Um Estudo sobre Ensino de Engenharia de Dados nas Universidades Brasileiras: Estado Atual e Perspectivas de Mercado

Tarsis Azevedo, Altigran da Silva
{tarsis.azevedo,alti}@icomp.ufam.edu.br
Instituto de Computação – Universidade Federal do Amazonas (UFAM) 69.077-000 – Manaus – AM – Brasil

RESUMO

O termo "Engenharia de Dados"(ED) tem sido utilizado frequentemente na literatura e em propostas curriculares atuais para se referir aos processos de adquirir, organizar e preparar dados para serem consumidos em análises exploratórias, como entrada de sistemas e aplicações ou outros contextos similares. Com o surgimento da área de Ciência de Dados, esse termo tem sido usado para englobar o que tradicionalmente era conhecido como gerenciamento de dados. Neste estudo, exploramos a ED no contexto acadêmico e industrial brasileiro, destacando a crescente relevância dessa área na sociedade atual e a necessidade de habilidades relacionadas a ela nos profissionais da área de computação. Este estudo foi motivado pela percepção dos autores de que os avanços de, pelo menos, uma década na indústria em ED ainda não foram adequadamente absorvidos pelo ensino de graduação das universidades. Através de levantamentos realizados sobre as disciplinas, bibliografias e ementas relacionadas à ED, em 23 universidades brasileiras e junto a empresas de tecnologia do país, construímos uma taxonomia dos tópicos ensinados atualmente e uma outra taxonomia de tópicos considerados relevantes para a indústria. O estudo comparativo destas taxonomias revelou que existe uma lacuna entre o ensino de ED e as demandas do mercado, com currículos acadêmicos frequentemente desatualizados quanto a tópicos considerados relevantes para a indústria contemporânea. Em particular, tópicos relacionados a plataformas de dados de alto desempenho, gerência de dados em nuvem e workflow de dados são destacados como grandes necessidades atuais da indústria, mas que são pouco explorados nos currículos atuais. Nosso objetivo, com esse estudo, é subsidiar mudanças nos currículos que possam contribuir para a formação de profissionais mais qualificados e alinhados às necessidades modernas do mercado.

CCS CONCEPTS

• **Social and professional topics** → Computing education.

PALAVRAS-CHAVE

Engenharia de dados, Educação de computação, Brasil, EduComp

EduComp24, Abril 22-27, 2024, São Paulo, São Paulo, Brasil (On-line)

1 INTRODUÇÃO

A digitalização acelerada de praticamente todas as atividades da sociedade moderna teve como uma de suas maiores consequências a produção de gigantescos repositórios de dados, em um volume sem precedentes na história. Atualmente, é tamanha a prevalência dos dados que eles são comparados em importância a commodities como o petróleo, ou mesmo a bens de capital, como o trabalho [4]. De fato, a quantidade de dados gerados globalmente aumenta diariamente, com estimativa de crescimento de mais de cinco vezes entre 2018 e 2025, sendo que, pelo menos, 30% desse volume será produzido pela comunicação máquina-máquina [12]. Esse fluxo gigantesco de dados está acelerando a economia e estima-se que, em 2025, o mercado de dados alcançará 2,5 trilhões de dólares globalmente [10]. Além disso, novas arquiteturas de dados estão surgindo para lidar com esse volume, como o *Data Lakehouse* [16].

Assim, um importante desafio do nosso tempo é a disponibilidade de mão de obra capacitada para lidar com esse volume de dados, e, mais ainda, ser capaz de desenvolver soluções para processá-los de forma efetiva e escalável para gerar, a partir desses dados, conhecimento que possa ser usado em benefício das pessoas, instituições e empresas. Em particular, é sabido que 80% do trabalho envolvido nesse tipo de atividade está associado a tarefas como integrar, preparar, transformar e manipular, em grande e pequena escala, os dados a serem utilizados. Neste artigo, empregamos o termo de Engenharia de Dados (ED) para nos referirmos a essas atividades [5]. Assim, nosso ponto de vista nessa pesquisa é que para atender adequadamente às demandas da sociedade e do mercado, é muito importante que os profissionais de computação formados pelas universidades dominem os principais métodos e técnicas de ED.

Nesse sentido, apresentamos neste artigo um estudo que busca responder a duas questões bastante relacionadas entre si. A Primeira é: Que assuntos têm sido ensinados em Engenharia de Dados nas Universidades brasileiras?. E a Segunda: Como estes assuntos abrangem as necessidades da sociedade e do mercado brasileiro em Engenharia de Dados? O nosso estudo se concentra nos cursos de graduação na área de computação das universidades brasileiras, onde tipicamente os tópicos de ED são cobertos em disciplinas com nome de "Bancos de Dados" ou uma de suas variações. Nós não cobrimos aqui nenhum curso ou mesmo disciplinas no escopo do que se convencionou chamar de "Ciência de Dados". Isso se deve a várias razões, entre elas, ao fato de que essa área abrange um corpo de conhecimento que intersecta ao da computação, mas que é distinto desse. Além disso, existem diversos estudos em propostas para criação de cursos de graduação específicos para essa área, cobrindo esse corpo de conhecimento. Finalmente, entendemos que as atividades de ED são inerentes da área de computação, uma vez que sua aplicação finalística não se limita a da assim chamada "Ciência de Dados", mas inclui também sistemas como busca, recomendação,

enriquecimento de documentos, processamento de imagens e texto, além de diversas outras aplicações.

Para determinar quais assuntos sobre ED têm sido ensinados nas universidades brasileiras fizemos um recorte com 23 das universidades com grupos de pesquisa mais tradicionais em ED, e analisando as disciplinas da área, fizemos um levantamento da bibliografia usada em cada um e também de suas ementas. De posse dessas informações, construímos uma taxonomia de tópicos ensinados nessas disciplinas. Ao mesmo tempo, para determinar como esses assuntos abrangem as necessidades da sociedade e do mercado brasileiros, fizemos um levantamento com várias das principais empresas do ramo de tecnologia para levantar quais habilidades técnicas seriam necessárias para um Engenheiro de Dados, recém-graduado, atender às demandas do trabalho. De novo, de posse dessas informações, construímos outra taxonomia de tópicos, que seriam essenciais para a indústria brasileira de tecnologia na área de dados. Depois, fizemos uma comparação sistemática das duas taxonomias para entender as diferenças entre os assuntos já cobertos pela Universidade e quais faltam para a indústria. Encontramos desafios no levantamento desses dados, tais como bibliografias mal formatadas; ementas mal escritas ou nenhuma ementa do curso; dificuldade quanto à disponibilidade das pessoas para responder ao levantamento, etc.

É importante mencionar, que, conforme o levantamento bibliográfico que fizemos, e que detalhamos na Seção 2, existem poucos estudos no mundo nesse sentido, e nenhum no Brasil. De fato, essa pesquisa é relevante, por fornecer uma visão das lacunas no ensino e na formação dos profissionais de dados no país. Com base nos resultados, as universidades podem adaptar seus currículos de dados, e assuntos relacionados, para melhor atender às necessidades da indústria e preparar melhor os estudantes para o mercado de trabalho. Além disso, as empresas de tecnologia podem usar os resultados para ajustar suas expectativas em relação aos recém-formados e identificar áreas onde podem fornecer treinamento adicional. A métrica de cobertura criada a partir dos resultados pode ser usada para avaliar a eficácia dos programas de educação em engenharia de dados e guiar futuras iniciativas educacionais.

Este artigo está organizado da seguinte forma, na seção 2, apresentamos um levantamento de trabalhos relacionados a nossa pesquisa, suas contribuições e em que ela se difere dos primeiros. Na seção 3, apresentamos detalhes sobre o levantamento dos dados das universidades brasileiras, mostramos como foi realizado um levantamento com as empresas de tecnologia e, como, a partir disso, criamos as taxonomias de tópicos atuais e necessários. Na seção 4, detalhamos o levantamento dos dados das disciplinas de dados das universidades brasileiras, como foi criada a taxonomia de tópicos, baseadas nas ementas e bibliografias de tais matérias, e quais conclusões tiramos fundamentados nesse levantamento. Na seção 5, detalhamos como foi feita a pesquisa com as empresas, como foi criada a taxonomia de tópicos baseadas nas respostas e quais conclusões podemos tirar desses dados. Na seção 6, detalhamos como as duas taxonomias anteriores interagem, suas principais diferenças e destaques. Por fim, temos a seção 7, onde concluímos o trabalho.

2 TRABALHOS RELACIONADOS

Nesta seção, apresentamos alguns dos estudos mais relevantes na área de engenharia de dados quanto à definição dos seus conceitos-chave e características, e também sobre sua integração em cursos de Ciência da Computação. Avaliamos suas principais contribuições e realizamos comparações com a nossa própria pesquisa.

De acordo com o documento da *Association for Computing Machinery* (ACM), que estabelece diretrizes para currículos de Ciência da Computação [3], desde a década de 1970, a área de gerenciamento de dados tem se concentrado principalmente em estudos de bancos de dados relacionais, com foco em tópicos como modelagem de dados, construção de consultas, processamento de consultas e estrutura interna dos sistemas de gerenciamento de bancos de dados (SGDBs). Podemos observar isso quando analisamos trabalhos que visam catalogar conceitos-chave da computação, como *Great Principles of Computing* de Denning [6], em que o campo de gerenciamento de dados está restrito à *Recollection*, tendo, na sua descrição, os importantes aspectos de Hierarquias, Persistência ou Compartilhamento, porém muitos outros aspectos de gerenciamento de dados parecem sub-representados [8]. Por outro lado, o volume de dados disponíveis e a serem gerenciados têm crescido exponencialmente a cada ano, aumentando a necessidade de profissionais qualificados para analisar e lidar com esses dados. Nesse contexto, o mercado busca cada vez mais esses profissionais, enquanto as universidades se esforçam para formá-los [9].

Para expandir os conceitos-chave do campo de gerenciamento de dados (GD), [8] usaram uma abordagem empírica, baseada no trabalho de Denning[6], que consistiu em gerar um sistema de categorias analisando os 6 livros mais usados no ensino de GD, e mais alguns livros com assuntos mais atualizados. A categorização dos conceitos foi feita durante o processo de análise da literatura, com todos os termos relevantes sendo inicialmente listados, seguidos da remoção de termos muito abrangentes ou muito detalhados. Durante o desenvolvimento do sistema de categorias, os termos foram selecionados a partir do livro mais abrangente, seguido dos outros, e então, manualmente agrupados. Para validar esses termos, foram analisados automaticamente 305 documentos, totalizando 9447 páginas e slides, resultando em 229 termos mencionados, pelo menos, 300 vezes. O modelo final foi baseado nas 4 perspectivas de Denning: tecnologias centrais, práticas, princípios de design e mecânicas. A partir desse processo, um modelo com vários conceitos-chave foi gerado, facilitando a inclusão de novos tópicos para ensino e a validação dos tópicos ensinados atualmente. No nosso trabalho utilizamos uma abordagem semelhante para criar as taxonomias de tópicos das bibliografias e ementas, porém analisamos uma base bem maior de livros, com 373 entradas e 108 ementas. Usamos o modelo gerado pelo trabalho de [8] para validar as taxonomias geradas, inclusive a da indústria.

A ACM também notou tais mudanças na área de gerenciamento de dados, tendo o seu relatório de 2020 [2] enfatizado a importância do envolvimento da indústria na formulação de competências para o trabalho e a necessidade de conselheiros consultivos do mercado para o desenvolvimento de um currículo significativo. A "Ciência de Dados"(CD), um novo campo da computação intimamente relacionado com a análise de dados e a engenharia de dados, é um exemplo da necessidade de uma abordagem de ensino atualizada. E em seu

novo currículo de computação [3], ainda em fase Beta, a ACM destaca que hoje, espera-se que os graduados tenham o conhecimento de um usuário, e não de um desenvolvedor. Isso implica uma maior ênfase no aprendizado de construção de consultas e modelagem de dados. Além disso, é vital que os alunos estejam cientes das tecnologias emergentes, como *NoSQL*, bancos de dados em nuvem, *mapreduce* e *dataframes*. O documento também ressalta que existe uma tensão entre o foco curricular na preparação profissional e o estudo de uma área de conhecimento como cientista, que é especialmente palpável em gerenciamento de dados. Como exemplo, a prova da completude dos axiomas de Armstrong é fundamental, mas a maioria dos graduados nunca usará tal conceito durante a carreira. Em nossa pesquisa, usamos o currículo de computação em beta da ACM para complementar nossa taxonomia de tópicos para a indústria.

Para suprir a necessidade de profissionais para lidar com tarefas de análises de dados, algumas iniciativas estão acontecendo no sentido de se criar uma nova área chamada "Ciência de Dados". A ACM [1] e a SBC [13] definiram currículos para criação de um curso de ciência de dados que visa formar profissionais capazes de suprir a demanda da indústria quanto à análise do volume exorbitante de dados gerados a cada ano. Em [1], a ACM define que "Ciência de Dados"é uma área interdisciplinar, que combina o domínio dos dados, a ciência da computação e a estatística para interrogar os dados e extrair informações úteis. Cientistas da computação desempenham um papel crucial nesse campo, trazendo métodos para armazenar, proteger a privacidade e integrar dados. Além disso, eles fornecem expertise em aplicar computação de alto desempenho em sistemas distribuídos de maneira eficiente e oferecem ferramentas para analisar, e consumir, todos os tipos de dados. No entanto, no próprio relatório a ACM admite que implementar a Ciência de Dados em um ambiente acadêmico pode apresentar desafios administrativos devido à sua natureza interdisciplinar. Também identificaram, junto à indústria, que características de computação são mais relevantes para pessoas que lidam com dados do que estatística e matemática.

Por isso, alguns trabalhos verificaram se seria possível formar profissionais com expertise em engenharia de dados, nos cursos de Ciência da Computação, sem alterar sua grade significativamente. Em [9], os autores identificaram seis competências essenciais para a CD, incluindo pensamento computacional, pensamento estatístico, fundamentos de matemática, construção e avaliação de modelos, fundamentos de algoritmos e software, e curadoria de dados. Concluíram, assim, que tais competências podem ser cobertas perfeitamente pelo currículo atual de Ciência da Computação, com pequenos ajustes em disciplinas específicas para incluir os tópicos relacionados.

Em [14], destaca-se a recente importância do que chamam de *Big Data Management Systems* (BDMS), no panorama atual de bancos de dados, ressalta-se a necessidade de integrar o estudo desses sistemas no currículo de computação, especialmente no âmbito do Gerenciamento de Dados (GD). Os autores propõem a inclusão de novas unidades de aprendizado, como *MapReduce* (com foco em *Hadoop*, *Dryad* e *MapReduce*), *NoSQL* (abrangendo *Cassandra*, *BigTable*, *MongoDB*), e *NewSQL* (por exemplo, *VoltDB*, *Google Spanner*), com objetivo de que, ao final do estudo, os alunos reconheçam as principais propriedades, forças e limitações de cada tipo de BDMS; construam aplicações usando esses sistemas e; entendam quando é mais adequado utilizar cada um deles. Essas unidades de aprendizagem propostas podem ser integradas em cursos introdutórios ou avançados. Por exemplo, uma introdução ao *MapReduce* pode ser incorporada em cursos introdutórios de bancos de dados ou sistemas distribuídos. Da mesma forma, uma visão geral de alto nível sobre *NoSQL* e *NewSQL* pode ser incluída em um curso introdutório de banco de dados, usando uma tarefa focada na exploração de tecnologias de gerenciamento de dados além dos bancos de dados relacionais.

Os autores do artigo avaliaram a integração de três módulos propostos e a eficácia dos objetivos de aprendizagem, por meio de uma pesquisa com uma turma avançada de uma disciplina de bancos de dados. Apesar do tamanho reduzido da turma, os resultados foram promissores, com distribuições de pontuação semelhantes para os módulos e a maioria dos alunos reconhecendo as propriedades de tecnologias como *MapReduce, NoSQL e NewSQL* e a utilidade de sistemas como *Hadoop, HBase e VoltDB*. Além disso, os alunos valorizaram a utilização de exercícios práticos e diagramas ilustrativos. Diante da crescente demanda por processamento escalável e distribuído de grandes conjuntos de dados, os chamados BDMS são amplamente utilizados, e os formandos frequentemente encontram oportunidades de emprego em empresas que utilizam ou planejam utilizar esses sistemas para extrair insights a partir de seus dados. Portanto, é crucial integrar o estudo desses sistemas nos currículos de computação.

Diante desses trabalhos, decidimos concentrar nossa pesquisa em Engenharia de Dados, conforme definido na Seção 3. Identificamos que as habilidades requeridas para um engenheiro de dados são totalmente abrangidas pelos currículos típicos dos cursos de Ciência da Computação. Isso nos permitiria formar profissionais de modo eficaz e rápido, sem a necessidade de criar um novo curso de graduação, evitando a burocracia e os custos associados a isso.

3 MÉTODOS

Esta seção detalha a metodologia adotada em nossa pesquisa, que envolveu a definição do que consideramos Engenharia de Dados (ED); o levantamento das disciplinas relacionadas à ED em algumas das principais instituições de ensino superior do Brasil; a normalização das bibliografias e ementas destas disciplinas; e a construção de taxonomias de tópicos, baseadas nestas informações. Também inclui um levantamento realizado com líderes de empresas de tecnologia para identificar as habilidades e tópicos mais relevantes na contratação de engenheiros de dados. Por fim, comparamos as taxonomias das ementas e das necessidades da indústria, a fim de identificar possíveis lacunas e divergências entre o ensino acadêmico e as demandas do mercado de trabalho.

A Engenharia de Dados (ED) está definida nesse artigo como o processo de adquirir, organizar e preparar dados para serem consumidos em análises exploratórias, como entrada de algoritmos ou outros contextos [5]. Em um nível mais abstrato, a ED resolve problemas relacionados à organização e à qualidade dos dados, bem como à extração de características de entidades do mundo real representadas em meio digital [11].

Com base nessa definição, selecionamos disciplinas que se relacionam diretamente com esses problemas, tais como “Bancos de Dados 1”, “Bancos de Dados 2”, “Mineração de Dados”, “Tópicos

Especiais em Bancos de Dados", "Computação em Nuvem", entre outras. Disciplinas que não abordam diretamente esses problemas, como Inteligência Artificial e Ciência de Dados, foram excluídas de nossa lista. A lista completa das disciplinas selecionadas pode ser encontrada no nosso hotsite[1].

Para determinar quais assuntos sobre Engenharia de Dados (ED) têm sido ensinados nas universidades brasileiras fizemos um recorte de 23 das universidades que mantêm vários dos grupos de pesquisa mais tradicionais na área. Fizemos um levantamento com base na grade curricular dos cursos de graduação em computação oferecidos (Ciência da Computação, Engenharia de Computação, Sistema de Informação, etc.) disponibilizada no site de cada instituição para separar as disciplinas relacionadas a "dados" de maneira geral como "Bancos de Dados", "Mineração de Dados", etc. Com isso, conseguimos levantar 108 disciplinas, e dessas, listamos suas ementas e suas bibliografias, que podem ser verificadas aqui. Filtramos as disciplinas que consideramos não relacionadas, com base na definição anterior de Engenharia de dados, ficando com 90 disciplinas na lista. Nesse levantamento, algumas universidades não puderam ser incluídas devido a diferentes motivos, tais como falta de informações disponíveis no site, indisponibilidade de currículo, links inativos no site, e currículos sem ementa ou bibliografia. Além disso, outros desafios nesse levantamento foram sites confusos e de difícil navegação, ementas incompletas, muito vagas ou indisponíveis, e bibliografias não padronizadas ou indisponíveis. Sabemos que tais problemas podem impactar na confiabilidade da pesquisa, mas acreditamos ter gerado dados suficientes para seguir com a análise.

Para normalizar os itens das bibliografias, decidimos usar o seu ISBN. Porém, nem todas as bibliografias disponibilizadas pelas universidades continham esse identificador. Para levantar os faltantes, fizemos uma busca no site da Agência Brasileira do ISBN, que faz parte da Câmara Brasileira do Livro (CBL), porém, nem todos os livros estavam disponíveis nesse site, alguns, por serem edições internacionais e outros por algum outro motivo desconhecido. Por isso, usamos o site "ISBN Search!"para completar as buscas. Nessas buscas nem sempre foi possível achar a edição exata do livro, e por isso, decidimos usar o ISBN da edição mais próxima à desejada, e colocamos essa informação na coluna "ISBN Aproximado", no arquivo disponibilizada em nosso hotsite [2]. Também, nem sempre foi possível encontrar a edição brasileira do livro, como listada na bibliografia, e, por isso, optamos por utilizar o ISBN da versão original, e colocamos essa informação na coluna "Edição Brasileira" no arquivo disponibilizado. Em algumas disciplinas, estão listadas como bibliografia, periódicos, revistas, congressos e notas de aula. Para esses, não colocamos dados de ISBN, autor e título, e descartamos das análises para taxonomia.

A fim de construir uma taxonomia dos tópicos ensinados nas disciplinas de Engenharia de Dados nas universidades brasileiras com base em suas bibliografias, fizemos um levantamento dos capítulos de cada um dos livros da bibliografia filtrada e normalizada, totalizando 874 capítulos, e investigamos os assuntos abordados em cada um deles através de seus sumários. Organizamos essas informações em um arquivo, que pode ser visto em nosso hotsite[3], que serviu como base para a criação da taxonomia de tópicos. A taxonomia tem 3 níveis, de tópicos mais genéricos a mais específicos, e pode ser vista completamente em nosso hotsite[4]. Ressalta-se que a intenção deste trabalho não foi criar uma taxonomia definitiva para tópicos relacionados a "dados", mas sim gerar uma visão geral do que é ensinado nas universidades brasileiras na área. Embora não tenhamos encontrado o sumário de 10 desses livros, consideramos que isso não prejudicou a representatividade do trabalho, uma vez que acreditamos que os assuntos tratados por esses já estão incluídos em outros livros que encontramos.

A fim de construir uma taxonomia de tópicos ensinados nas disciplinas ligadas à Engenharia de Dados nas universidades brasileiras, com base nas ementas disponibilizadas, tomamos os assuntos tratados em cada uma e mapeamos para a taxonomia de tópicos da bibliografia descrita acima. Isso foi feio porque entendemos que os assuntos tratados nas ementas são um subconjunto de todos os assuntos tratados na bibliografia de cada disciplina. Organizamos essas informações em uma imagem, que pode ser vista completa em nosso hotsite[5], que serviu como base para a criação da taxonomia de tópicos. Ressalta-se que essa taxonomia não tem intenção de afirmar o que é exatamente ensinado em cada uma dessas disciplinas, dado que a ementa é só uma representação simplista do conteúdo dado em sala de aula. Embora não tenhamos ementas em 16 das disciplinas, consideramos que isso não prejudicou a representatividade da nossa análise.

Após a construção da taxonomia de tópicos, com base nas ementas disponibilizadas, realizamos uma análise de frequência nas ementas, para compreender como os tópicos listados foram utilizados pelas disciplinas. Essa análise pode ser vista completa em nosso hotsite[6]. Para isso, procuramos os tópicos listados na taxonomia em cada ementa, e para cada ocorrência, somamos 1 à frequência. Alguns termos foram agrupados sob um mesmo tópico, como é o caso dos tópicos "Normalização", "Formas Normais", ""1FN", "2FN"73FN", que estão todos sob o tópico "Formas Normais". Embora reconheçamos que essa análise é superficial e limitada, consideramos ser relevante para este estudo, uma vez que fornece indícios do que é ensinado nas disciplinas.

Para determinar como esses assuntos abrangem as necessidades da sociedade e do mercado brasileiro, foi realizado também um levantamento com várias das principais empresas do ramo de tecnologia da informação do país. Nesse levantamento, fizemos 3 perguntas para mapear essas necessidades. A primeira pergunta foi feita em forma de múltipla escolha para determinar, de uma lista de tópicos pré-selecionados por nós, quais as habilidades técnicas necessárias para Engenharia de Dados e o nível de proficiência esperado em cada uma dessas habilidades. Também fizemos uma pergunta discursiva para ser possível justificar as escolhas na primeira pergunta; e uma terceira pergunta, também discursiva, para ser possível listar habilidades não listadas por nós na primeira pergunta. Tal levantamento foi distribuído por e-mail para pessoas

[1] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/lista-disciplinas.pdf
[2] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/bibliografia-normalizada.pdf
[3] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/capitulos_filtrados.pdf
[4] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/taxonomia-topicos-bibliografia.png
[5] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/taxonomia-topicos-ementas.png
[6] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/taxonomia-topicos-ementas-frequencia.pdf

escolhidas por nós, em posições de liderança dentro dessas empresas. E seu resultado bruto pode ser visto em nosso hotsite[7].

A partir dos resultados desse levantamento, geramos uma taxonomia de tópicos considerados relevantes para a indústria na contratação de engenheiros de dados, que pode ser conferida em nosso hotsite[8]. Para completar essa taxonomia, utilizamos outras duas fontes. A primeira fonte foi o conjunto de conceitos-chave definidos por [8], que propõem um conjunto de conceitos que descrevem o campo de gerenciamento de dados de maneira geral, englobando tecnologias recentes. A segunda fonte foi a própria experiência do primeiro autor deste artigo, como professor durante 5 anos em um bootcamp de dados que capacitou mais de 1000 alunos nas áreas de Análise de Dados, Machine Learning e Engenharia de Dados. É importante ressaltar que essa taxonomia não pretende definir o que deve abranger a formação de um engenheiro de dados, mas sim apresentar um recorte dos assuntos considerados relevantes nesta área no mercado do Brasil.

Na etapa seguinte, realizamos uma análise comparativa entre a taxonomia de ementas, que consta a frequência de assuntos abordados nas disciplinas de engenharia de dados e afins, e a taxonomia de tópicos da indústria, que foi construída a partir do levantamento com profissionais-chave da indústria e revisão de artigos relacionados. Nesse sentido, identificamos e destacamos as diferenças entre as duas taxonomias, visando aperfeiçoar a nossa compreensão sobre os tópicos mais relevantes na área de engenharia de dados, e identificar diferenças relevantes entre o que é ensinado e as necessidades da indústria.

Nas próximas seções, detalharemos os resultados dos levantamentos feitos para construir as taxonomias de tópicos das bibliografias em ementas (Seção 4), a taxonomia de tópico da indústria (Seção 5) e análise comparativa das taxonomias (Seção 6), respectivamente.

4 LEVANTAMENTO DAS BIBLIOGRAFIAS E EMENTAS DAS DISCIPLINAS

Nesta seção, apresentamos a pesquisa feita sobre a bibliografia e as ementas utilizadas nas disciplinas de Engenharia de Dados (ED). A intenção é compreender as tendências, práticas e, potencialmente, lacunas existentes na estruturação dos currículos de ED. À pesquisa incluiu um levantamento de livros de referência, a criação de uma taxonomia de tópicos abordados e uma análise das ementas das disciplinas. A investigação também procurou identificar a frequência com que diferentes tópicos são abordados nas ementas, visando revelar quais tópicos são mais enfatizados e quais podem ser negligenciados.

O levantamento da bibliografia adotada nas disciplinas pesquisadas resultou em uma lista inicial de 373 livros. Após filtragem por disciplinas relacionadas à ED, a lista foi reduzida para 250 e, em seguida, foram removidas duplicatas, gerando uma lista final de 80 livros. Dessas referências, 10 não tinham informações de sumário disponíveis. A partir dessa bibliografia, foi criada uma taxonomia de tópicos, que incluiu um universo de 245 tópicos em três níveis de profundidade. Os tópicos de cada nível variaram de genéricos a específicos. Essa taxonomia completa pode ser encontrada em nosso hotsite[9]. Na Figura 1 apresentamos uma parte dessa taxonomia com os tópicos somente até o segundo nível.

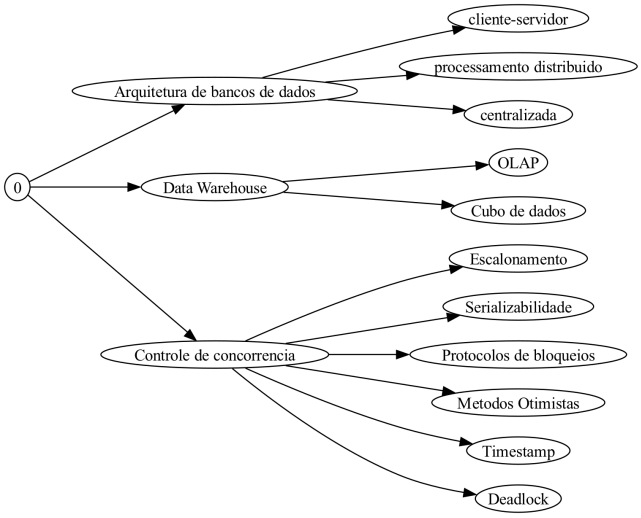

Figura 1: Amostras das taxonomias com tópicos de bibliografia.

Embora a bibliografia listada nas disciplinas de ED não represente todo o conteúdo ensinado em sala de aula, consideramos que a taxonomia resultante fornece uma base para todo o trabalho e representa o universo de assuntos que poderiam ser abordados nas disciplinas. A partir da deduplicação dos dados pelo ISBN-13 dos livros, foi possível observar que a maioria das universidades usa os mesmos livros de referência, reduzindo o universo de 250, para 80 livros (68% de redução). Alguns livros e autores dominam a bibliografia, com os top 5 livros em número de aparições nas bibliografias representando 46% de toda a lista.

A pesquisa das ementas resultou em uma lista inicial de 108 ementas, reduzida para 90 após a filtragem por disciplinas de ED, como definido na Seção 3. A partir dessas ementas, foi criada uma taxonomia de tópicos com 138 tópicos, sendo esse um subconjunto da taxonomia de tópicos da bibliografia. Foi feita uma medida da frequência da ocorrência do tópico nas ementas para entender a quantidade de vezes que cada tópico era mencionado. É importante notar que para isso os nomes dos tópicos foram normalizados. Por exemplo, os tópicos “Normalização” e “Formas normais” foram colocados sob o tópico “Formas Normais”. A taxonomia completa com as frequências dos tópicos são apresentadas no nosso hotsite[10]. Na Figura 2 apresentamos uma parte dessa taxonomia com os tópicos somente até o segundo nível.

Com essa taxonomia, foi possível obter uma lista dos tópicos mais frequentes tratados nas disciplinas de ED das universidades pesquisadas. É importante destacar que algumas disciplinas têm ementas muito abertas, especialmente as disciplinas como “Tópicos Avançados” ou "Tópicos Especiais” em bancos de dados. Essas

[7] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/resultado-pesquisa.xlsx
[8] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/taxonomia-topicos-industria.png
[9] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/taxonomia-topicos-bibliografia.png
[10] https://github.com/tarsisazevedo/sbbd23-paper/blob/main/taxonomia-topicos-ementas-frequencia.pdf

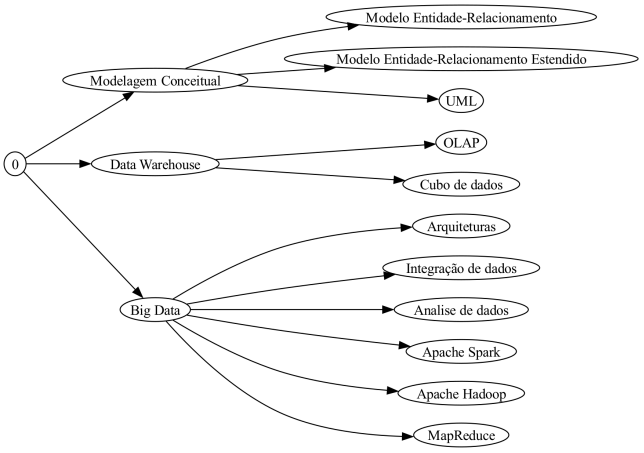

Figura 2: Amostras das taxonomias com tópicos de ementas.

disciplinas podem ser usadas para tratar de temas relevantes e atuais, embora não possamos afirmar isso com certeza nesta pesquisa, pois não tivemos acesso ao material dado em sala pelos professores de cada disciplina. Além disso, ressaltamos que essa taxonomia não representa exatamente os assuntos tratados em sala. Para uma pesquisa que reflita exatamente o que é ensinado na sala de aula, precisamos de acesso ao material usado por cada professor.

Com base na frequência de citações dos tópicos nas ementas, podemos observar que, como esperado, os tópicos relacionados a bancos de dados relacionais são amplamente abordados. Tópicos como “SQL”, “Modelo Relacional”, "Formas Normais", "Álgebra Relacional”, “Restrições”, "Gatilhos", "Projeto Físico", "Processamento e Otimização de Consultas", "Controle de Concorrência", “Transações", "Gerenciamento de Bancos de Dados Relacionais", “Arquitetura de Bancos de Dados” e “Modelo Objeto-Relacional” estão entre os mais mencionados, destacando a importância desses temas no ensino de bancos de dados atualmente. Por outro lado, alguns tópicos considerados em desuso atualmente, como modelos de dados hierárquico e em rede, bancos de dados heterogêneos, XML e bancos de dados orientados a objeto, ainda ocupam espaço nas ementas. Não podemos afirmar como tais tópicos são ensinados, se são usados para contextualização histórica ou ainda como tecnologias a serem utilizadas no dia a dia.

Tópico	**Frequência**
Bancos de Dados Relacionais	20
Controle de Concorrência	20
Processamento e otimização de consultas	18
Projeto Físico de Bancos de Dados Relacionais	16
Gerenciamento de Bancos de Dados Relacionais	16

Tabela 1: Tabela de Top 5 Tópicos e suas Frequências

Uma observação que foi possível fazer logo depois deste levantamento, foi a baixa frequência de citações para tópicos mais recentes, como ” Processamento de Dados Massivos (big data)”, "Data Streaming", "Cloud Computing e “Bancos de Dados Paralelos”, todos com 1 citação.

5 ANÁLISE DAS NECESSIDADES DA INDÚSTRIA BRASILEIRA EM RELAÇÃO À ENGENHARIA DE DADOS

Nesta seção, discutiremos os resultados de uma pesquisa realizada junto a líderes de empresas de tecnologia brasileiras, cujo objetivo foi identificar as necessidades específicas dessas empresas em relação aos engenheiros de dados. Além disso, faremos uma análise detalhada das respostas obtidas. Com base nessa análise, combinada com o conhecimento do primeiro autor deste artigo como instrutor de cursos de engenharia e análise de dados, bem como com as referências bibliográficas relevantes, desenvolvemos uma taxonomia de tópicos abrangente.

Para entender a realidade atual da indústria brasileira e suas necessidades quanto à engenharia de dados, fizemos um levantamento com dez pessoas-chave em cargos de liderança na área de tecnologia de várias empresas relevantes no Brasil. Com base na experiência dos próprios autores na indústria, partimos de uma lista de 14 habilidades que consideramos relevantes para engenheiros de dados recém-graduados, que estão listadas na Tabela 2.

Em cada uma dessas habilidades, demos 3 opções para definir o nível de conhecimento desejável em cada uma: Conhecimento Teórico, Conhecimento Prático Básico, Conhecimento Prático Avançado. Também deixamos uma pergunta discursiva para detalhar as escolhas feitas na primeira pergunta e uma terceira pergunta para listar habilidades que não estavam incluídas na lista da primeira, mas que os participantes do levantamento achavam relevantes serem mencionadas. Note-se que o objetivo da pesquisa não foi cobrir toda a indústria brasileira, mas sim, termos uma noção inicial do que é esperado pela indústria. Com as respostas desse levantamento, foi possível iniciar a construção de uma taxonomia de tópicos, que, como esperado, é um subconjunto da taxonomia de tópicos das ementas e tópicos recentes necessários para a realidade atual, citados nas respostas. O resultado do levantamento pode ser visto na Tabela 2.

Habilidade	Teórico	Prático Básico	Prático Avançado
SQL	20%	60%	20%
Modelagem de Dados Relacionais	30%	70%	0%
NoSQL Databases	70%	20%	10%
Criação e gerenciamento de ETLs	0%	80%	20%
Streaming Processing	50%	40%	10%
Batch Processing	10%	80%	10%
Distributed Computing	60%	30%	10%
Distributed File System	70%	20%	10%
Cloud Computing	20%	80%	0%
Arquitetura de Data Lake	60%	20%	20%
Arquitetura de Data Warehouse	50%	30%	20%
Arquitetura de Data Lakehouse	70%	10%	20%
Programação	0%	40%	60%
Automatização de Infraestrutura	50%	50%	0%

Tabela 2: Levantamento das habilidades esperadas dos engenheiros de dados

Os resultados do levantamento indicaram que o conhecimento em SQL é uma habilidade importante, com 60% dos entrevistados relatando ser necessário conhecimento prático básico e 20% relatando ser necessário conhecimento prático avançado. A modelagem de dados relacionais também foi considerada relevante, com 70% dos entrevistados relatando ser necessário conhecimento prático básico. Conhecimento em Bancos de Dados *NoSQL* foi considerado uma habilidade para o qual é necessário conhecimento teórico por 70% dos entrevistados. Conhecimento sobre Criação e Gerenciamento de ETLs[11] tem 80% das respostas indicando a necessidade de conhecimento prático básico, mostrando uma clara necessidade de que engenheiros de dados saibam usar ferramentas e desenvolver pipelines de dados. Batch processing e cloud computing tem 80% das respostas indicando a necessidade de conhecimento prático básico, mostrando que tratar dados em lote é a principal técnica esperada de um engenheiro de dados, em comparação ao processamento de dados via Streaming, e que o trabalho no dia a dia será feito em um ambiente de Cloud (pública ou privada) e que esse domínio é relevante. Este fato é corroborado quando vemos que para Automatização de Infraestrutura foi indicado como necessidade de conhecimento básico em 50% dos casos. Isso pode ser explicado por estar se tornando cada dia mais comum que os engenheiros cuidem de suas ferramentas nesses ambientes, logo eles devem ser responsáveis por configurar, disponibilizar e otimizar seu uso. A expectativa sobre o conhecimento de arquiteturas de Data Lake e Data Warehouse ficou equilibrada, sendo esperado mais conhecimento teórico e menos prático. O último ponto interessante que queremos destacar é a relevância de habilidade em programação no levantamento, tendo 40% de respostas em conhecimento prático básico e 60% em conhecimento prático avançado. Concluímos que tal relevância se dá, pois a construção de pipelines ETLs é feita por meio de programação, utilizando principalmente Python e Scala. Mesmo as ferramentas "no code" ou "low code"[12] precisam de customizações, as tarefas de processamento de dados se dão por meio de códigos mais elaborados que o SQL é capaz, principalmente com Apache Spark, e, por fim, também é exigido, desse engenheiro, que ele construa algoritmos, APIs e programas auxiliares para tarefas do dia a dia.

Além do levantamento acima, também foi usado o modelo de conceitos-chave para gerenciamento de dados criado por [8], que traz quatro categorias principais: Práticas, Princípios de Projeto, Mecanismos e Tecnologias Centrais, com termos genéricos e bastante abrangentes, porém, que definem o estado da engenharia de dados atual. A taxonomia também foi complementada pelos tópicos propostos em [Til De Bie 2022] onde são identificadas necessidades do dia a dia de engenharia de dados. Outra fonte consultada para criação da taxonomia foi a versão beta do currículo de computação da ACM, em 2023 [3]. Por fim, também foi usado o conhecimento do primeiro autor deste artigo na indústria como engenheiro de dados com mais de 12 anos de experiência em grandes empresas, e também sua experiência tendo fundado e ensinado em um bootcamp de engenharia e análise de dados, durante 6 anos, tendo a criação desse bootcamp sido motivada pela falta de profissionais capacitados na área de engenharia de dados no mercado, em 2017.

Com todas essas informações, criamos uma taxonomia de tópicos com 79 entradas, dividida em 3 níveis, do mais genérico ao mais específico. Essa taxonomia tem como principais diferenças da taxonomia de ementas o foco maior em NoSQL na parte teórica, com tópicos específicos para bancos de dados dessa categoria, como Tipos de Bancos, Projeto Físico, Gerenciamento de Bancos de Dados NoSQL, entre outros. Isso ocorre porque tais bancos de dados têm diferenças cruciais dos bancos de dados relacionais. Tais tópicos foram retirados do [3]. Também foram reorganizados os subtópicos de bancos de dados relacionais, agrupando Projeto Físico, Bancos de dados distribuídos, Otimização de consulta e Gerenciamento de Bancos de Dados abaixo deste tópico, pois queremos destacar que tais subtópicos dizem respeito somente ao universo de dados relacionais, e por isso, não podem ocupar o primeiro nível da taxonomia. No tópico de Mineração de Dados, foi adicionado os subtópicos de Processamento em Lote, Streaming, ETLs, Ingestão de Dados e Transformação, pois segundo [5], esse trabalho de minerar dados tem tais tarefas e tecnologias envolvidas. Em Big Data, foram adicionados os tópicos de Data Lake e Data Lakehouse como especificidade de Arquitetura, dado que a primeira é estabelecida na indústria como padrão e a segunda é um tópico emergente [16]. Também foram adicionados o subtópico de Bancos de dados NewSQL [14], Governança de Dados, Armazenamento e Computação distribuída e Evolução de Esquema, por serem assuntos vitais para o armazenamento e processamento de grandes massas de dados. Em Cloud Computing, foram adicionados os tópicos de Automatização e Infraestrutura, pois se revelou importante no levantamento feito junto à indústria, Containers, pois, hoje, é o principal meio para rodar aplicações em ambientes de cloud computing, Segurança, Custos, Monitoramento e Técnicas de SRE. Em Modelagem Conceitual, foram adicionados sub-tópicos de modelagem de dados para bancos de dados não relacionais, como bancos de dados de documentos e grafos. Na Figura 3, apresentamos uma parte dessa taxonomia com os tópicos somente até o segundo nível.

6 DISCUSSÃO

Nesta seção, com base nos levantamentos apresentados nas seções anteriores, apresentamos um panorama da situação atual do ensino de Engenharia de Dados nas universidades brasileiras e da cobertura dos tópicos ensinados frente às necessidades da indústria brasileira por profissionais formados nessas universidades.

Inicialmente, pode-se notar que a taxonomia das ementas (Seção 4) tem 138 tópicos, enquanto a da indústria (Seção 5) é um pouco menor, com 79 tópicos. Isso ocorre principalmente porque as ementas trazem muitos tópicos explicitamente, enquanto, na indústria, esses tópicos podem ter perdido seu destaque, e/ou terem sido agrupados em termos mais genéricos. Um exemplo disso é o tópico "Gerenciamento de Bancos de Dados", que, na primeira taxonomia, está no primeiro nível, com 5 subtópicos e, na segunda taxonomia, é um subtópico de "Bancos de Dados Relacionais"e "Bancos de Dados Não-relacionais". Isso se deu pelo fato de a pesquisa com a indústria indicar que não seria necessário um aprofundamento nesse

[11] Abordagem de processamento de dados, na qual os dados são extraídos de várias fontes, carregados em um ambiente de armazenamento e, em seguida, transformados e preparados para análise e uso posterior.
[12] Plataformas de desenvolvimento que permitem criar aplicativos ou sistemas sem a necessidade de escrever código complexo, utilizando interfaces gráficas e recursos visuais intuitivos.

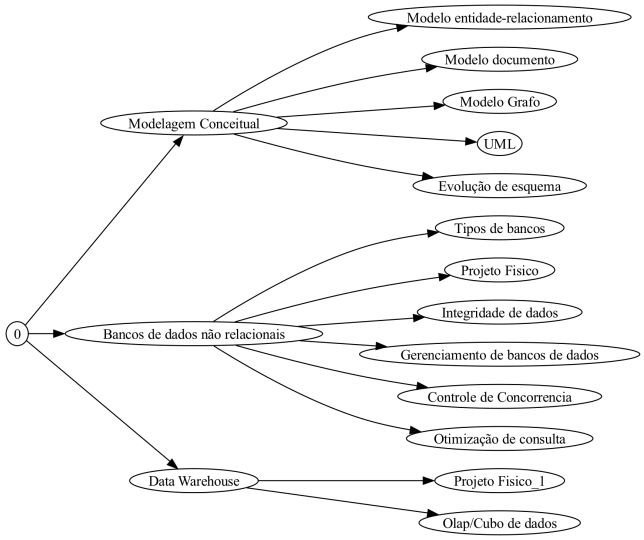

Figura 3: Amostras das taxonomias com tópicos da indústria

tópico. Também acontece com outros tópicos como "Controle de Concorrência", "Processamento e Otimização de Consultas", entre outros.

A importância do conhecimento teórico e prático em bancos de dados não-relacionais fica evidente na pesquisa feita junto à indústria, e é um tópico fundamental atualmente, pois tanto na academia quanto na indústria, cada vez mais se considera que somente bancos de dados relacionais não são mais suficientes para lidar com a quantidade e a variedade de dados com as quais é necessário lidar hoje em dia [7] [15] [14]. Por isso, expandimos o tópico listado no primeiro nível da taxonomia das ementas para detalhar e deixar evidente que, conforme o levantamento junto à indústria, esses tipos de bancos de dados devem ser estudados numa profundidade semelhante aos tradicionais bancos de dados relacionais.

O tópico "Modelagem Conceitual"está presente nas duas taxonomias e sua importância não desapareceu, no levantamento junto à indústria, esse tópico teve destaque para 70% dos respondentes como importante habilidade prática. Tal resposta se dá pela prevalência dos bancos de dados relacionais no mercado, e também pela sua importância na construção de um "Armazém de Dados".

O tópico "Mineração de Dados"passou a incluir os métodos de Processamento de dados em lote e via Streaming, que serão usados para Ingestão de dados, Transformação de dados, garantir Qualidade dos dados. Esses subtópicos representam o trabalho do Engenheiro de Dados no dia a dia e são essenciais para manipulação dos dados [5]. Na taxonomia das ementas, Mineração de Dados ainda trata de técnicas específicas para tratamento dos dados, como KDD, detecção de outlier e tratamento de dados faltantes (missing values). Esses subtópicos ainda fazem parte do tópico, mas agora estão incluídos nos termos citados acima.

O tópico "Big Data"inclui todos os tópicos da taxonomia de ementas, com a adição de tópicos relevantes para a indústria. O subtópico de Arquitetura recebe dois sub-subtópicos, data lake e data lakehouse, o primeiro, por ser a segunda geração de plataformas de dados; e o segundo, por ser a evolução dessa plataforma [16]. Foi adicionado também um subtópico de Armazenamento Distribuído, pois atualmente esse é o padrão da indústria para Big Data. Um novo tipo de bancos de dados foi incluído, os Bancos NewSQL, os quais são SGBDs para Big Data com características relacionais. Os subtópicos "Governança de Dados"e "Evolução de Esquema"foram incluídos, por serem essenciais em ambientes de "Big Data"para mantê-los coesos ao longo do tempo.

O tópico "Data Warehouse"se mantém relevante entre as taxonomias, dada a sua importância para as análises de negócio, sendo usado por praticamente todas as empresas Fortune 500 juntamente com Data Lake. No entanto, essa arquitetura está evoluindo para algo que está sendo chamado de "Data Lakehouse"[16]. Esse novo conceito foi incluído no tópico "Big Data", no terceiro nível, abaixo do tópico Arquitetura. Ainda em "Big Data", foi expandido tal tópico para tratar de armazenamento distribuído, tipos de dados estruturados e não-estruturados, governança de dados e bancos de dados NewSQL [14].

Durante a pesquisa, ficou evidente que a engenharia de dados tem muita aderência a ciência da computação, com tópicos que vão além de lidar com os dados em si, como Computação em Nuvem, para processar dados em larga escala, disponibilizar dados para consumo, unir ferramentas por meio de APIs [9], etc. Também foi detectada a necessidade do conhecimento em Computação Distribuída para entender seus conceitos e como podem ser aplicados no contexto de dados.

O tópico de Computação em Nuvem tem uma frequência baixa na taxonomia de ementas, porém, no levantamento junto à indústria, ele se mostrou muito relevante, por isso, foi mantido e expandido. A Automatização de Infraestrutura(infrastructure as code) é um caso a se notar, pois no levantamento com a indústria, ela se mostrou relevante e necessária para criar e manter serviços em plataformas de computação em nuvem e essa tarefa é cada vez mais comum no dia a dia de um engenheiro de dados.

Tópico	Academia	Indústria
Computação em Nuvem	Baixa	Alta
Data Streaming	Baixa	Alta
Modelagem Conceitual	Alta	Alta
Data Lake	Baixa	Alta
Data Warehouse	Média	Média

Tabela 3: Comparação de importância dos tópicos para Academia vs Indústria

Adicionamos um tópico chamado genericamente de "Programação"para deixar explícito o conhecimento necessário em microserviços, padrões de projeto e Estrutura de dados. Ficou muito evidente, na pesquisa, que esse tópico é extremamente relevante para a indústria, quando falamos de engenharia de dados, pois tais engenheiros precisam integrar diversas ferramentas heterogêneas, e, muitas vezes, criar soluções para consumo desses dados via APIs.

Podemos notar que a academia tem um grande foco em bancos de dados relacionais e suas tecnologias, o que pode ser explicado por sua importância, desde a década de 1970. Poucas universidades procuraram agregar tecnologias mais modernas em seu currículo, e

quando o fizeram, as colocaram com pouco destaque no currículo ou em cursos de pós-graduação, o que acabou gerando uma distância e um descompasso entre a academia e a indústria.

Essa distância entre academia e indústria impacta diretamente a colocação de recém-graduados no mercado de trabalho para trabalhar com dados, sendo necessário um investimento grande, tanto da indústria, para treinar tais profissionais, quanto em relação aos alunos para buscar cursos para aprenderem essas habilidades. Essa formação é geralmente dada por cursos rápidos chamados de "bootcamps", que fazem um bom papel de introdução dos assuntos, mas dificilmente aprofundam o conteúdo como a universidade tem o poder de fazer.

7 CONSIDERAÇÕES FINAIS

É fato que os dados se tornaram um elemento importantíssimo na indústria e na sociedade e, para gerenciá-los eficientemente, é preciso conhecimento e experiência em vários tópicos da Ciência da Computação (CC) [5]. Quando olhamos para esse problema, tendo em vista o que é ensinado nos cursos de CC, vemos que existe uma distância entre o gerenciamento de dados ensinado nos cursos de CC e o que é necessário na indústria. Enquanto o ensino em CC, no contexto de dados se concentra em tópicos tradicionais como bancos de dados relacionais e modelagem de dados relacionais, outros aspectos importantes atualmente quase não são considerados.

À medida que as competências e habilidades básicas de gerenciamento de dados estão se tornando cada vez mais necessárias, os alunos devem conseguir adquiri-las em sua formação em CC [8]. Examinando a taxonomia da indústria, fica evidente que existe uma grande correlação com o que chamamos de "Engenharia de Dados"(ED) e a graduação de CC e com algumas modificações, em algumas disciplinas conseguiríamos atender às necessidades latentes do mercado, de forma mais rápida e eficiente, evitando toda a burocracia para criar um novo curso.

Como trabalhos futuros, pretendemos, a partir dos resultados aqui apresentados, propor mudanças no currículo de graduação de uma universidade, modificando a disciplina de *Tópicos Especiais em Bancos de Dados* de dados para incluir os Tópicos de *Processamento em Lote e Streaming*, *Cloud Computing*, *ETLs*, entre outros, para melhor atender aos requisitos da indústria quanto ao ensino de ED e analisar os efeitos dessas mudanças na formação dos alunos e de sua aceitação no mercado.

REFERÊNCIAS

[1] ACM. 2021. Data Science Curricula 2021. https://www.acm.org/binaries/content/assets/education/curricula-recommendations/dstf_ccdsc2021.pdf

[2] ACM and IEEE. 2020. ACM Computing Curricula 2020. https://www.acm.org/binaries/content/assets/education/curricula-recommendations/cc2020.pdf

[3] ACM, IEEE and AAAI. 2023. Computer Science Curricula 2023 - Version Beta. https://csed.acm.org/wp-content/uploads/2023/03/Version-Beta-v2.pdf

[4] Imanol Arrieta-Ibarra et al. 2018. Should We Treat Data as Labor? Moving beyond "Free". *AEA Papers and Proceedings* 108, 38–42.

[5] Tijl De Bie et al. 2022. Automating data science. *Commun. ACM* 65, 3, 76–87.

[6] Peter J. Denning. 2003. Great principles of computing. *Commun. ACM* 46, 11, 15–20.

[7] Andreas Grillenberger and Ralf Romeike. 2014. Big Data - Challenges for Computer Science Education. In *Informatics in Schools. Teaching and Learning Perspectives - 7th International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, ISSEP*. 29–40.

[8] Andreas Grillenberger and Ralf Romeike. 2017. Key Concepts of Data Management – an Empirical Approach. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research*. 30–39.

[9] Ismail Bile Hassan and Jigang Liu. 2019. Embedding Data Science into Computer Science Education. In *IEEE International Conference on Electro Information Technology EIT*. 367–372.

[10] Nicolaus Henke et al. 2016. The age of analytics: Competing in a data-driven world.

[11] Alfredo Nazábal et al. 2020. Data Engineering for Data Analytics: A Classification of the Issues, and Case Studies. *CoRR* abs/2004.12929.

[12] David Reinsel et al. 2018. The Digitization of the World - From Edge to Core.

[13] SBC. 2021. Ref. Curricular: Bacharelado em Ciência de Dados. https://www.sbc.org.br/documentos-da-sbc/send/131-curriculos-de-referencia/1402-ref-curricular-bacharelado-em-ciencia-de-dados

[14] Yasin N. Silva et al. 2014. Integrating big data into the computing curricula. In *The 45th ACM Technical Symposium on Computer Science Education, SIGCSE*. 139–144.

[15] Michael Stonebraker and Ugur Çetintemel. 2005. "One Size Fits All": An Idea Whose Time Has Come and Gone (Abstract). In *Proceedings of the 21st International Conference on Data Engineering, ICDE*. 2–11.

[16] Matei Zaharia et al. 2021. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In *11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, Online Proceedings*.