Ensino de Engenharia de Dados nas Universidades Brasileiras: Estado Atual e Perspectivas de Mercado*

Tarsis Azevedo, Altigran da Silva

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
69.077-000 – Manaus – AM – Brasil

{tarsis.azevedo,alti}@icomp.ufam.edu.br

Resumo. *O termo* Engenharia de Dados (ED) *tem sido utilizado para se referir aos processos de adquirir, organizar e preparar dados para serem consumidos em análises ou aplicações. 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 sua crescente relevância e a necessidade de habilidades relacionadas a ela nos profissionais de computação. Nossa motivação foi percepção de que os avanços de pelo menos uma década na indústria em ED ainda não foram adequadamente absorvidos pelo ensino nas universidades. Para esse trabalho, construímos e comparamos duas taxonomias de tópicos, que resultaram, respectivamente, de levantamentos das disciplinas, bibliografias e ementas relacionadas a ED em universidades brasileiras e junto a empresas de tecnologia do país. Identifcamos uma lacuna entre o ensino e mercado, com currículos desatualizados quanto a tópicos considerados relevantes para a indústria contemporânea. Em particular, tópicos sobre 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. Nosso objetivo é subsidiar mudanças nos currículos que possam contribuir para a formação de profissionais mais qualificados e alinhados às necessidades modernas do mercado.*

1. Introdução

A digitalização acelerada de praticamente todas as atividades da sociedade moderna teve como uma de suas maiores consequências a produção de dados em um volume sem precedentes na história. É tamanha a prevalência dos dados atualmente que eles são comparados em importância a *commodities* como o petróleo, ou mesmo a bens de capital, como o trabalho [Arrieta-Ibarra et al. 2018]. De fato, a quantidade de dados gerados globalmente aumenta diariamente, com estimativa de crescimento de mais de 5 vezes entre 2018 e 2025, sendo que pelo menos 30% deste volume de ser produzido pela comunicação maquina-maquina [Reinsel et al. 2018]. Assim, um grande desafio contemporâneo é a disponibilidade de mão-de-obra para lidar com este volume de dados e ser capaz de desenvolver soluções para processá-los de forma efetiva e escalável. É bem conhecido que 80% do trabalho envolvido neste tipo de atividade é associado a tarefas como integrar, preparar, transformar e manipular os dados a serem utilizados. Neste artigo, empregamos o termo de *Engenharia de Dados (ED)* para nos referirmos a estas atividades [Bie et al. 2022].

*Este trabalho foi parcialmente apoiado pela Jusbrasil, pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001, pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD, e pelo CNPq através de uma bolsa PQ para Altigran da Silva (Proc. 307248/2019-4).

Assim, nosso ponto de vista é que para atender adequadamente as 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.

Apresentamos neste artigo um estudo que busca responder duas questões relacionadas. (1) *que assuntos têm sido ensinados em ED nas Universidades brasileiras?*. (2) *como estes assuntos abrangem as necessidades da sociedade e do mercado brasileiro em ED*. O nosso estudo se concentra nos cursos de graduação em computação das universidades brasileiras, onde tipicamente os tópicos de ED são cobertos em disciplinas como *Bancos de Dados* ou uma de suas variações. Não cobrimos aqui nenhum curso ou mesmo disciplinas no escopo de *Ciência de Dados*. Isso se deve a várias razões, entre elas o fato de que essa área cobre um corpo de conhecimento que intersecta a da computação, mas que é distinto deste. Além disso, existem diversos estudos em propostas para criação de cursos de graduação específicos para esta área, cobrindo esse corpo de conhecimento[1][2]. Finalmente, entendemos que as atividades de ED são inerentes à área de computação, uma vez que sua aplicação finalística não se limita a de Ciência de Dados, mas inclui também sistemas como busca, recomendação, enriquecimento de documentos, processamento de imagens e texto, entre outras aplicações.

Para determinar que assuntos sobre ED tem sido ensinados no país, fizemos um recorte com 23 das universidades mais tradicionais em ED, e analisamos as disciplinas da área, suas bibliografias e ementas, construindo uma taxonomia de tópicos dessas disciplinas. Para determinar como esses assuntos abrangem as necessidades do mercado, fizemos um levantamento com várias das principais empresas de tecnologia no país para levantar quais habilidades seriam necessárias para um Engenheiro de Dados recém-graduado. De posse dessas informações, construímos outra taxonomia de tópicos. Finalmente, fizemos uma comparação das duas taxonomias para entender as diferenças entre os assuntos já cobertos pela Universidade e quais faltam para a indústria.

O *Computing Curricula 2023* da ACM destaca que, desde os anos 70, o foco de estudos em gerenciamento de dados tem sido em BDs relacionais, mas o aumento exponencial do volume de dados demanda uma atualização nesse enfoque. Ele também salienta a importância do envolvimento da indústria na formulação de currículos, para preparar melhor os profissionais para o mercado. Usamos esse currículo para complementar nossa taxonomia de tópicos para a indústria. A ACM e SBC tem currículos para criação de cursos de Ciência de Dados, porém em [Hassan and Liu 2019] os autores concluíram 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. Embora existam os estudos acima nesse sentido, nenhum foi feito para a realidade do Brasil, então, nossa pesquisa é relevante por fornecer uma visão das lacunas na formação dos profissionais de dados no país.

2. Metodologia

A Engenharia de Dados (ED) aborda problemas relacionados à organização e qualidade dos dados, bem como à extração de características de entidades do mundo real representadas em meio digital [Nazábal et al. 2020]. Com essa definição, selecionamos disciplinas que se relacionam diretamente com esses problemas, tais como *Bancos de Dados*, *Mineração de Dados*, *Tópicos Especiais em Bancos de Dados*, *Computação em Nuvem*, entre outras. Disciplinas que não abordam diretamente esses problemas, como

[1]ACM (2021). Data science curricula 2021.

[2]SBC (2021). Ref. curricular: Bacharelado em ciência de dados.

Inteligência Artificial e *Ciência de Dados*, foram excluídas de nossa lista. A lista final das disciplinas está em um *hot site* anonimizado que criamos para apoio ao trabalho[3].

Para determinar os assuntos sobre ED ensinados nas universidades brasileiras fizemos um recorte de 23 destas que mantêm grupos de pesquisa tradicionais na área e fizemos um levantamento com base nos currículos de graduação em computação disponibilizados no site de cada instituição. Levantamos 108 disciplinas e listamos suas ementas e suas bibliografias, que podem ser vistas no nosso hot site[4]. Filtramos as disciplinas que consideramos não-relacionadas, com base na nossa definição de ED, ficando com 90 disciplinas.

A fim de construir uma taxonomia dos tópicos das disciplinas de ED, 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[5], que serviu como base para a criação da taxonomia de tópicos. Ressalta-se que a intenção deste trabalho foi criar uma visão geral do que é ensinado nas universidades brasileiras na área.

Para construir uma taxonomia de tópicos ensinados nas disciplinas, tomamos os assuntos tratados em cada ementa e mapeamos para a taxonomia de tópicos da bibliografia descrita acima. Isso foi feito porque entendemos que os assuntos tratados nas ementas são um sub-conjunto de todos os assuntos tratados na bibliografia de cada disciplina. Realizamos também uma análise de frequência nas ementas para compreender como os tópicos listados foram utilizados pelas disciplinas, esta analise pode ser vista no nosso hotsite[6]. Para isso, procuramos os tópicos listados na taxonomia em cada ementa, e para cada ocorrência, somamos 1 a frequência. Alguns termos foram agrupados sob um mesmo tópico, como os tópicos *Normalização*, *Formas Normais*, *1FN*, *2FN*,*3FN*, que estão todos sob o tópico *Formas Normais*.

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 referentes a habilidades técnicas necessárias para Engenharia de Dados e o nível de proficiência esperado em cada uma dessas habilidades (Teórico, Prático Básico e Prático Avançado). 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 escolhidas por nós em posições de liderança dentro dessas empresas. O seu resultado bruto pode ser visto no nosso *hot site*[7].

A partir deste levantamento, geramos uma taxonomia de tópicos. Para completar essa taxonomia, utilizamos outras duas fontes. A primeira fonte foi o conjunto de conceitos-chave definidos por [Grillenberger and Romeike 2017], 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

[3] https://bit.ly/sbbd-23-disciplinas
[4] https://bit.ly/sbbd-23-biblio-ement
[5] https://bit.ly/sbbd-23-cap-filtr
[6] https://bit.ly/sbbd-23-tax-ement-freq
[7] https://bit.ly/sbbd-23-resultado-pesquisa

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 apresenta um recorte dos assuntos considerados relevantes nesta área no mercado do Brasil.

Na etapa seguinte, realizamos uma análise comparativa entre a taxonomia de ementas e a taxonomia de tópicos da indústria. 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.

3. Levantamento das Bibliografias e Ementas

Nesta seção, apresentamos os resultados do levantamento feito sobre a bibliografia e as ementas das disciplinas de Engenharia de Dados (ED), com a metodologia descrita na Seção 2.

A taxonomia completa da bibliografia pode ser encontrada em nosso hot site[8]. A taxonomia de ementas e a frequência dos tópicos completa com as frequências dos tópicos pode ser vista no nosso hot site[9].

Com base na frequência de citações dos tópicos nas ementas, pudemos observar que, como esperado, os tópicos relacionados a bancos de dados relacionais são amplamente abordados. Tópicos como *SQL*, *Modelo Relacional*, *Formas Normais*, e outros assuntos mais tradicionais 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. Uma observação que foi possível fazer logo depois deste levantamento, foi a baixa frequência de citações para tópicos mais recentes, como *Big Data*, *Workflow de Dados*, *Computação em Nuvem* e *Bancos de Dados Paralelos*, todos com uma citação.

4. Análise das Necessidades da Indústria Brasileira

Nesta seção, apresentamos o levantamento com líderes de empresas de tecnologia brasileiras, como descrito na Seção 2. O objetivo do levantamento não foi cobrir toda a indústria brasileira, mas sim dar uma noção inicial do que é esperado por ela. O resultado do levantamento pode ser visto na tabela que está no nosso hot site[10].

Com todas essas informações, criamos uma taxonomia de tópicos, seguindo a metodologia descrita na Seção 2, com 79 entradas, dividida em 3 níveis, do mais genérico ao mais específico. A taxonomia completa pode ser encontrada em nosso hot site[11]. Ela da destaque a sistemas NoSQL na parte teórica, com tópicos específicos para bancos de dados dessa categoria, devido a sua relevância no levantamento junto a industria, como *Tipos de Bancos*, *Projeto Físico*, *Gerenciamento de Bancos de Dados NoSQL*, entre outros. Isso ocorre porque tais sistemas têm diferenças cruciais dos SGBDs relacionais, pois tanto na academia quanto na indústria cada vez mais se considera que somente bancos de dados relacionais não são mais suficientes a quantidade e a variedade de dados com as quais é necessário lidar hoje em dia [Grillenberger and Romeike 2014,

[8]https://bit.ly/sbbd-23-tax-biblio-comp
[9]https://bit.ly/sbbd-23-tax-ement-comp
[10]https://bit.ly/tabela-resultado
[11]https://bit.ly/sbbd-23-tax-ind-comp

Stonebraker and Çetintemel 2005, Silva et al. 2014]. Tais tópicos foram obtidos do relatório da ACM[12].

Também foram reorganizados os sub-tópicos de bancos de dados relacionais, agrupando *Projeto Físico*, *Bancos de Dados Distribuídos*, *Otimização de consulta* e *Gerenciamento de Bancos de Dados* embaixo deste tópico, pois queremos destacar que tais sub-tópicos dizem respeito somente ao universo do modelo relacional. No tópico de *Mineração de Dados*, foram adicionados os sub-tópicos de *Processamento em Lote*, *Streaming*, *ETLs*, *Ingestão de Dados* e *Transformação*, de acordo com [Bie et al. 2022]. 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 [Zaharia et al. 2021].

Também foram adicionados o sub-tópico de *Governança de Dados*, *Bancos de Dados NewSQL* [Silva et al. 2014], *Armazenamento e Computação distribuída* e *Evolução de Esquema*, por serem assuntos vitais para o armazenamento e processamento de grandes massas de dados. Em *Computação em Nuvem*, foram adicionados os tópicos de *Automatização de Infraestrutura*, pois se revelou importante no levantamento feito junto a indústria, *Containers*, pois hoje é o principal meio para rodar aplicações em ambientes de cloud computing, entre outros topicos relacionados. 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.

Adicionamos um tópico chamado genericamente de *Programação* para deixar explícito o conhecimento necessário em micro-serviços, padrões de projeto e *Estruturas 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, desenvolver soluções próprias para consumo desses dados via APIs.

5. O Ensino de Engenharia de Dados e a Indústria

Nesta seção, com base nos resultados das 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.

Inicialmente, pode-se notar que a taxonomia das ementas (Seção 3) tem 138 tópicos, enquanto a da indústria (Seção 4) é um pouco menor, com 79 tópicos. Isso ocorre principalmente porque as ementas trazem muitos tópicos explicitamente, enquanto na da indústria esses tópicos podem ter perdido seu destaque, e/ou ter sido agrupados em termos mais genéricos. Um exemplo é o tópico *Gerenciamento de Bancos de Dados*, que na primeira taxonomia está no primeiro nível, com 5 sub-tópicos e na segunda taxonomia é um sub-tópico de *Bancos de Dados Relacionais*.

A importância do conhecimento teórico e prático em bancos de dados não-relacionais fica evidente na pesquisa feita junto a indústria, e é um tópico fundamental atualmente, por isso, o colocamos no primeiro nível da taxonomia da indústria para detalhar e deixar evidente que, conforme o levantamento junto à indústria, estes tipos bancos de dados devem ser estudados numa profundidade semelhante aos tradicionais bancos de dados relacionais.

[12]ACM, IEEE and AAAI (2023). Computer science curricula 2023 - version beta.

No que diz respeito às arquiteturas de armazenamento e consumo de dados, atualmente observamos três gerações distintas no mercado. A primeira geração é representada pelo conceito de *Data Warehouse*, que ainda se mantém como um tópico de primeiro nível na taxonomia de arquiteturas de dados. A segunda geração é caracterizada pelo *Data Lake*, que se encontra categorizado como um sub-tópico de *Arquitetura* na área de *Big Data*. Essas duas gerações compõem a maior parte das arquiteturas de dados empregadas pelas empresas listadas na Fortune 500 [Zaharia et al. 2021]. A inovação mais recente, denominada *Data Lakehouse*, é considerada a terceira geração. Esta se posiciona também como um sub-tópico de *Arquitetura* em *Big Data* [Zaharia et al. 2021].

Ainda em *Big Data* foi adicionado também um sub-tópico de *Armazenamento Distribuído*, pois atualmente esse é o padrão da indústria para Big Data. Um novo tipo de bancos de dados foi incluído, os chamados *NewSQL*, que correspondem a SGBDs para Big Data com características relacionais. Os sub-tópicos *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, na experiência do primeiro autor do artigo.

Ainda em *Big Data*, temos os sub-tópicos de *MapReduce* e *Apache Spark*, que são frameworks de processamento de dados distribuídos. Esses frameworks são citados no tópico de *Mineração de Dados*, que passou a incluir os métodos de *Processamento de Dados em Lote* e *Processamento em Streaming*. Os sub-tópicos de *Ingestão de Dados*, *Transformação de Dados* e *Qualidade dos dados* representam o trabalho do Engenheiro de Dados no dia a dia e são essenciais. [Bie et al. 2022]. Também foi incluído o sub-tópico de *ETLs*, que representa a construção de pipelines de processamento de dados.

O tópico de *Computação em Nuvem* tem uma frequência baixa na taxonomia de ementas, porém, no levantamento junto a indústria ele se mostrou muito relevante, por isso foi mantido e expandido. A *Automatização de Infraestrutura* 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.

6. Conclusão

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) [Bie et al. 2022]. 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 [Grillenberger and Romeike 2017]. 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 as necessidades latentes do mercado.

Como trabalhos futuros, pretendemos a partir dos resultados aqui apresentados propor mudanças nos currículos de graduação de uma universidade para melhor atender os requisitos da industria quanto ao ensino de ED e analisar os efeitos destas mudanças na formação dos alunos e sua aceitação no mercado.

Referências

Arrieta-Ibarra, I. et al. (2018). Should we treat data as labor? moving beyond "free". *AEA Papers and Proceedings*, 108:38–42.

Bie, T. D. et al. (2022). Automating data science. *Commun. ACM*, 65(3):76–87.

Grillenberger, A. and Romeike, R. (2014). Big data - challenges for computer science education. In *Proceedingso of the 7th Int. Conf. on Informatics in Schools: Situation, Evolution, and Perspectives, ISSEP*, pages 29–40.

Grillenberger, A. and Romeike, R. (2017). Key concepts of data management – an empirical approach. In *Proceedings of the 17th Koli Calling Int. Conf. on Computing Education Research*, page 30–39.

Hassan, I. B. and Liu, J. (2019). Embedding data science into computer science education. In *IEEE Int. Conf. on Electro Information Technology EIT*, pages 367–372.

Nazábal, A. et al. (2020). Data engineering for data analytics: A classification of the issues, and case studies. *CoRR*, abs/2004.12929.

Reinsel, D. et al. (2018). The digitization of the world - from edge to core.

Silva, Y. N. et al. (2014). Integrating big data into the computing curricula. In *The 45th ACM Technical Symposium on Computer Science Education, SIGCSE*, pages 139–144.

Stonebraker, M. and Çetintemel, U. (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*, pages 2–11.

Zaharia, M. 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*.