeto ilustrado na Figura 12. A matriz de barramento é uma ferramenta que serve de guia para a arquitetura do modelo de dados, que fornece uma abordagem incremental e integrada para DW, usada como entrada de um projeto de sistema de apoio à decisão. Em cada linha da matriz é verificado se uma ou mais dimensões são associadas a um determinado processo de negócio (Ross *et al.*, 2013). Uma definição de requisitos do negócio eficaz é crucial, pois se estabelece como fundação para todas as atividades posteriores do ciclo de vida do projeto. A probabilidade

de sucesso de uma iniciativa de *data warehousing/business intelligence* aumenta com a compreensão dos usuários de negócio e seus requisitos (Kimball *et al.*, 2015).

Figura 12 – Matriz de barramento

	Data Pesquisa	Data Conclusão	Sexo	Faixa Salarial	Tipo Convite	Possui LinkedIn	Localidade	Está Empregado	Local Sede Org.	Recomenda BSI	Demora p/1º trab.	1ª Faixa Salarial
Egressos respondentes	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Egressos ñ respondentes	⃠	✔	✔	⃠	⃠	⃠	⃠	⃠	⃠	⃠	⃠	⃠
Indique o seu grau de satisfação com relação:												
à qualidade geral do curso.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
aos professores do curso.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
à infraestrutura do curso.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ao seu atual emprego.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
às atividades desempenhadas no emprego.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
à sua profissão como bacharel em SI.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Indique o grau de importância que você atribui a cada um dos itens a seguir p/ avanço na carreira profissional:												
formar-se em um curso de nível superior.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ter se formado no BSI.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
falar inglês.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ter realizado estágio(s) não obrigatório(s).	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ter realizado iniciação científica.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ter realizado cursos/obtido certificações adicionais.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ter participado de um projeto de extensão.	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Fonte: elaborada pelo autor

Antes de se começar a construir a matriz de barramento, foram obtidas informações por meio de entrevistas junto a professores representantes do Núcleo Docente Estruturante (NDE) e de um estudo de artefatos preliminares: relatórios do sistema acadêmico, dados recuperados do *LinkedIn* e pesquisa realizada com os egressos, conforme Quadro 1.

O objetivo desse mapeamento de necessidades informacionais da Figura 12, foi de identificar o que se deseja medir e elicitar quais dimensões poderiam existir, sendo que as dimensões fornecem contexto ou informações de suporte textual à área de assunto.

Na pesquisa realizada com os egressos do BSI da UTFPR, *campus* Curitiba, cada egresso foi convidado a emitir o seu grau de concordância com treze afirmações contidas em um formulário *Google Forms*.

Ao serem analisados os dados brutos, após o encerramento de recebimento de respostas, percebeu-se que os graus de concordância estavam no formato texto, sendo assim, os

respectivos valores foram mapeados para os níveis de respostas, conforme formato típico de itens da escala *Likert*, como mostra o Quadro 4, a seguir.

Quadro 4 – Itens da escala *Likert*

Nível	Item Likert	Grau de Satisfação	Grau de Importância
1	Discordo totalmente	Muito insatisfeito	Não importante
2	Discordo parcialmente	Insatisfeito	Pouco importante
3	Indiferente	Neutro	Neutro
4	Concordo parcialmente	Satisfeito	Importante
5	Concordo totalmente	Muito satisfeito	Muito importante

Fonte: elaborado pelo autor

Com base na Figura 12 e nas respostas dos egressos, antevendo a entrega do produto final, foram definidos alguns índices que poderiam existir para análise dos egressos e o fomento da tomada de decisão coletiva. Dessa forma, preparou-se o Quadro 5, a seguir, como forma de se estabelecer uma avaliação preliminar com relação aos índices de satisfação, importância e empregabilidade, os quais poderiam ser, posteriormente, ajustados pelos decisores.

Quadro 5 – Mapa de KPIs

KPI	Fórmula	UOM	Meta	Ruim	Bom
Índice de respondentes	total de respondentes / total de egressos	Percentual	0%	50	80
Índice de satisfação	soma de todos os graus de satisfação / total de respondentes	Valor	4	2	>3
Índice de importância	soma de todos os graus de importância / total de respondentes	Valor	4	2	3
Índice de empregabilidade	total de egressos empregados / total de egressos	Percentual	0%	30	60
Índice de respondentes empregados	total de respondentes com emprego / total de egressos que responderam à pesquisa	Percentual	0%	30	60

Fonte: elaborado pelo autor

Esse modelo de artefato de definição de indicadores, apresentado no Quadro 5, auxilia tanto na construção da engenharia de dados bem como na camada de apresentação dos dados, quando forem criados os objetos de visualização de dados. Portanto, entende-se como um elemento importante de um projeto de engenharia de dados.

Geralmente as organizações utilizam-se de *Balanced Scorecard* (BSC) como metodologia de gestão estratégica, mensuração, avaliação de desempenho, mapas estratégicos visando implementar o monitoramento de performance das estratégias (Kaplan; Norton, 2018). No entanto, mesmo que os indicadores de desempenho do BSC possam se enquadrar a uma

universidade pública, existem necessidades de adaptação para a implantação do BSC no setor público, estabelecendo outros indicadores de desempenho relevantes para o aprimoramento dos serviços envolvidos, visando à satisfação da sociedade (Oliveira; Izelli, 2018).

Na classificação das decisões, pode-se dizer que no NDE do BSI algumas decisões podem ser consideradas como programadas, pois espera-se uma melhora no mercado de trabalho brasileiro, bem como o aumento de ingressos no curso de BSI. Mas também haverá decisões não programadas, devido ao cenário de pandemia ter apresentado impactos inesperados na conclusão de cursos, impactando no número de egressos.

Um engenheiro de dados bem-sucedido sempre se esforça para entender o panorama geral e como aumentar o valor para o negócio. Muitas vezes equipes de dados são bem-sucedidas com base em sua comunicação com outras partes interessadas. Saber como navegar em uma organização, definir o escopo, reunir requisitos e aprender continuamente diferencia engenheiros de dados que dispõem dessas habilidades dos que dependem exclusivamente de suas habilidades técnicas para levar adiante sua carreira (Reis; Housley, 2022).

4.2 Modelagem multidimensional

Baseando-se na definição de requisitos de negócio e na análise das fontes de dados, é possível avançar para a definição da arquitetura de dados, realizando a modelagem de destino dos dados. O nível físico do modelo multidimensional implementado está representado na Figura 13, onde é aplicada a técnica *star schema* de modelagem de dados para sistemas de DW.

Figura 13 – Modelo dimensional de dados

Fonte: elaborada pelo autor

Ao implementar a modelagem dimensional do DW deste projeto, os fatos e os elementos que participam dos fatos foram modelados, sendo que os elementos são referências para os fatos analisados e estão dentro das várias dimensões. Na dimensão “tempo”, os membros podem ser organizados em níveis hierárquicos e se permite analisar até o dia do fato. Desta forma, é possível estabelecer a granularidade mínima de tempo neste projeto, respondendo quando o egresso se formou e participou da pesquisa de satisfação sobre o curso. Sendo a sumarização uma das operações de um DW, o modelo enfatiza o uso de valores agregados sobre os egressos, não se preocupando em expor quem é o aluno. A realização da anonimização ocorre por meio da aplicação de somas das quantidades analisadas e média nas avaliações coletadas na pesquisa

com os egressos. Na dimensão “localidade”, é gerado o ponto geométrico das coordenadas aplicadas para cada localização coletada, que foi informada nos perfis dos egressos no *LinkedIn*. Foi gerada uma *surrogate key* para cada instância na dimensão e também armazenadas as coordenadas em formato texto e formato numérico para subsidiar qualquer tipo de instrumento de visualização de dados na camada superior do sistema de *data warehouse*, *front-end* com o usuário final.

4.3 Arquitetura em camadas

As decisões relevantes vão ser derivadas de informações que precisam ser pautadas em dados confiáveis com qualidade e que realmente representem a verdade dos fatos, porque ao relacionar esses dados e gerar uma informação, essa informação pode ser contextualizada, tornando-se útil para quem irá tomar as decisões. Sendo assim, a base é o alicerce, é a fundamentação do sistema para apoio à tomada de decisão.

Com uma compreensão das informações de negócios que foram capturadas e serão integradas, nesta seção são abordados brevemente alguns desafios de implementar um sistema OLAP. Apresenta-se uma arquitetura em camadas para entender a funcionalidade necessária para implementar com sucesso a tecnologia de *data warehousing* com base no que foi desenvolvido nesse trabalho. A Figura 14 ilustra a arquitetura em camadas.

Figura 14 – Processo de arquitetura em camadas

Fonte: elaborada pelo autor

Nas subseções a seguir discorre-se brevemente sobre o propósito de cada camada mencionada no processo de arquitetura de sistema OLAP, descrito na Figura 14.

4.3.1 Camada de sistemas fonte

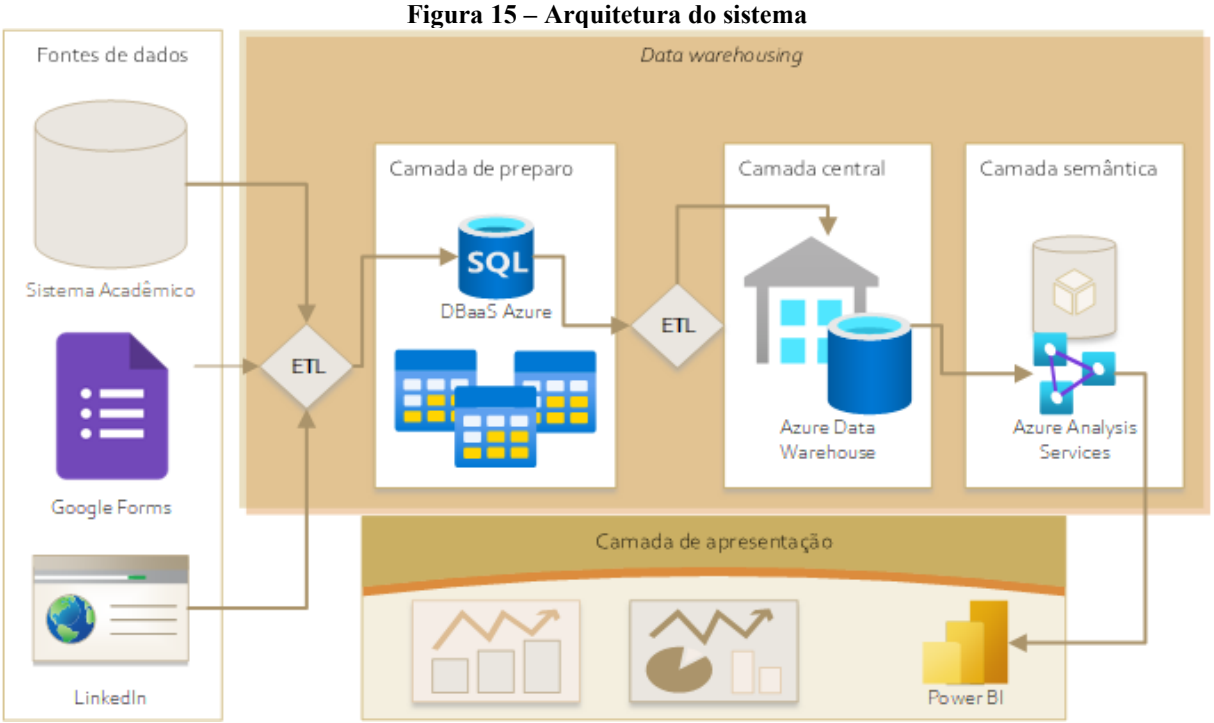
A camada de sistemas fonte apresenta os dados em estado bruto e originário, frutos dos sistemas OLTP, fontes externas, ou de arquivos considerados fidedignos como fontes de informação. Isso implica ao profissional de engenharia de dados acabar tendo que lidar com vários formatos de fonte de dados usados comumente por alguns sistemas e plataformas de processamento de *big data* para armazenar ou trocar dados, incluindo *Apache Parquet/ORC*,

formatos de armazenamento colunar e *Apache Avro*, formato de arquivo aberto de armazenamento de dados orientado por linha (Zeydan *et. al.*, 2022), formatos otimizados para recuperação rápida de dados.

4.3.2 Camada de integração de dados

Esta camada se refere à onde acontecem as integrações de dados, combinando os dados das distintas fontes, tratamento, limpeza, organização, transformações e carregamento, fazendo uso de técnicas e implementações de *pipelines* de ETL, conforme ilustra a Figura 15 baseada na arquitetura da implementação aplicada neste estudo dos dados empíricos do ambiente da organização acadêmica.

Figura 15 – Arquitetura do sistema



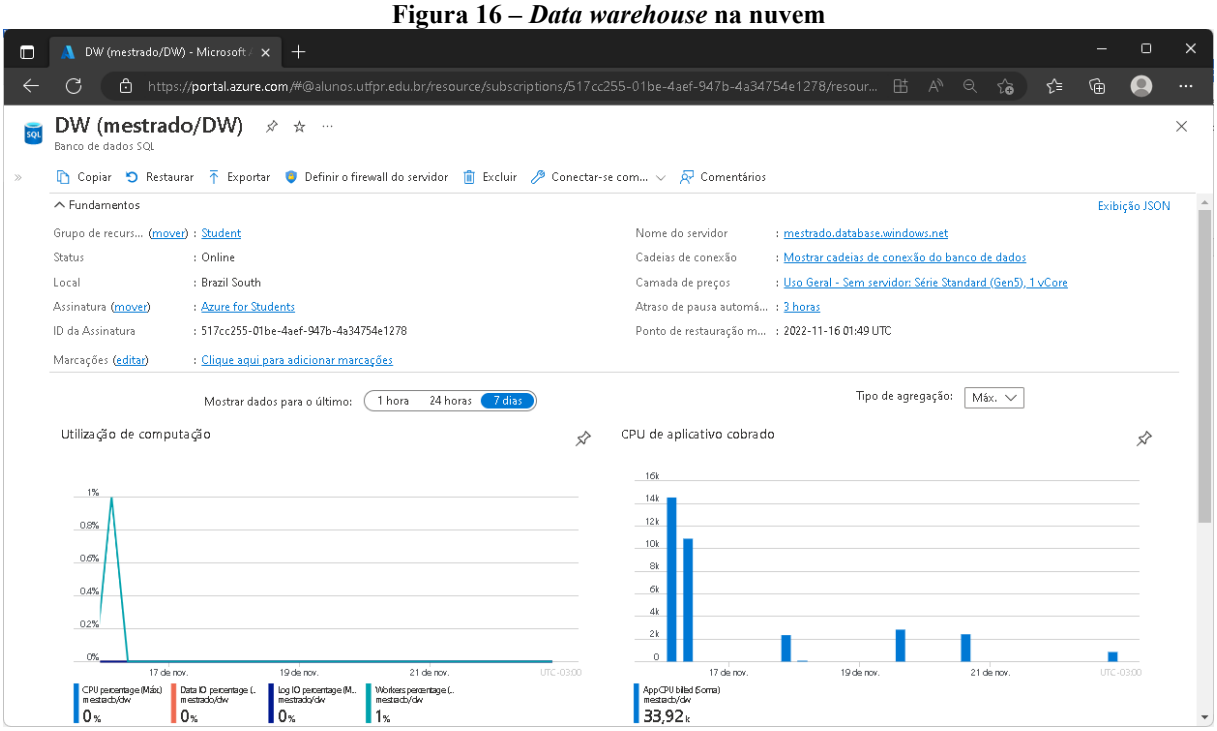
Fonte: elaborada pelo autor

Na camada de preparo de dados, os dados primários contidos nos conjuntos de dados preliminares, que estavam em formato de arquivo *flat*, passaram por limpeza, transformação e foram armazenados em um esquema que foi implementado exclusivamente para *staging* no sistema de gerenciamento de banco de dados utilizado no projeto.

Como componente do sistema de apoio à decisão deste trabalho, foi implantado um servidor de banco de dados SQL como serviço de computação em nuvem, utilizando uma conta

no serviço *Azure* de *cloud computing,* visando fluidez e baixo custo. A Figura 16 mostra a utilização da tecnologia da implementação em nuvem aplicada neste estudo.

Figura 16 – *Data warehouse* na nuvem



Fonte: Tela capturada do Azure em nov/22

4.3.3 Camada de preparo de dados

A camada de preparo é uma área para um estágio anterior ao destino dos dados, onde dados brutos são copiados, preparados e alocados para depois avançarem para o DW (Van Der Lans, 2012). Nesta camada pode haver materialização de visões sobre os sistemas de produção, significando que o resultado da consulta é armazenado como uma tabela concreta que precisa ser mantida após atualizações incrementais em lote conforme necessidades do projeto (Lwin; Thida, 2009).

Fazendo uso dessa abordagem, a extração de dados materializados na *staging* para a camada principal do *data warehousing*, agiliza os *pipelines* de ETL e assegura a confiabilidade, visto que os dados originários já passaram por limpeza; No entanto, conforme Van Der Lans (2012, p. 157), “em um mundo ideal, a limpeza de dados é totalmente tratada pelos próprios sistemas de produção, sendo que esses sistemas devem ser desenvolvidos de modo a que se torne quase impossível para os utilizadores introduzir dados incorretos”.

Engenheiros de dados também podem usar *data lakes* como áreas de preparação ou servir de fontes para os *data warehouses*, sendo áreas de preparação, que podem armazenar cópias dos dados brutos, como uma etapa no caminho para o DW, ou armazenar conjuntos de resultados temporários, como parte de processos de ETL (Llave, 2018).

4.3.4 Camada central

A camada principal de *data warehousing* refere-se ao armazém de dados corporativo em si, onde de fato acontecem as transformações necessárias para o armazenamento de dados de forma definitiva em arquitetura multidimensional de dados, sendo que a capacidade de armazenamento aumenta conforme a granularidade dos dados. Após os dados serem limpos e transformados na área de preparo eles são destinados ao repositório central, *data warehouse*, conforme o modelo físico de dados implementado neste projeto apresentado na Figura 13.

4.3.5 Camada semântica

Esta camada diz respeito a uma alta abstração em termos de negócio. Nessa modelagem, realizam-se cálculos e agregações. Cada cubo formado de um subconjunto de dados do DW visa atender um determinado cenário do negócio na tomada de decisão. Os sistemas de DW desempenham um papel importante nos cenários de TI das empresas atuais e para lidar com os complexos processos de transformação e grandes volumes de dados. Eles precisam ser complementados por metadados (Reisser; Priebe, 2009).

Quando se utiliza o termo “semântica”, tem-se em mente um modelo lógico formal para representar o conhecimento, meios para analisar e explorar com eficiência grandes quantidades de dados com recursos OLAP, dados apresentados formalmente que devem ser organizados em um esquema multidimensional bem definido (Nebot; Berlanga, 2012).

O conceito de camada semântica foi patenteado pela *Business Objects*, que menciona que se trata de um sistema de acesso a banco de dados relacional usando objetos semanticamente dinâmicos, permitindo realizar consultas sem se preocupar com a estrutura relacional ou a linguagem de consulta (Cambot; Liautaud, 1996).

Mundy (2013) realça que a camada semântica é um dos principais componentes da arquitetura de sistema de apoio a decisão baseado em dados, fornecendo uma tradução das estruturas subjacentes em termos orientados ao usuário de negócios. Assim, um modelo

semântico de dados é criado para representar e aprimorar os dados para uso em aplicativos de relatório e análise.

4.3.6 Camada de apresentação

A camada de apresentação é a camada superior da arquitetura do sistema de DW. Refere-se à forma como os dados serão apresentados aos clientes finais, onde se constroem objetos de visualização de dados, *dashboards*, análises visuais e relatórios, a fim de fornecer uma interface ao usuário do negócio como produto final. Esta camada da arquitetura é o local em que a interface *front-end* apresenta visualmente os dados processados do sistema OLAP, que estarão dispostos para que os usuários da organização realizem suas análises e consultem os formatos de relatórios de BI.

4.4 **Demonstração**

O primeiro protótipo do SAD em homologação foi apresentado a um professor do NDE/colegiado do BSI em Curitiba.

Este professor atuou na fase de validação fornecendo *feedbacks* e tirando dúvidas do pesquisador, que atuou realizando demonstrações técnicas de usabilidade da ferramenta e explicação sobre as informações dispostas no artefato. No meio deste caminho, algumas alterações foram realizadas no protótipo no sentido de *design* e novas análises implementadas, a fim de que o objeto a ser disponibilizado atendesse ao máximo o grupo de usuários e mitigasse quaisquer dúvidas de entendimento de indicadores e usabilidade.

Esta etapa de validação das telas e dos dados durou alguns meses, havendo várias reuniões técnicas do desenvolvedor com o professor representante do grupo de decisores que se incumbia de trazer algumas informações internas para subsidiar o desenvolvimento do SAD. Ele também solicitou algumas alterações.

Essas alterações eram coletadas pelo pesquisador que, de forma ágil, tentava já implementar, gerando novas versões até alinhar o entendimento com o utilizador sobre os indicadores e visões de análise implementadas no painel. Assim, a prototipação incremental foi o método de desenvolvimento de *software*, por meio do qual o artefato evoluiu até chegar em uma versão madura para liberação, significando que o protótipo pôde ser refeito e adaptado até se alinhar às necessidades de informação e uso (Sommerville, 2016).

Em seguida, pensando em apresentar os dados ao grupo de decisores do NDE/colegiado do curso BSI, foi implementado um *dashboard* baseado em computação em nuvem, dispondo de uma *interface* que centraliza várias visões de análise, onde cada página *web* busca focar em um determinado objetivo de análise. A intenção era permitir ao grupo de usuários a possibilidade de interagir com os dados e realizar análises de forma dinâmica e com vários modos de visualização de dados, a fim de melhor entender os indicadores e dados sobre o objeto de estudo.

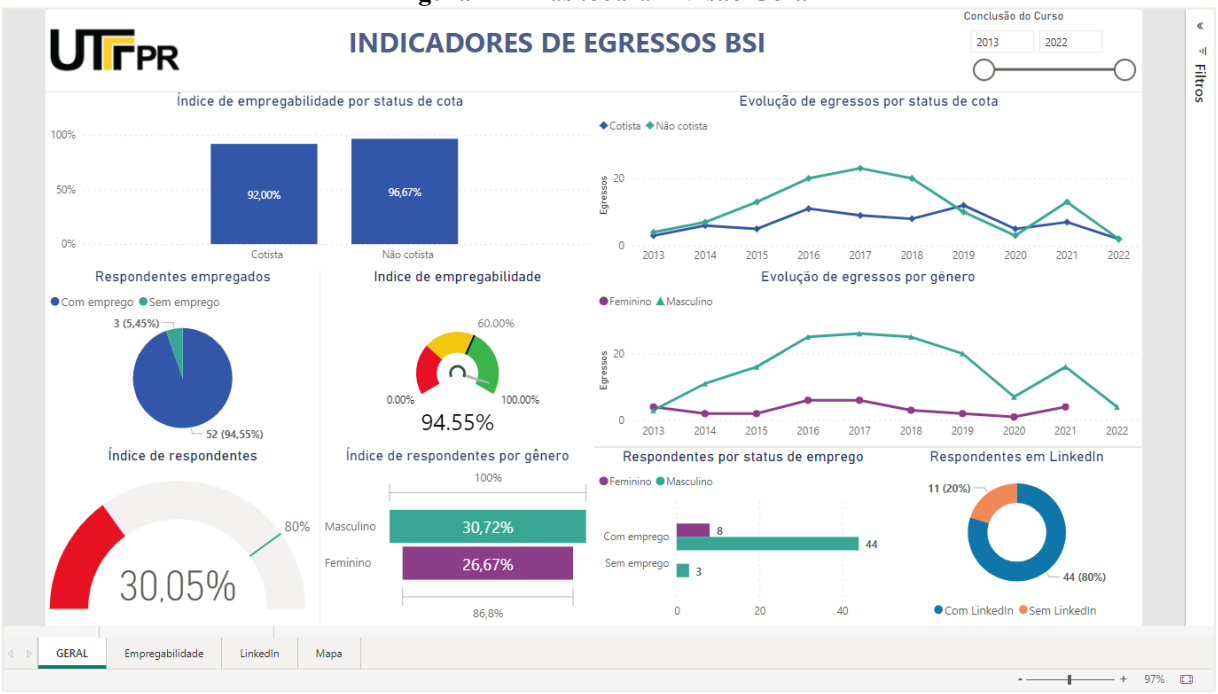
Visando realizar a implantação do artefato no ambiente corporativo de TI, foram realizadas implementações no ambiente integrado da *Microsoft* de plataforma de computação em nuvem, buscando garantir a segurança da informação e objetivando que o grupo de usuários composto por catorze professores, tivesse acesso ao objeto usando credenciais de *e-mail* corporativo e de rede.

O artefato foi disposto por meio dos canais *web*: https://utfpredubr.sharepoint.com/sites (portal *web* implementado dentro da rede *intranet* da organização) e https://bit.ly/bsidashboard (contendo o *dashboard* de forma incorporada em página *web* na *internet*).

Acompanhando os *links* foi produzido um material em vídeo orientativo sobre como acessar e utilizar os recursos da ferramenta para análise de dados, filtros e informações contextuais para apoiar o decisor.

A Figura 17, apresenta a tela principal do artefato. Nesta visualização buscou-se reunir os principais indicadores-chave de desempenho sobre o tema de análise, almejando trazer uma visão geral sobre os egressos do curso BSI aos decisores, representando graficamente os dados presentes na arquitetura implementada, conforme mostrado na Figura 13.

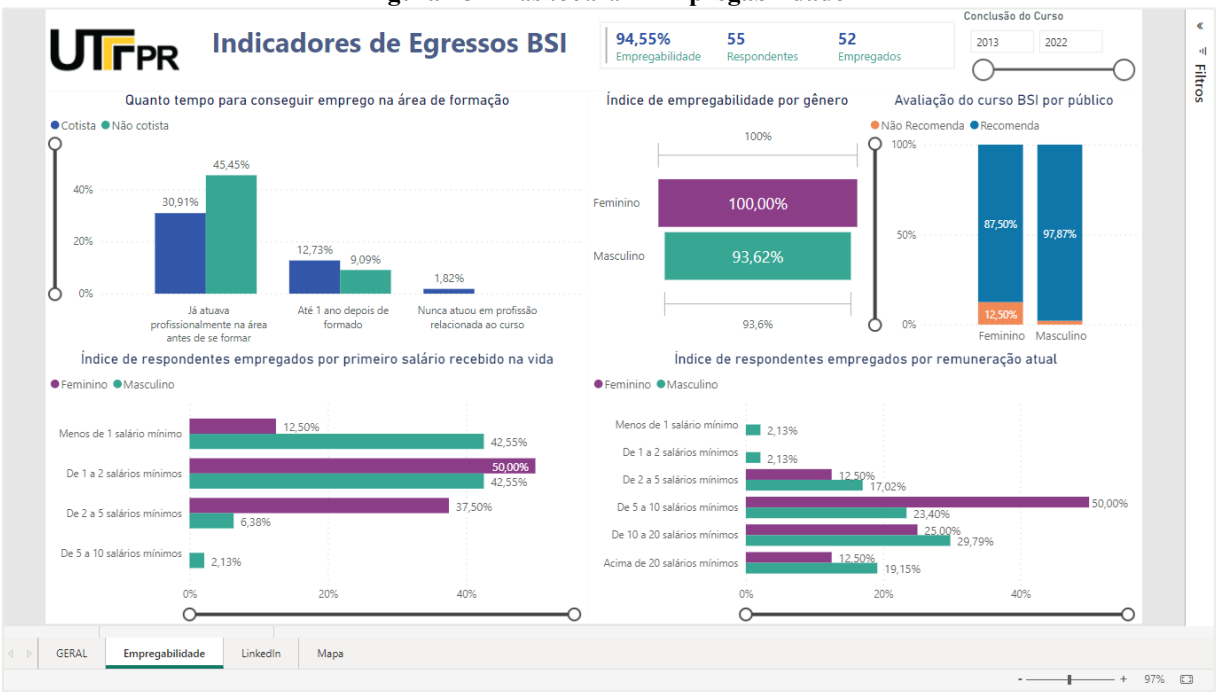
Figura 17 – *Dashboard* – Visão Geral



Fonte: elaborada pelo autor

A partir desta página (Figura 17) buscou-se oferecer uma visão geral sobre os egressos ao decisor e responder, da esquerda para direita: índice de empregabilidade por *status* do cotista; situação de emprego dos egressos que responderam à pesquisa; índice de empregabilidade sinalizando a meta e o *status*; índice de respondentes sinalizando o percentual de atingimento da meta; índice de respondentes por gênero; evolução dos egressos por *status* de cota; evolução de egresso por gênero; situação de emprego dos egressos respondentes por gênero; e comportamento dos egressos respondentes no *LinkedIn*.

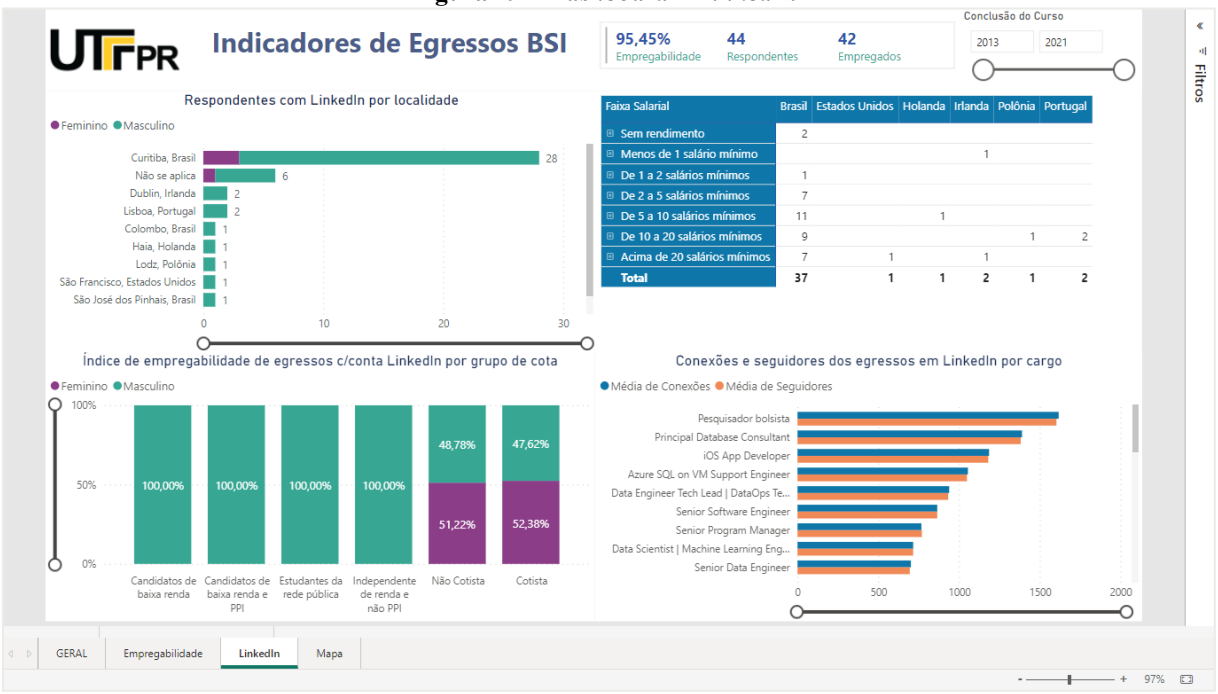
Figura 18 – *Dashboard* – Empregabilidade



Fonte: elaborada pelo autor

Na Figura 18, a visualização buscou ter o enfoque na análise de empregabilidade dos egressos e proporcionar informações, da esquerda para direita, sobre: a demora dos egressos para conseguir emprego na área de formação por *status* de cota; o índice de respondentes empregados por faixa de renda do primeiro salário recebido na vida e por gênero; o *ranking* do índice de empregabilidade por gênero; a avaliação do curso de BSI por tipo de público; e a faixa de salário atual.

Figura 19 – *Dashboard – LinkedIn*

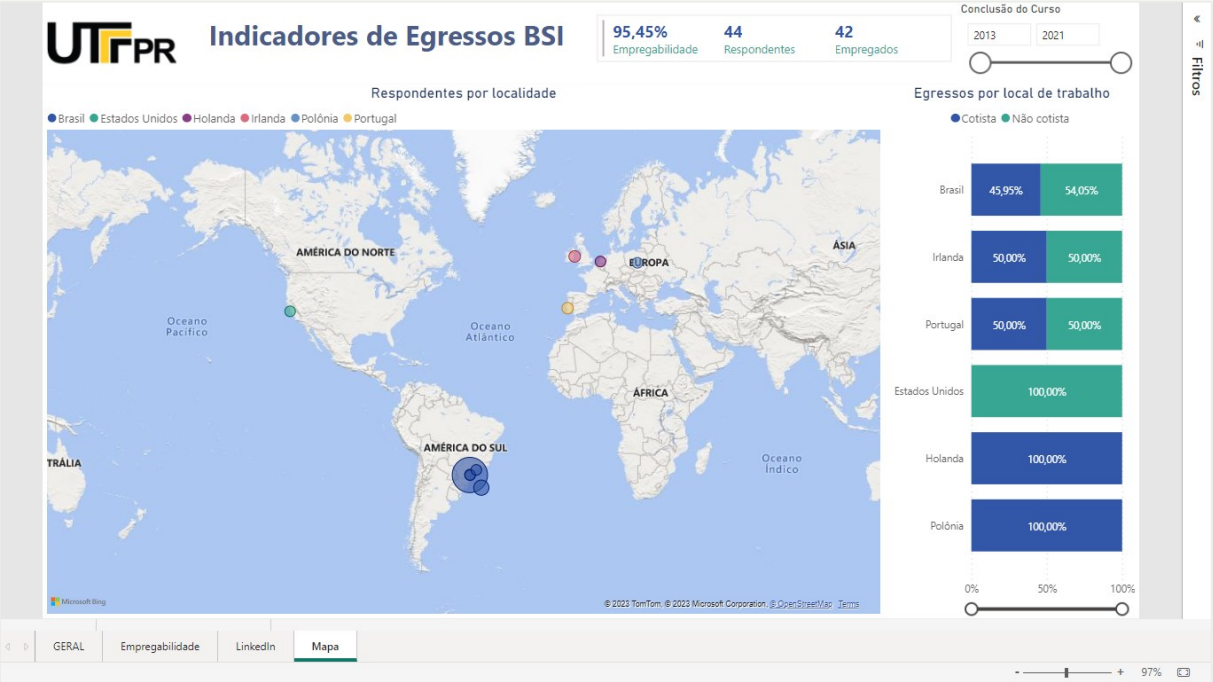


Fonte: elaborada pelo autor

Na Figura 19, a visualização focou no engajamento dos egressos no *LinkedIn* e em responder, da esquerda para direita: onde estão os egressos conforme *LinkedIn* por gênero; o índice de empregabilidade daqueles que possuem conta no *LinkedIn* por grupo de cota e tipo de cota ao realizar *drill down*; as faixas salariais desses egressos com emprego e a distribuição por país de localidade e cargo conforme informado na rede social *Linkedin*; e o comparativo de média de conexões e seguidores por cargo do perfil do egresso no *Linkedin*.

Já a Figura 20, a seguir, de forma sucinta, buscou se concentrar na geolocalização, apresentando a distribuição dos egressos por localidade no mapa *mundi*, os locais de trabalho dos egressos cotistas e não cotistas e a possibilidade de se analisar o índice de empregabilidade por país de trabalho, quando o usuário desejar realizar filtros na análise de forma interativa.

Figura 20 – *Dashboard* – Geolocalização



Fonte: elaborada pelo autor

No meio o desenvolvimento da aplicação *web* para interface com usuário final, foram enfrentadas dificuldades no que tange ao ambiente corporativo de TI de computação em nuvem. Seria necessário o suporte técnico dos profissionais de TI da instituição para deixar todos os *e-mails* corporativos dos usuários em um mesmo domínio na nuvem *Azure*, pois algumas contas de usuários não apareciam no *active directory*, por não estarem disponíveis dentro do domínio principal na *Azure,* de modo que fosse possível adicioná-las em um grupo que o implementador havia criado para deixar vinculado aos objetos construídos ou até mesmo adicionar as contas de *e-mail* corporativas em objetos criados no ambiente *Office* 365 da nuvem. Optou-se por não acionar a equipe de suporte institucional que estava assoberbada com outras demandas da própria universidade.

Como forma paliativa de liberar o artefato para o maior número de usuários da instituição acadêmica que faziam parte do grupo de decisores ou partes interessadas, foi implementado um portal dentro da rede interna (*intranet*) e autorizados todos os *e-mails* que o ambiente em nuvem permitia. O artefato também foi disponibilizado por meio de uma página *web*, que não exigia autenticação, implementada em um ambiente de nuvem *Azure* para disponibilização dos *dashboards* e gerenciamento dessas tecnologias para demonstração do artefato ao grupo de decisores.

Sendo assim, o pesquisador gerenciou dois modos de disposição do artefato aos decisores, um hospedado dentro da *Azure* da própria organização e um outro implementado apenas para fins de demonstração da interface do sistema, evitando impeditivos de acesso para com os usuários. No entanto, devido aos custos, a segunda opção teve caráter temporário, apenas para a demonstração, em decorrência do padrão de licenciamento das plataformas utilizadas na implementação pela *Microsoft*, provedor de nuvem utilizado.

4.5 Avaliação

A avaliação do artefato fornece informações de *feedback* e uma melhor compreensão do problema para melhorar tanto a qualidade do produto quanto o processo de *design* (Hevner *et al.*, 2004).

Seguindo a abordagem exposta na Figura 10, no final do ciclo desta pesquisa, buscou-se realizar avaliações considerando a finalização do ciclo de vida de desenvolvimento do sistema, foi realizada a disponibilização do protótipo inicial aos usuários e, após isso, foram elencadas algumas perguntas principais, documentadas conforme o Quadro 2.

O formulário com as questões apresentadas no Quadro 2 foi disponibilizado por meio da plataforma *Microsoft Forms* destinada ao grupo de decisores formado por membros do Colegiado e NDE do BSI, a fim de coletar depoimentos sobre a utilização, qualidade e eficácia do artefato (Hevner *et al.*, 2004).

Após a coleta das avaliações do grupo de decisores do NDE/colegiado do BSI, os resultados da *survey* foram agregados a fim de comprovar a potencial contribuição do artefato projetado para o fomento de decisões coletivas e apoio no processo decisório e de planejamento estratégico do curso de bacharelado em sistemas de informação.

Os principais resultados da avaliação pelos decisores estão documentados na Figura 21, que apresenta a análise das médias das questões em escala *Likert* de 1 a 7 pontos.

Sobre a participação dos entrevistados, foram 42,86% de mulheres e 57,14% de homens na pesquisa, onde 50% do total dos catorze membros do grupo de decisores foram respondentes e, sobre o tempo de experiência atuando como professores existe uma mediana de 11 anos, com média de 15 anos, mínimo de 6 anos e máximo de 25 anos.

Figura 21 – Média de avaliações dos decisores

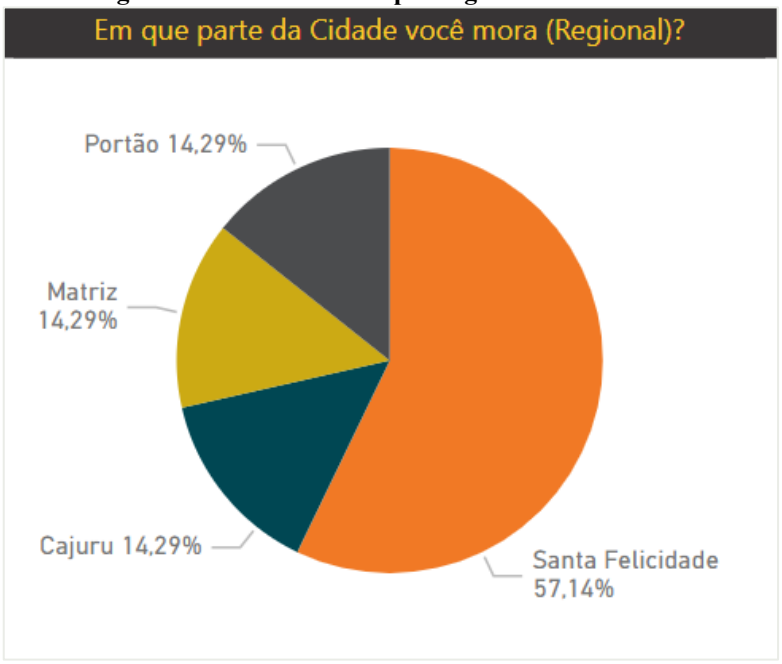
Gênero do Decisor	O dashboard de indicadores BSI ajuda a compreender a situação dos egressos do curso BSI?	O dashboard facilita as tomadas de decisões de forma coletiva sobre o curso BSI?	Os dados da forma que foram estruturados permitem analisar/explorar os temas em várias perspectivas?	Os indicadores de empregabilidade ajudaram no entendimento da situação atual dos egressos BSI?	O dashboard ajuda na tomada de decisão do colegiado/NDE do BSI?
Mulher	6,67	5,67	5,33	6,67	5,67
Homem	6,50	5,25	6,25	6,75	6,00
Média Geral	**6,57**	**5,43**	**5,86**	**6,71**	**5,86**

Fonte: elaborada pelo autor

Percebe-se a satisfação de grupo de decisores que responderam ao questionário em relação à usabilidade e solução aos problemas do negócio que o artefato buscou atender neste campo de estudo de dados empíricos.

A Figura 22 mostra a distribuição da quantidade de respondentes por regional da cidade de Curitiba, a fim de avaliar se usando o artefato de forma remota auxiliaria em decisões coletivas em reuniões feitas via ferramentas de videoconferência, considerando que o artefato faz uso de tecnologia de computação em nuvem.

Figura 22 – Entrevistados por regional de residência

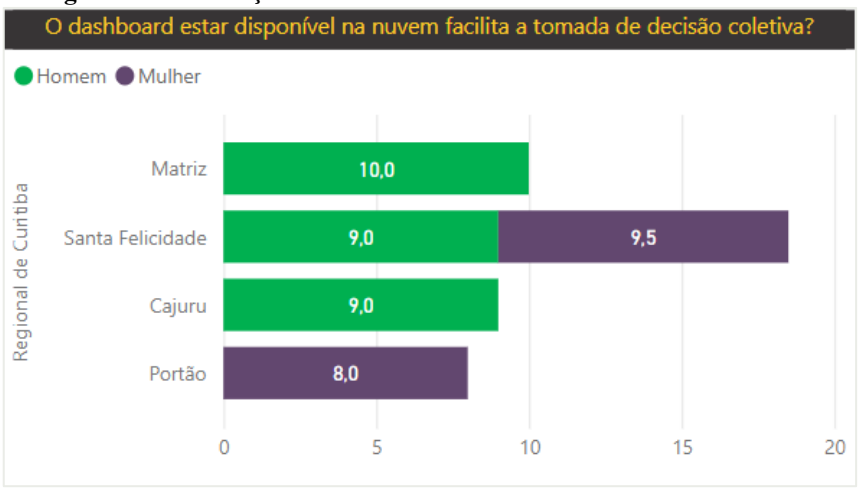


Fonte: elaborada pelo autor

Sobre o a questão de o artefato estar disponível em nuvem facilitando a tomada de decisão em grupo baseando em dados, a média das avaliações dos respondentes na questão que utilizou escala de 1 a 10 pontos também aponta para a satisfação dos usuários,

independentemente do gênero dos entrevistados e da regional de residência, conforme mostra a Figura 23.

Figura 23 – Avaliação sobre o fomento de tomada de decisão coletiva



Fonte: elaborada pelo autor

Infere-se que o artefato ser baseado em nuvem fomenta a tomada de decisão em grupo, pois facilita a colaboração nas decisões tanto em reuniões presenciais quanto de forma remota, em casos de adesão de forma de trabalho *home office* ou em casos de força maior, como por exemplo o caso da pandemia global do coronavírus em que se houve a obrigatoriedade do distanciamento social na população brasileira. Isto aumentou o uso de ferramentas de reuniões remotas *on-line* e de aprendizagem à distância, que levaram os professores a utilizarem sistemas de videoconferência e sistemas de aprendizagem nas rotinas de trabalho na interação com os discentes (Bernardo; Maia; Bridi, 2020).

A Figura 24 mostra as principais palavras mencionadas ao solicitarmos quais seriam as palavras que vinham à mente, relativas à experiência com o artefato. “Análise” foi a palavra com maior número de ocorrências nas avaliações dos entrevistados, fazendo jus ao tema de “*data analytics*”, que se refere ao termo de usar tecnologia para descobrir informações úteis a partir dos dados para suporte às tomadas de decisões (Goyal *et al.*, 2020).

Figura 24 – Nuvem de palavras das avaliações

Fonte: elaborada pelo autor

Sobre a alternativa na pesquisa “Quais indicadores chamaram mais atenção no *dashboard*? Por quê?”, os entrevistados alegaram que foram os seguintes itens:

a) A divisão por gêneros e a faixa salarial;
b) O índice de empregabilidade alto e a baixa diferença entre cotistas e não cotistas pois esse tipo de informação é muito difícil de acompanhar se não for a partir de *dashboards* de visualização como os implementados no trabalho;
c) A empregabilidade mostra que há ainda ampla demanda no mercado;
d) A relação salarial com gênero esclarece diferenças importantes sobre remunerações altas e a relação entre cotistas e empregabilidade mostra que essa diferença se dissipa;
e) A avaliação do curso pelos seus próprios egressos;
f) A análise de empregabilidade para tentar entender melhor quais as atuações dos egressos;
g) A evolução dos egressos mostra queda mesmo antes de período da pandemia decorrente da COVID-19;
h) Local de trabalho dos egressos.

Sobre a pergunta "Quais dos dados de empregabilidade te chamaram mais atenção e te fizeram rever alguma ideia anterior sobre o curso?”, os entrevistados mencionaram:

a) Em particular, a parte de gênero e salário;

b) Já se esperava uma empregabilidade alta, mas é sempre legal observar os índices com precisão. O período da pandemia decorrente da corona vírus COVID-19 tem características muito próprias e foi legal poder analisar com mais detalhamento;
c) Constatam-se evidências para percepções que já existiam entre os docentes do departamento;
d) O índice de empregabilidade;
e) Poucas pessoas sem emprego;
f) Local de trabalho.

Pontos negativos ressaltados: uma parte dos entrevistados esperava uma participação maior de respondentes, principalmente de mulheres.

Pontos positivos ressaltados: não houve dúvidas sobre como utilizar o artefato, realizar análises, explorar dados e obter o conhecimento nele disposto.

Sobre os comentários deixados na pesquisa, em síntese houve:

a) Sugestão de agregar mais os gráficos, divisão por seção, utilizar menos telas e deixar os filtros sempre na mesma região;
b) Expectativa de carregar os dados do PPGCA (Programa de Pós-Graduação em Computação Aplicada) para fazer análises dos dados e a curiosidade sobre a extração de dados do *LinkedIn* automatizada visto que a plataforma é complicada na questão de exportação de dados;
c) A avaliação de que construir argumentos embasados nesses dados torna o processo de tomada de decisão mais preciso;
d) A curiosidade sobre quais seriam os motivos daqueles respondentes que mencionaram que não recomendariam o curso de BSI.

Sobre o item “d” acima, fica o encaminhamento de enviar os dados dos motivos relatados na pesquisa para o professor representante do comitê do curso de BSI, visto que se tratou de um campo aberto na pesquisa onde os respondentes puderam explanar livremente suas opiniões sobre a satisfação com a capacitação no curso e atendimento.

No que tange às avalições acima, percebe-se que a ferramenta pode contribuir na análise dos egressos, promovendo *insights* relevantes. Contudo, uma limitação do trabalho é que não se explorou em profundidade a forma como o artefato pode apoiar a decisão por parte dos integrantes dos colegiados envolvidos, embora, nas respostas ao questionário, tenha ficado claro que os respondentes gostaram de poder dispor de um *dashboard* com informações relevantes sobre muitas das questões importantes para o planejamento do curso, relativas aos seus egressos, ao longo do tempo.

5 CONSIDERAÇÕES FINAIS

Organizações que percebam o valor de seus dados e fomentem a cultura de tomada de decisões baseadas em dados (DDDM) podem capacitar seus membros a tomarem melhores decisões com seus dados no dia a dia corporativo e a identificarem premeditadamente quais serão as próximas oportunidades estratégicas, diante de um mercado competidor e exigente.

Neste trabalho, mostrou-se um protótipo de sistema de apoio à decisão orientado a dados, que permite a análise de dados empíricos relacionados aos egressos do curso de BSI do *campus* Curitiba da UTFPR. Foi apresentado um ciclo de desenvolvimento da engenharia de dados, desde a fase da definição de requisitos até a disponibilização de interface aos usuários finais, representado pelo grupo de decisores do NDE/colegiado de curso do BSI.

Assim, objetivou-se desenvolver um artefato baseado em tecnologia da informação para contribuir na tomada de decisão. A avaliação do artefato demonstrou sua utilidade e eficácia no ambiente do problema.

Os principais *insights* foram inerentes à empregabilidade dos egressos sendo analisada por perfil de cotas, gênero, faixa salarial, local de trabalho e outras informações disponíveis e passíveis de cruzamentos.

Com base na avaliação dos decisores, o trabalho permitiu concluir que a engenharia de dados poderá contribuir para o processo de decisão daqueles que definem a estratégia do curso de bacharelado em sistemas de informação da UTFPR. O sistema demonstrou-se capaz de fomentar o processo decisório, subsidiando os envolvidos com informação adequada. O fato de operar em nuvem também contribui para participação coletiva e remota nas decisões, facilitando a análise dos KPIs e outros dados importantes para as decisões.

Por meio de reuniões do grupo de decisores, onde é possível operar os *dashboards* disponibilizados, apresentando KPIs sobre os egressos do BSI, contendo dados do sistema acadêmico e externos, foi possível perceber que o artefato contribui para o estímulo de tomada de decisões de forma coletiva e colaborativa pelo NDE/colegiado de curso.

No decorrer do trabalho, foi possível compreender o papel da engenharia de dados na estruturação do artefato pretendido, fornecendo suporte ao processo de tomada de decisão com a implementação do projeto. Compreendeu-se que um engenheiro de dados obtém dados de várias fontes e fornece valor a partir dos dados para uso em *analytics* e ciência de dados. Além disso, foi possível perceber a colaboração da computação em nuvem para a engenheira de dados, por meio da demonstração dos recursos da *Azure* para os serviços de dados e análise.

Entretanto, uma limitação da pesquisa é não acompanhar de fato se decisões serão tomadas com base no uso da ferramenta e nos dados dispostos por ela. Isso se dá por circunstância do pesquisador não estar presente no ambiente organizacional e o ato de analisar as tomadas de decisões não ser uma variável que pôde ser controlada por esta pesquisa.

Uma outra questão de limitação na pesquisa é que os dados do sistema da universidade foram disponibilizados em planilhas eletrônicas, assim, o implementador realizou a técnica de *data wrangling* na engenharia de dados para construção do artefato. Pensando em trabalhos futuros, se for do interesse da instituição ter um ou mais artefatos na estrutura computacional interna para a análise de egressos ou outros temas, seria ideal permitirem o acesso aos sistemas OLTP para facilitar a automatização na camada de integração de dados, garantindo atualização dos dados com maior recorrência para os sistemas de apoio à decisão (SAD).

Uma outra questão também é que, se for implantar o artefato ou um outro protótipo inteiramente em ambiente de produção da universidade, toda estrutura que foi implementada em ambiente paliativo por parte do pesquisador para engenharia de dados, codificação, sistemas de gerenciamentos de banco de dados, *data warehouse*, *middleware*, serviços de computação em nuvem deverão ser implantados na arquitetura computacional da instituição acadêmica.

Espera-se que este trabalho tenha contribuído para o entendimento das atividades ligadas à engenharia de dados e da sua importância para proporcionar os dados certos, de forma confiável, àqueles que os analisarão e se basearão neles para a tomada de decisões.

Como contribuições no âmbito científico e tecnológico, fica o modelo de processo de engenharia de dados para análise de egressos para que outros profissionais e pesquisadores possam aplicar em seus problemas de campo e de estudo, bem como o entendimento, delineamento e aprofundamento no que tange a engenharia de dados aplicada a um contexto de aplicação prática.

O presente trabalho contribui para a comunidade cientifica, trazendo fundamentos sobre a engenharia de dados, como disciplina distinta da ciência de dados, buscando mitigar pensamentos dúbios no que se refere à sua importância e atuação, bem como sanar dúvidas sobre os papéis dos profissionais na área tecnológica que atuam no universo de dados.

REFERÊNCIAS

ALBANO, Antonio. **Decision Support Databases Essentials.** Univ. Pisa, Dep. Comput. Sci.(2015), v. 138, 2015.

AMARAL, F. **Introdução a ciência de dados, mineração de dados e Big Data**. Rio de Janeiro: Alta Books, 2016.

ANDERSON, Jesse. **Data Teams:** A Unified Management Model for Successful Data-Focused Teams. 1. ed. Reno: Apress. 2020.

ASTROVA, I.; KOSCHEL, A.; EICKEMEYER, C.; OFFEL, N. **Comparison of dbaas architectures.** In: 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, 2018. p. 1-5.

ATWAL, Harvinder. **Organizing for dataops. Practical DataOps:** Delivering Agile Data Science at Scale, p. 191-211, 2020.

BALTZAN, Paige. **Tecnologia orientada para gestão.** 6. ed. Porto Alegre: AMGH, 2016.

BEACH, Daniel. **Introduction to Data Engineering:** Learn the skills needed to break into Data Engineering. Kindle Edition, 2022.

BEARMAN, C., PALETZ, S. B.; Orasanu, J.; THOMAS, M. J. **The breakdown of coordinated decision making in distributed systems.** Human factors, 2010.

BEAUCHEMINM, M. **The Rise of the Data Engineer**. Jan. 2017. Disponível em: <https://medium.freecodecamp.org/the-rise-of-the-data-engineer-91be18f1e603/>. Acesso em: 21 abr. 2020.

BERNARDO, Kelen A. S.; MAIA, Fernanda Landolfi; BRIDI, Maria A. **As configurações do trabalho remoto da categoria docente no contexto da pandemia Covid-19.** Novos Rumos Sociológicos, v. 8, n. 14, p. 8-39, 2020.

BRODBECK, Felix C. et al. **Group decision making under conditions of distributed knowledge:** The information asymmetries model. Academy of Management Review, v. 32, n. 2, p. 459-479, 2007.

CAI, Li; ZHU, Yangyong. **The challenges of data quality and data quality assessment in the big data era.** Data science journal, v. 14, 2015.

CAMBOT, Jean-Michel; LIAUTAUD, Bernard. **Relational database access system using semantically dynamic objects.** U.S. Patent n. 5,555,403, 10 set. 1996.

COOK, H.; ADRIAN, M.; GREENWALD, R., GU, X**. Magic Quadrant for Cloud Database Management Systems.** 2022. Disponível em: < https://www.gartner.com/doc/reprints?id=1-2C0ZJ2K2&ct=221216/>. Acesso em: 14 dez. 2022.

CÔRTES, Pedro Luiz. **Administração de sistemas de informação**. São Paulo: Saraiva, 2008.

CRESWELL, John W.; CRESWELL, J. David. **Projeto de pesquisa:** Métodos qualitativo, quantitativo e misto. Porto Alegre: Penso. 2021.

DATABRICKS. **Data Lakehouse Platform.** 2022. Disponível em: <https://www.databricks.com/product/data-lakehouse/>. Acesso em: 24 jul. 2022.

ECKERSON, Wayne W. **Performance dashboards:** measuring, monitoring, and managing your business. 2. ed. Wiley, 2010.

ECKERSON, Wayne. **Who ensures clean, consistent data?** The Data Warehouse Institute, 2009.

ELDOR, Liat. **How collective engagement creates competitive advantage for organizations:** A business-level model of shared vision, competitive intensity, and service performance. Journal of Management Studies, v. 57, n. 2, p. 177-209, 2020.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa.** 7. ed. São Paulo: Atlas, 2019.

GIL, Antonio de Loureiro. **Processos Decisórios:** SIG / ERP; Controle Interno; Metodologia / Projeto / Sistema Decisão. São Paulo: Kindle Edition, 2017.

GOYAL, Deepti; GOYAL, R.; REKHA, G.; MALIK, S.; YYAGI, A.K. **Emerging trends and challenges in data science and big data analytics.** In: 2020 International conference on emerging trends in information technology and engineering (ic-ETITE). IEEE, 2020. p. 1-8.

HEVNER, A. R., MARCH, S. T., PARK, J., RAM, S**. Design science in information systems research.** MIS quarterly, p. 75-105, 2004.

HOFFMAN, Jason. **CLOUD COMPUTING:** A Complete Guide on the Concepts and Design of Cloud Computing (SaaS, PaaS, IaaS, Virtualization, Business Models, Mobile, Security and More). Kindle Edition, 2020.

INMON, Bill; LEVINS, Mary; SRIVASTAVA, Ranjeet. **Building the Data Lakehouse.** 2021.

INMON, W. H.; LINSTEDT, D. **Data Architecture:** A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault. Elsevier. 2015.

INMON, William H.; STRAUSS, Derek; NEUSHLOSS, Genia. **DW 2.0:** The Architecture for the Next Generation of Data Warehousing. Elsevier, 2010.

KANSAL, S., SINGH, G., KUMAR, H., KAUSHAL, S. **Pricing models in cloud computing.** In: Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies. 2014. p. 1-5.

KAPLAN, Robert S.; NORTON, David P. **Mapas estratégicos:** balanced scorecard - Convertendo ativos intangíveis em resultados tangíveis. Brasil, Alta Books, 2018.

KAVIS, Michael J. **Architecting the Cloud** (Wiley CIO). Wiley. Kindle Edition, 2014.

KIMBALL, Ralph; ROSS, Margy. **The kimball group reader:** Relentlessly practical tools for data warehousing and business intelligence remastered collection. John Wiley & Sons, 2015.

KUMAR, Ratnesh; TAKAI, Shigemasa. **Inference-based ambiguity management in decentralized decision-making:** Decentralized control of discrete event systems. IEEE Transactions on Automatic Control, v. 52, n. 10, p. 1783-1794, 2007.

KUROSE, J. F; ROSS, K. W. R. **Redes de computadores e a Internet**: uma abordagem top-down/. 6. ed. São Paulo: Pearson Education do Brasil, 2013.

LAUDON, Kenneth C.; LAUDON, Jane P. **Sistemas de informações gerenciais**. 7. ed. São Paulo: Pearson Prentice Hall, 2007.

LINKEDIN. **Search Results for Free and Premium Members**. Disponível em: <https://www.linkedin.com/help/linkedin/answer/a520842/what-you-get-when-you-search-on-linkedin?lang=en/>. Acesso em: 09 mar. 2022.

LLAVE, Marilex Rea. **Data lakes in business intelligence:** reporting from the trenches. Procedia Computer Science, v. 138, p. 516-524, 2018.

LUELLEN, Eric. **Why Data Science Succeeds or Fails.** Towards Data Science [online], 2018.

LWIN, Yupar Kyaw; THIDA, Aye. **Data Extraction Using Materialized Views.** MERAL Portal, 2020. Disponível em: <https://meral.edu.mm/record/3858/files/55026.pdf/>. Acesso em: 20 set. 2022.

MACEY, Tobias. **97 Things Every Data Engineer Should Know:** Collective Wisdom from the Experts. O'Reilly Media, 2021.

MANN, Robert I.; WATSON, Hugh J. **A contingency model for user involvement in DSS development.** MIS Quarterly, p. 27-38, 1984.

MARTIN, Wendy L.; MCKELVIE, Alexander; LUMPKIN, G. Tom. **Centralization and delegation practices in family versus non-family SMEs:** a Rasch analysis. Small Business Economics, v. 47, p. 755-769, 2016.

MEDEIROS, L. F. de. **Banco de dados**: princípios de prática. Curitiba: InterSaberes, 2013.

MICROSOFT. **O que é computação em nuvem**. 2022. Disponível em: <https://azure.microsoft.com/pt-br/overview/what-is-cloud-computing/>. Acesso em: 19 set. 2021.

MUNDY, Joy. **Design Tip #158 Making Sense of the Semantic Layer.** 2013. Disponível em: <https://www.kimballgroup.com/2013/08/design-tip-158-making-sense-of-the-semantic-layer/>. Acesso em: 20 set. 2022.

NAGLE, Tadhg; REDMAN, Thomas C.; SAMMON, David. **Only 3% of companies' data meets basic quality standards.** Harvard Business Review, v. 95, n. 5, p. 2-5, 2017.

NEBOT, Victoria; BERLANGA, Rafael. **Building data warehouses with semantic web data.** Decision Support Systems, v. 52, n. 4, p. 853-868, 2012.

NEFF, Andrew; WOODYER, Adam. **Gartner Invest Analyst Insight:** The Cloud Computing Scenario: The Future Is Distributed Cloud. [S.I.] 2020. Disponível em: <https://www.gartner.com/en/documents/3992057>. Acesso em: 24 jun. 2022.

NIRSBERGER, Henry M. **Data In The Cloud:** A Conceptual Data Model For Amazon Database Services (Visual Cloud Computing Book 2). Kindle Edition, 2020.

O'BRIEN, James A. **Sistemas de Informação e as Decisões Gerenciais na Era da Internet.** São Paulo: Saraiva, 2001.

OLIVEIRA, J. M. de; IZELLI, R. C. **Indicadores de desempenho baseados no Balanced Scorecard:** um modelo adaptado à Administração Pública. Refas - Revista Fatec Zona Sul, [S. l.], v. 4, n. 2, p. 37–51, 2018.

PINHEIRO, A. F. Concursos Teorias e Questões: Análise de Informações. 1. ed. 2017.

PONNIAH, Paulraj. **Data warehousing fundamentals for IT professionals.** John Wiley & Sons, 2011.

PRIMAK, F. V. **Decisões com B.I. (Business Intelligence).** 1. ed. Rio de Janeiro: Ciência Moderna, 2008.

PROVOST, Foster; FAWCETT, Tom. **Data science and its relationship to big data and data-driven decision making.** Big data, v. 1, n. 1, p. 51-59, 2013.

RAMZAN, Shabana; BAJWA, Imran Sarwar; RAMZAN, Bushra; ANWAR, Waheed. **Intelligent data engineering for migration to NoSQL based secure environments.** IEEE Access, v. 7, p. 69042-69057, 2019.

REIS, Joe; HOUSLEY, Matt. **Fundamentals of Data Engineering.** O'Reilly Media, 2022.

REISSER, Andreas; PRIEBE, Torsten. **Utilizing semantic web technologies for efficient data lineage and impact analyses in data warehouse environments.** In: 2009 20th International Workshop on Database and Expert Systems Application. IEEE, 2009. p. 59-63.

ROSS, Margy; KIMBALL, Ralph. **The data warehouse toolkit:** the definitive guide to dimensional modeling. John Wiley & Sons, 2013.

ROSSI, Matti; SEIN, Maung K. **Design research workshop:** a proactive research approach. Presentation delivered at IRIS, v. 26, p. 9-12, 2003.

SALTZ, Jeffrey S.; GRADY, Nancy W. **The ambiguity of data science team roles and the need for a data science workforce framework.** In: 2017 IEEE international conference on big data (Big Data). IEEE, 2017. p. 2355-2361.

SALTZ, Jeffrey S.; YILMAZEL, Sibel; YILMAZEL, Ozgur. **Not all software engineers can become good data engineers.** In: 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016. p. 2896-2901.

SANDHU, Amanpreet Kaur. **Big data with cloud computing:** Discussions and challenges. Big Data Mining and Analytics, v. 5, n. 1, p. 32-40, 2021.

SARKAR, Pushpak. **Data as a service:** A framework for providing reusable enterprise data services**.** Wiley, 2015.

SCHLEGEL, K.; SUN, J.; PIDSLEY, D.; GANESHAN, A.; FEI, F; POPA, A.; MICLAUS, R.; MACARI, E.; QUINN, K.; LONG, C. **Magic Quadrant for Analytics and Business Intelligence Platforms.** 2023. Disponível em: <https://www.gartner.com/doc/reprints?id=1-2955ETOT&ct=220215/>. Acesso em: 7 abr. 2023.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business Intelligence e análise de dados para gestão do negócio.** 4. ed. Porto Alegre: Bookman, 2019.

SOMMERVILLE, Ian. **Software Engineering.** Global Edition. Alemanha: Pearson Education, 2016.

SORDI, José Osvaldo de. **Design Science Research Methodology:** Theory Development from Artifacts. Kindle Edition, 2021.

SUNYAEV, Ali. **Internet computing:** Principles of Distributed systems and emerging internet-based technologies. Springer Nature, 2020.

SUSANTO, H.; LEU, F. Y.; CAESARENDRA, W.; IBRAHIM, F.; HAGHI, P. K.; KHUSNI, U.; GLOWACZ, A. **Managing cloud intelligent systems over digital ecosystems:** revealing emerging app technology in the time of the COVID19 pandemic. Applied System Innovation, v. 3, n. 3, p. 37, 2020.

TAMIR, Mike; MILLER, Steven; GAGLIARDI, Alessandro. **The data engineer.** Available at SSRN 2762013, 2015.

TURBAN, Efraim; McLEAN, Ephraim; WETHERBE, James; **Tecnologia da Informação para Gestão.** 3.ed. Porto Alegre: Bookman, 2004.

VAN DER LANS, Rick. **Data Virtualization for business intelligence systems:** revolutionizing data integration for data warehouses. Elsevier, 2012.

VERMESAN, O.; HARRISON, M.; VOGT, H.; KALABOUKAS, K.; TOMASELLA, M.; WOUTERS, K.; GUSMEROLI, S.; HALLER, S. **Internet of Things Strategic Research Roadmap**. IoT European Research Cluster: Brussels, Belgium, 2009.

WHITE, K. Sarah; OLAVSRUD, Thor. **What is a data engineer?** An analytics role in high demand. Ago. 2022. Disponível em: <https://www.cio.com/article/3292983/what-is-a-data-engineer.html/>. Acesso em: 10 ago. 2022.

YOUNG, Norbert. **Cloud Computing:** A to Z of Cloud Computing. Kindle Edition. 2019.

YU, Abraham Sin Oih; et al. **Tomada de decisão nas organizações:** uma visão multidisciplinar. 1. ed. São Paulo: Saraiva, 2017.

ZAKI, Asadulla Khan. **NoSQL databases:** new millennium database for big data, big users, cloud computing and its security challenges. International Journal of Research in Engineering and Technology (IJRET), v. 3, n. 15, p. 403-409, 2014.

ZEYDAN, Engin; MANGUES-BAFALLUY, Josep. **Recent Advances in Data Engineering for Networking.** IEEE Access, 2022.