

IMPLEMENTAÇÃO DE UM DATA WAREHOUSE PARA ANÁLISE DE DADOS ABERTOS GOVERNAMENTAIS DA EDUCAÇÃO A DISTÂNCIA

Roberta Macêdo Marques Gouveia*

Charles Nicollas Cavalcante Freitas**

Resumo: O artigo propõe a implementação de um banco de dados educacional multidimensional voltado à análise do ensino superior brasileiro, com foco na Educação a Distância – EaD. Para dar suporte ao processo de apoio à decisão foi projetado um *Data Warehouse* – DW, tendo como base a modelagem dimensional constelação de fatos e a aplicação de tecnologias *OnLine Analytical Processing* – OLAP. Foram utilizados dados abertos de alta granularidade dos sistemas e-MEC e Universidade Aberta do Brasil. O trabalho fundamenta-se no desenvolvimento de um ambiente computacional analítico visando traçar o perfil de instituições de ensino atuantes na EaD, bem como estudantes universitários e cursos, em especial da área de conhecimento “ciências exatas e da terra” e subárea “computação”. Após as análises realizadas no trabalho por meio do DW, constata-se que a modalidade de educação a distância está cada vez mais acessível no Brasil, com vistas à interiorização e democratização do ensino superior.

Palavras-chave: Educação a Distância. Dados Abertos. *Data Warehouse*. Modelagem Dimensional.

1 Introdução

Este artigo aborda o desenvolvimento de um banco de dados educacional multidimensional voltado para análise do ensino superior brasileiro. O objetivo do trabalho consiste no desenvolvimento de um ambiente computacional analítico, a fim de compreender e traçar o perfil da Educação a Distância (EaD) no Brasil. Para isso foram utilizadas as tecnologias de *Data Warehouse* (DW) e *OnLine Analytical Processing* (OLAP), com foco em apoiar decisões estratégicas no âmbito da EaD

Um DW é um tipo especial de banco de dados que facilita a exploração de grandes volumes de dados, proporcionando agregações e sumarizações de forma otimizada. De acordo com Inmon (2005), o termo é definido como “um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar decisões gerenciais”. As tecnologias

* Professora Doutora do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco – UFRPE, Campus Recife/PE. Formação acadêmica em Ciência da Computação. E-mail: roberta.gouveia@ufrpe.br.

** Mestrando do Programa de Pós-Graduação em Informática Aplicada (PPGIA) da Universidade Federal Rural de Pernambuco – UFRPE. E-mail: nicollasfreitas21@gmail.com.

OLAP, por sua vez, são projetadas para apoiar consultas não triviais, além de auxiliar a sintetizar informações através de comparações, visões personalizadas e dados históricos. As tecnologias OLAP têm como característica principal permitir visões intuitivas dos dados multidimensionais, por meio de análises em diferentes perspectivas. OLAP também auxilia na exploração dos dados armazenados no DW, fornecendo funcionalidades para análise interativa de dados em diferentes dimensões e granularidades (INMON, 2005).

Para construção do DW foi utilizada a modelagem dimensional constelação de fatos, que é uma técnica de projeto lógico definida sobre dois pilares: tabelas fato e tabelas dimensão. De acordo com Kimball e Ross (2011), fato é uma tabela primária em um modelo dimensional, onde as medidas numéricas de desempenho são armazenadas, já as tabelas de dimensão são complementos integrais para uma tabela fato, contendo descritores textuais do domínio analisado.

O DW foi implementado com dados abertos de alta granularidade obtidos dos sistemas e-MEC e Universidade Aberta do Brasil (UAB), visando obter uma visão contextualizada e realista da EaD no Brasil. Conforme constatado nesse trabalho, a aplicação de tecnologias OLAP em bases de dados educacionais possui grande potencial de diagnóstico, sendo úteis na exploração, análise e recuperação de informações, tendo um papel relevante no entendimento do perfil da EaD nas regiões geográficas do Brasil.

Com este estudo, buscou-se compreender os desafios da EaD, com vistas à obtenção de indicadores relacionados à qualidade das instituições e cursos a distância. Dentre as informações investigadas por meio do DW e das tecnologias OLAP tem-se: verificação das preferências de cursos e áreas de conhecimento na graduação e especialização; perfil da EaD no Brasil; investigação dos índices de avaliação dos cursos (Conceito de Curso – CC, Conceito Preliminar de Curso – CPC, Índice Geral de Cursos – IGC), índices de avaliação dos estudantes (notas do Exame Nacional de Desempenho dos Estudantes – ENADE) e índices de avaliação das Instituições Federais de Ensino Superior (IFES), por meio do Conceito Institucional (CI); exploração das regiões com oferta de cursos a distância; e análises dos índices de qualidade dos cursos à distância, por regiões e áreas de conhecimentos.

A EaD no Brasil está em constante expansão, devido ao aumento considerável de instituições de ensino superior adotando a modalidade a distância em seus cursos de graduação e especialização. Segundo ABED (2016), das 368 instituições analisadas pelo Censo da EaD, 79 (21,5%) também atuam fora de seu estado de origem, dispondo de dezenas de polos. O Censo da EaD contabilizou aproximadamente 5 (cinco) milhões de estudantes, sendo 1,1 milhão em cursos regulamentados totalmente a distância e semipresenciais (*b-*



learning), e aproximadamente 4 (quatro) milhões em cursos livres. Foi constatado no Censo da EaD 2016 um aumento de 23,4% de estudantes matriculados em cursos a distância no Brasil, isto é 1,2 milhão de estudantes a mais do que no Censo da EaD 2015.

Ainda de acordo com ABED (2016), no Brasil os cursos são ofertados em todos os níveis e áreas de conhecimento e as instituições contam, em média, com 1.000 a 5.000 estudantes, com destaque para as áreas de Ciências Sociais Aplicadas, com 608 ofertas de cursos regulamentados totalmente a distância. Entre os cursos semipresenciais a preferência é pela área de Ciências Humanas. Dentre as instituições educacionais com ofertas de cursos na modalidade a distância, 64% delas estão sediadas nas capitais dos estados brasileiros e 36% nos interiores, com concentração de 42% na região Sudeste, seguida pela região Sul (21%), Nordeste (18%), Centro-Oeste (13%) e Norte (6%).

A maioria das IFES que oferta cursos na modalidade presencial, também já oferta cursos a distância, ou está começando a ofertá-los (ABED, 2016). Embora tenha aumentado o número de instituições com adesão à EaD, o Brasil ainda está em fase de transição nessa modalidade, visto que muitas instituições estão se limitando a reproduzir para o ambiente virtual pequenas adaptações do ensino presencial. Muitas vezes as aulas são disponibilizadas do ensino presencial para o virtual sem qualquer alteração didático-pedagógica nos processos de ensino-aprendizagem. Essas práticas, contribuem para aumentar os índices de evasão e retenção dos estudantes. Em se tratando das taxas de evasão reportadas nos cursos a distância, ABED (2016) registra uma evasão de 26% a 50% – alertando que a desistência dos estudantes da EaD é maior que nos cursos presenciais. As instituições apontam o fator tempo como o mais influente no fenômeno da evasão, seguido do fator financeiro.

2 Trabalhos relacionados

Nos últimos anos várias instituições e empresas de diferentes segmentos – tais como educação, telecomunicações, transporte, saneamento, segurança, saúde, entre outros – vêm implantando processos computacionais de *Business Intelligence* (BI), com vistas ao planejamento estratégico, por meio de tecnologias de DW, OLAP, MD e *Machine Learning*. Tais organizações buscam otimizar o armazenamento e integração de seus dados, além de automatizar a extração de informações, reconhecimento de padrões e descoberta de novos conhecimentos. Seguem abaixo alguns trabalhos relacionados a DW com aplicação em diferentes áreas, inclusive na análise e gestão de instituições de ensino que é a base do presente trabalho.



O artigo de José Júnior, Bastos e Kaestner (2015) apresenta o desenvolvimento de um DW para instituições de educação superior que utilizam sistemas de gestão acadêmica e cujos dados gerados por esses sistemas transacionais não têm sido explorados de forma eficaz. Dessa forma, o intuito do DW é subsidiar a tomada de decisão pelos gestores educacionais. A abordagem é genérica e pode ser aplicada a uma grande quantidade de instituições. O trabalho relata um estudo de caso aplicado na Universidade Tecnológica Federal do Paraná, cuja construção do DW baseia-se no uso de visões materializadas e no uso de *table functions* para agregação das regras de negócio. Como resultados, os autores afirmam que a construção do DW baseado em visões materializadas propicia flexibilidade e facilidade de manutenção, visto que se alguma estrutura no banco de dados for alterada é de fácil verificação, além disso o uso das *table functions* permite uma independência da camada de apresentação.

Os serviços de abastecimento de água quando fornecidos de forma adequada, constituem-se em um dos principais indicadores de qualidade de vida da população. Diante desse contexto, a dissertação de mestrado de Gouveia (2009) relata o desenvolvemos de um sistema de apoio à decisão para gestores do setor de saneamento. A metodologia consiste em organizar os dados operacionais em um DW, aplicar as tecnologias OLAP e os algoritmos de mineração de dados, a fim de proporcionar aos gestores um melhor entendimento da companhia e perfil dos consumidores. Portanto, o trabalho tem como objetivos projetar um DW departamental referente ao setor comercial, aplicar as tecnologias OLAP sobre os cubos de dados multidimensionais, e executar algoritmos de mineração de dados de classificação e associação com vistas à melhor compreender e minimizar as perdas de água, inadimplências e ligações clandestinas no sistema de abastecimento urbano de água.

O trabalho desenvolvido por Kanashiro (2007) insere-se no contexto de uma ferramenta inteligente de apoio à pesquisa, nomeada de FIP. A ferramenta proposta tem como propósito a recuperação, organização e mineração de grandes conjuntos de documentos científicos da área de computação. Nesse contexto, foi projetado um DW de artigos para a ferramenta FIP e, adicionalmente, realizado experimentos com técnicas de mineração e aprendizado de máquina para automatizar o processo de indexação das informações, descoberta de tópicos e documentos armazenados no DW.

Prado (2006) desenvolveu um sistema de informações gerenciais utilizando tecnologia de DW e ferramenta OLAP para análise de dados. Como resultado da dissertação foi implantado um *Data Mart* com o objetivo de proporcionar a geração de consultas *ad hoc*. Esse sistema foi aplicado em um estudo de caso envolvendo uma empresa de transporte



rodoviário, e apresentou boa capacidade na constituição do diagnóstico dos procedimentos da operação e apoio adequado ao processo decisório.

Pêgas e Yano (2004) afirmam que a tendência no processo de aprendizagem é o professor exigir do aluno a capacidade de exploração e criação de novos conhecimentos e não, simplesmente, memorização do conteúdo de uma disciplina. Entretanto, essa nova abordagem requer novos instrumentos para acompanhamento e avaliação do processo de aprendizagem. Uma maneira para entender o processo de aprendizagem de programação é através de dados coletados durante o desenvolvimento do projeto. Dessa forma, os autores propõem a implementação de um DW, obtendo respostas a questões que podem direcionar os professores a redefinirem estratégias para melhorar o desempenho de cursos de programação.

Na literatura existem vários trabalhos relacionados à temática de análise de dados no contexto educacional, dentre eles tem-se o artigo de Guércio et al. (2014) e Gottardo et al. (2014), que enumera técnicas de MD propondo reflexões fundamentadas sobre os desafios da informática na educação, com objetivo de potencializar o uso de tecnologias de informação e comunicação no processo de ensino e aprendizagem. O presente estudo pode ser considerado uma extensão do artigo publicado por Freitas et al. (2015), onde os autores aplicaram o processo de integração e análise de dados educacionais, visando análises bem fundamentadas acerca de instituições de ensino superior e cursos realizados em ambientes *b-learning*, bem como traçar o perfil e revelar os principais desafios da EaD.

3 Metodologia

O ambiente computacional analítico implementado nesse trabalho é formado por três componentes: (I) Cubos de dados do DW; (II) SGBD *MySQL Workbench* – considerados a memória do sistema de apoio a decisão; (III) ferramentas de BI da *suíte Pentaho* – consideradas a inteligência do ambiente proposto. A implementação do sistema consiste em integrar os dados da EaD oriundos de diversas fontes, com vistas a facilitar as análises de grandes volumes de dados. Tal sistema permite que se observem tendências; que se detectem cursos e áreas de conhecimento por região preferenciais dos estudantes; além de cursos e IFES que possuem os melhores indicadores de qualidade.

A primeira etapa para análise dos dados consistiu na criação do Modelo Dimensional, cujo propósito é facilitar a exploração dos dados, por isso a forma como os dados são organizados difere do modelo tradicional – Modelo Relacional. O formato do Modelo Dimensional é composto de tabelas fato e tabelas dimensão, com relacionamentos simples entre elas, para melhor manuseio e criação dos cubos de dados. Para a modelagem

dimensional foi utilizado o *Fact Constellation Schema* – Esquema Constelação de Fatos, que favorece a otimização de consultas, fazendo com que os comandos da linguagem MDX tenham o mínimo de junções. Nesse trabalho fez-se necessário a modelagem dimensional de 2 esquemas do DW, um representando o sistema UAB e outro o e-MEC.

O método utilizado para análise e descoberta de conhecimento iniciou-se com o processo *Extract, Transform and Load* (ETL) e prosseguiu com a implementação do DW. O pré-processamento, transformação, seleção e integração de dados foi realizado visando a limpeza e enriquecimento da base de dados. A implementação do DW se deu com a geração das tabelas fato e dimensão no *SGBD MySQL*, em seguida foi realizada a carga do DW com os dados abertos disponibilizados por dois repositórios governamentais relacionados ao Ministério da Educação: sistemas UAB e e-MEC.

O trabalho prossegue com a aplicação de tecnologias OLAP, a fim de realizar uma análise criteriosa e precisa dos dados armazenados no DW. As tecnologias OLAP foram utilizadas com foco nas diversas operações que podem ser realizadas sobre os cubos de dados, tais como funções de agregações, estatísticas e visualizações gráficas. As operações OLAP permitem materializar diferentes perspectivas, também conhecidas como visões, além de simulações e análises interativas sobre as dimensões do DW.

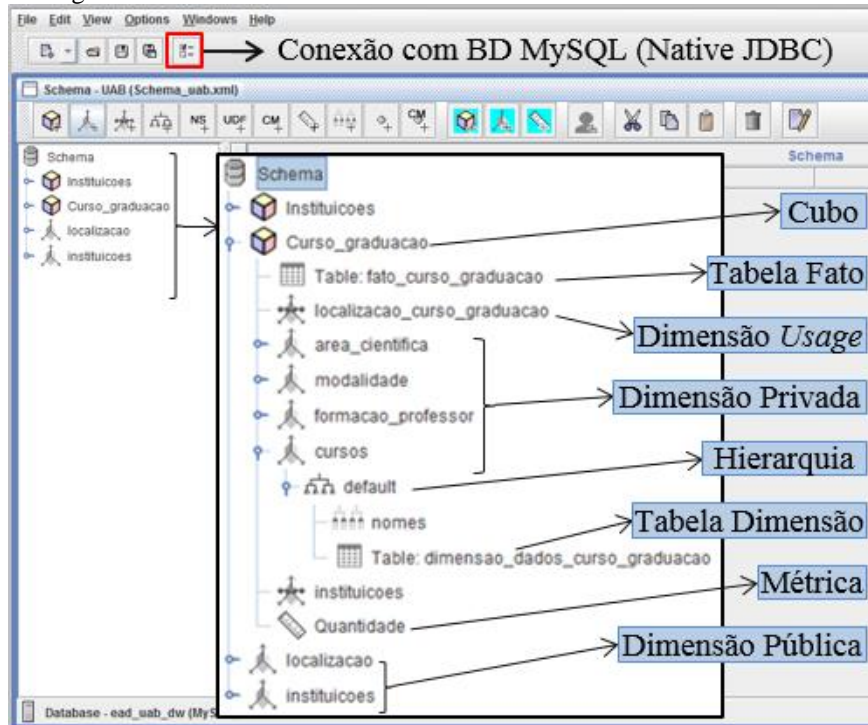
Para construção e manipulação dos dois cubos de dados foram utilizadas as seguintes ferramentas da *Suíte Pentaho*: *Pentaho Schema Workbench* (PSW) e *Pentaho BI Server*, mais especificamente o *plugin Saiku Analytics View* (SAV). A ferramenta PSW foi utilizada para criação dos cubos de dados (tabelas fato, tabelas dimensão e métricas) dos esquemas multidimensionais do E-MEC e UAB.

A Figura 1 ilustra o esquema multidimensional da UAB criado por meio da ferramenta PSW. O mesmo foi feito para o esquema do e-MEC. Para execução das operações OLAP sobre os dois cubos de dados, utilizou-se a ferramenta SAV, que por sua vez permite a visualização e análise dos dados, fornecendo ao analista informações detalhadas, sumarizadas e agregadas da área de estudo.

A Quadro 1 apresenta as dimensões e fatos pertencentes ao cubo de dados multidimensional, tanto do e-MEC quanto da UAB. Pode-se observar na Quadro 1 os *fatos* e seus relacionamentos com as dimensões, por exemplo, as dimensões “*organização_academica*” e “*categoria_adm*” estão associadas apenas a uma tabela de fato (*instituição_ies*), tanto para o cubo do e-MEC quanto para o cubo da UAB. Já as dimensões “*dados_ies*” e “*localização*” estão associadas a três tabelas de fato em relação ao cubo do e-

MEC (curso_graduação, curso_especialização e instituição_ies), já para o cubo da UAB, essas três dimensões estão relacionadas a duas tabelas de fato (curso_graduação e instituição_ies).

Figura 1 – Schema Workbench: cubo de dados multidimensional da UAB.



Fonte: Elaborada pelos autores a partir do software Pentaho Schema Workbench.

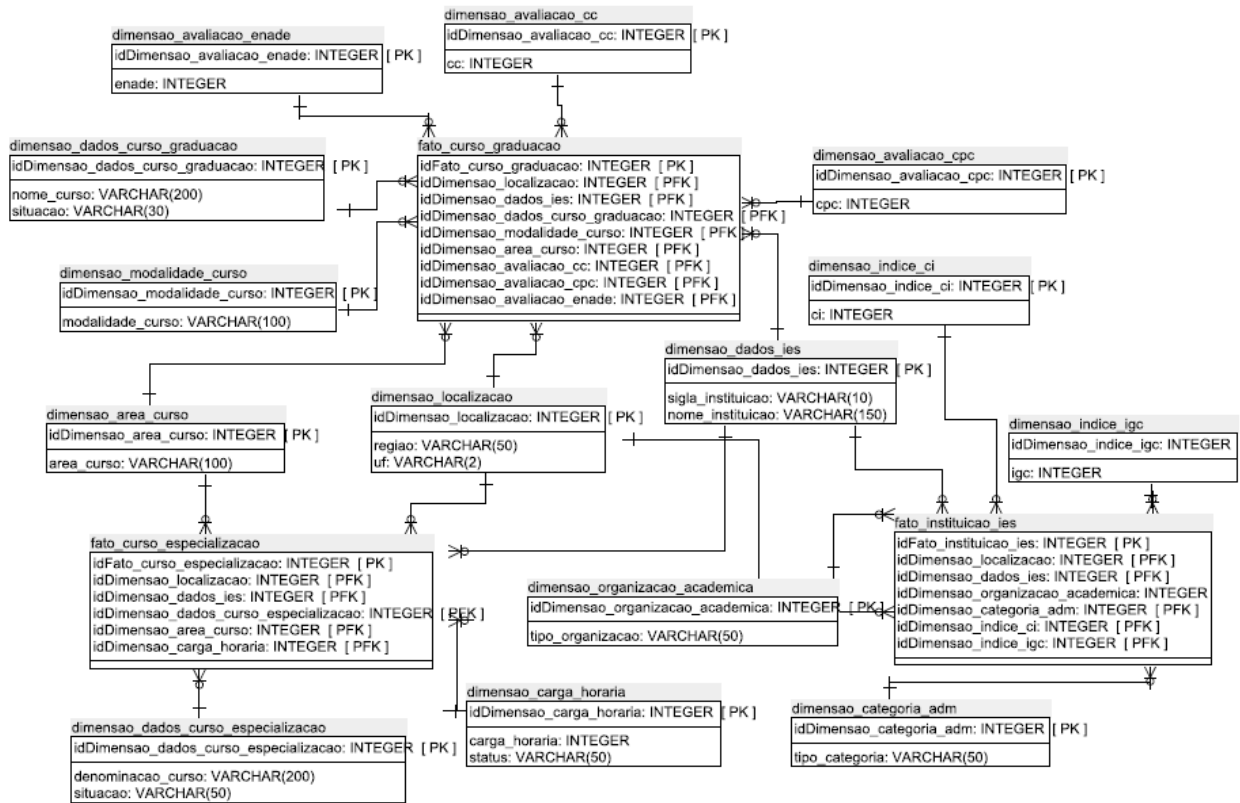
Quadro 1 – Fatos e dimensões dos cubos de dados do e-MEC e UAB.

		Fatos		
		Cursos de Graduação	Cursos de Especialização	Instituições de Ensino Superior (IES)
Dimensões	dados_curso_graduação	e-MEC / UAB		
	modalidade_curso	e-MEC / UAB		
	avaliação_curso_cc	e-MEC		
	avaliação_curso_cpc	e-MEC		
	avaliação_curso_enade	e-MEC		
	área_curso	e-MEC	e-MEC	
	dados_ies	e-MEC / UAB	e-MEC	e-MEC / UAB
	Localização	e-MEC / UAB	e-MEC	e-MEC / UAB
	dados_curso_especialização		e-MEC	
	carga_horária		e-MEC	
	organização_acadêmica			e-MEC / UAB
	categoria_adm			e-MEC / UAB
	índice_ci			e-MEC
	índice_igc			e-MEC
	área_científica	UAB		
formação_professor	UAB			

Fonte: elaborado pelos autores.

A modelagem dimensional do sistema e-MEC está representada no diagrama Entidade Relacionamento Estendido (EER) disposto na Figura 2.

Figura 2 – Modelagem dimensional dos dados do sistema e-MEC.



Fonte: elaborada pelos autores.

O cubo do e-MEC possui 14 (quatorze) tabelas de dimensão e 3 (três) tabelas fato. As tabelas fato pertencentes ao cubo do e-MEC são: curso_graduação, curso_especialização e instituicao_ies, dispondo das seguintes métricas: (I) quantidade de cursos de graduação, oriunda da dimensão “dados_curso_graduação”; (II) Min, Max, AVG do Conceito do Curso (CC), Conceito Preliminar de Curso (CPC) e Exame Nacional de Desempenho dos Estudantes (ENADE); (III) Min, Max, AVG do Conceito Institucional (CI) e Índice Geral de Cursos (IGC) da instituição; (IV) quantidade de cursos de especialização, oriunda da dimensão “dados_curso_especialização”; (V) quantidade de instituições de ensino superior, oriunda da dimensão “dados_ies”. As tabelas de dimensão pertencentes ao cubo do e-MEC são: dados_curso_graduação, modalidade_curso, avaliacao_curso_cc, avaliacao_curso_cpc, avaliacao_curso_enade, area_curso, dados_ies, localizacao, dados_curso_especializacao, carga_horaria, organizacao_academica, categoria_adm, indice_ci, indice_igc e área_científica.

O cubo da UAB possui 8 (oito) tabelas de dimensão e 2 (duas) tabelas de fato. As tabelas de fato são: curso_graduação e instituição_ies, dispondo das seguintes métricas: (I) quantidade de cursos de graduação, oriunda da dimensão “dados_curso_graduação”; (II) quantidade de instituições de ensino superior, oriunda da dimensão “dados_ies”. As tabelas de dimensão do cubo de dados multidimensional da UAB são: curso_graduação, modalidade_curso, dados_ies, localização, área_científica, organização_academica, categoria_adm e formação_professor.

4 Resultados e discussão

No panorama atual da educação superior, faz-se necessário que as IFES desenvolvam inovações tecnológicas a fim de aperfeiçoar suas práticas administrativas e pedagógicas. Diante desse contexto, o presente trabalho emprega as tecnologias DW e OLAP como facilitadoras para descoberta de novos conhecimentos sobre dados abertos educacionais, afinal, quando as tecnologias de informação e comunicação são bem aplicadas, tornam-se um diferencial para as IFES que buscam excelência nos serviços prestados.

A EaD encontra-se em expansão, e tal crescimento da modalidade a distância no cenário brasileiro deve-se, em grande parte, aos programas governamentais que impulsionam e incentivam o acesso de estudantes ao ensino superior. Há uma crescente necessidade das instituições educacionais em implementar e melhorar as metodologias praticadas na EaD. Isso ocorre em virtude de a educação *online* flexibilizar os processos de ensino e aprendizagem, facilitar a disseminação de informações e interação entre docentes e discentes, além de expandir as universidades. A EaD se destaca em virtude de suas plataformas computacionais adaptativas, que possibilitam aos estudantes uma aprendizagem mais interativa e colaborativa, sem afetar ou interferir na identidade, qualidade e valores da instituição responsável pela oferta de cursos a distância.

Inicialmente foi realizado um levantamento quantitativo de instituições públicas com oferta de cursos a distância por região, sendo a região sudeste a que possui mais instituições de ensino superior (50,2%), seguida pela região nordeste com 24%, região sul com 10,8%, região norte com 8,7%, e região centro-oeste com 6,3%.

Ao aplicar as operações OLAP – entre elas *slice and dice*, *roll-up*, *drill-up*, *drill-down*, *drill-across*, *drill-through* e *pivot* – no cubo multidimensional formado pelos dados do e-MEC, implementado no *software Saiku Analytics View*, chegou-se a um montante de 287 instituições públicas ofertando cursos de EaD no Brasil no ano de 2014. Desse conjunto, 37,3% são universidades, 47,7% são faculdades – representam a maioria das instituições, 14%

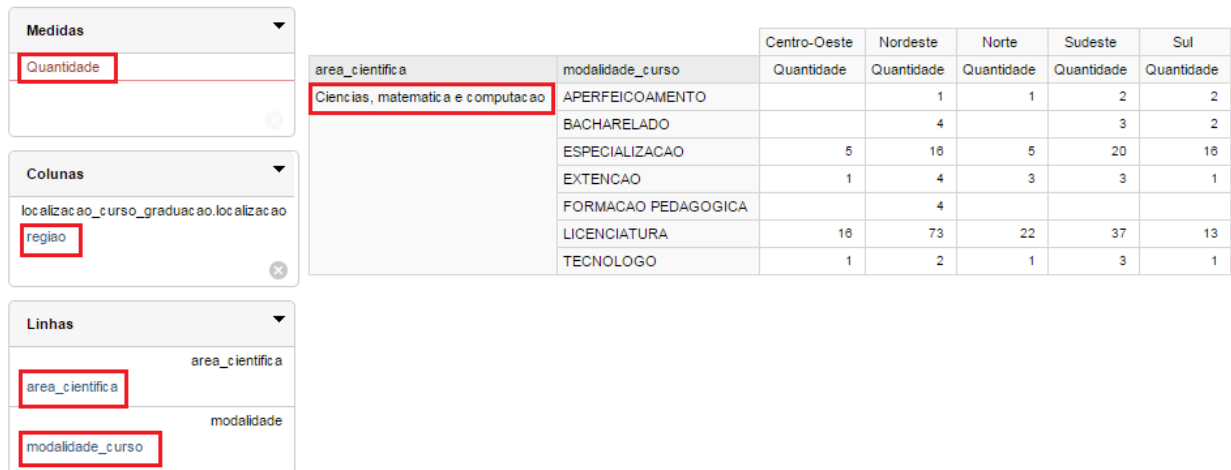
são Institutos Federais de Educação, Ciência e Tecnologia (IFET) e 1% são centros universitários. As organizações acadêmicas estão distribuídas pelas regiões geográficas do Brasil da seguinte forma: Centro-Oeste (18 organizações acadêmicas); Nordeste (69); Norte (25); Sudeste (143); e Sul (30). Das 287 instituições, 44% são instituições pública estadual, 37,2% públicas federal e 18,8 pública municipal.

Ao realizar as operações OLAP no cubo multidimensional referente aos dados da UAB, também implementado no *software Saiku Analytics View*, obteve-se informações quantitativas acerca dos cursos da EaD por região geográfica e grau de ensino (ou modalidade do curso), sendo a maior concentração de cursos na região Nordeste (38,1%), enquanto que a menor concentração está no Norte (10,2%). Existem 8 (oito) graus de ensino, sendo a especialização o grau de ensino mais ofertado, com 38,1%, seguido por licenciatura (30,1%), aperfeiçoamento (15,5%), bacharelado (8,3%), extensão (5,9%), tecnológico (1,1%), formação pedagógica (0,9%) e sequencial (0,1%).

Ao considerar as informações sobre os cursos de especialização por região geográfica e áreas científicas, dispostas no cubo de dados do e-MEC, verificou-se que o cenário se difere em comparação à análise realizada com todos os graus de ensino, já que o Sudeste se sobressaiu em relação ao Nordeste, isto é, há uma maior concentração de cursos de especialização no Sudeste (57%), seguida pela região Sul (21%), Nordeste (11%), Centro-oeste (9%), enquanto que a menor concentração continua sendo o Norte (2%). Em se tratando da área científica com mais cursos de especialização a distância no Brasil, tem-se “Ciências sociais, negócios e direito”, representando 46% do total de cursos de especialização, seguida pela área “Educação” com 32,9%, “Ciências, matemática e computação” com 6,4%, “Saúde e bem-estar social” com 5,5%, “Humanidades e artes” com 4,7%, “Engenharia, produção e construção” com 3%, “Serviços” com 0,8%, e “Agricultura e veterinária” com 0,7%.

Em se tratando da distribuição numérica de cursos de graduação por região geográfica e área científica, verifica-se por meio do cubo de dados da UAB que a área científica “Educação” possui mais cursos (40,2%), seguida pela área “Ciências sociais, negócios e direito” (31,7%), e pela área “Ciência, matemática e computação” (23,7%). As demais áreas científicas correspondem a 4,5% dos cursos. Ao verificar a distribuição numérica da área científica “Ciências, matemática e computação” por região geográfica e grau de ensino, conforme ilustra a Figura 3, pode-se constatar que a região Nordeste dispõe de uma maior concentração de cursos de “Ciências, matemática e computação”, com 39,7%, enquanto que a menor concentração está no Centro-Oeste (8,82%). A licenciatura é o grau de ensino que prevalece na referida área, com 61,5%, seguida por especialização com 23,7%.

Figura 3 – Cursos da área científica “Ciências, matemática e computação” por grau de ensino e região geográfica do Brasil.



Fonte: elaborada pelos autores a partir do *software Saiku Analytics View*.

Ainda de acordo com cubo de dados da UAB, foi possível analisar os 26 cursos de graduação em Computação ofertados na modalidade a distância no Brasil, com as seguintes denominações: Ciências da Computação, Computação, Informática, Sistemas de Informação, Tecnologia em Análise e Desenvolvimento de Sistemas, Tecnologia em Desenvolvimento de Sistemas de Informação, Tecnologia em Sistema para Internet, Tecnologia em Sistemas de Computação, e Tecnologia em Sistemas para Internet a Distância. Na região Nordeste encontra-se metade dos cursos de Computação, seguida pela região Sudeste (23%), Centro-oeste e Norte (11,6% cada) e Sul (3,8%). O grau de ensino “Licenciatura” é o que prevalece na área de Computação, com 65%, seguido por “Tecnólogo” com 19%, e “Bacharelado” com 16%.

Em relação à visão quantitativa geral de instituições e cursos da EaD, a região Nordeste é a que possui maior número de instituições com oferta de cursos em EaD (33 instituições), maior quantidade de cursos em EaD (421) e maior quantidade de cursos de graduação em Computação na modalidade a distância (13 cursos de Computação). Seguida pela região Sudeste, que possui 28 instituições com oferta de cursos em EaD, 272 cursos em EaD e 6 cursos de graduação em Computação na modalidade a distância. As referidas informações foram obtidas do cubo de dados da UAB implementado no *Saiku Analytics View*.

Também foram analisadas informações referentes aos maiores e menores índices “Conceito de Curso (CC)” e “Exame Nacional de Desempenho dos Estudantes (ENADE)”, cujos indicadores variam de 1 a 5. As instituições que ofertam cursos da área “Ciências, Matemática e Computação” que possuem os maiores índices CC são: Universidade de São Paulo e Universidade Federal de Juiz de Fora, com índice CC igual a 5 (cinco). E as

instituições com os maiores índices ENADE são: Universidade Federal de Pelotas e Universidade Federal Fluminense, com índice ENADE igual a 5 (cinco).

A média do índice CC na região Centro-oeste é 3,75, na região Nordeste é 3, na região Norte é 1,9, na região Sudeste é 4,2 e na região Sul é 4,3. Segue a média do índice CC por área científica: (I) área “Serviços”: média CC igual a 4,3; (II) área “Educação”: média CC igual a 4,2; (III) área “Ciências sociais, negócios e direito”: média CC igual a 4,1; (IV) área “Humanidades e artes”: média CC igual a 4; (V) área “Saúde e bem-estar social”: média CC igual a 4; (VI) área “Ciências, matemática e computação”: média CC igual a 3,7; (VII) área “Agricultura e veterinária”: média CC igual a zero; (VIII) área “Engenharia, produção e construção”: média CC igual a zero.

Dentre as 287 instituições de ensino superior analisadas, 4,5% (13 instituições) possuem a pontuação máxima no Índice Geral de Cursos (IGC = 5), são elas: Escola de Governo Professor Paulo Neves de Carvalho, Faculdade de Tecnologia de Ourinhos, Fundação Universidade Federal do ABC, Instituto Militar de Engenharia, Instituto Tecnológico de Aeronáutica, Universidade Estadual de Campinas, Universidade Federal de Lavras, Universidade Federal de Minas Gerais, Universidade Federal de Santa Catarina, Universidade Federal de São Carlos, Universidade Federal de São Paulo, Universidade Federal de Viçosa e Universidade Federal do Rio Grande do Sul. Apenas 0,3% das instituições possuem IGC mínimo (IGC = 1), são elas: Faculdade de Ciências Agrárias de Araripina, e Faculdade de Ciências Humanas e Sociais Aplicadas do Cabo de Santo Agostinho. Todas as informações referentes aos IGC foram obtidas do cubo de dados do e-MEC, implementado no *software Saiku Analytics View*.

Uma das dimensões do cubo de dados da UAB aborda a formação do professor, cuja temática foi abordada por Ponticelli et al. (2013). Segundo os autores, faz-se necessário aumentar o número de professores com graduação em *licenciatura* em sua área específica de atuação, e que uma das soluções que o governo federal vem adotando é o fomento de diversas políticas públicas de incentivo à formação desses profissionais por intermédio da modalidade de educação a distância. Ainda nessa temática, Purificação e Pessoa (2015) refletem sobre os desafios dos programas de formação de professores frente às inovações tecnológicas, chegando à conclusão que ao se fazer uso da tecnologia, o professor se familiariza com a tecnologia e que seu domínio pode levá-lo a um novo olhar sobre ela, auxiliando-o na produção de mudanças de sua prática pedagógica.

5 Conclusões

Após análises realizadas no trabalho, percebe-se que a modalidade de educação a distância tem crescido no âmbito nacional, isso mostra que a EaD está cada vez mais acessível no Brasil, com vistas à interiorização e democratização do ensino superior. Contudo, faz-se necessário continuar avaliando o desempenho dos cursos, a fim de elucidar os principais obstáculos enfrentadas pelas instituições e estudantes da modalidade a distância.

Com o aumento na quantidade de vagas e formas de ingresso na educação superior do Brasil, seja por meio do sistema de cotas, Exame Nacional do Ensino Médio – ENEM, Sistema de Seleção Unificada – SISU, bolsas e financiamentos, tais como Fundo de Financiamento Estudantil – FIES, Programa Universidade para Todos – ProUni, entre outros, muitos estudantes que antes não eram admitidos pelos processos seletivos tradicionais, passaram a ter oportunidade de ingressar em uma universidade/faculdade. Com essa forma de ingresso ampliada e mais simplificada, o estudante tende a mudar de curso e instituição com mais frequência, e essas mudanças refletem nos índices de avaliação dos cursos (Conceito de Curso – CC, Índice Geral de Cursos – IGC), índices de avaliação dos estudantes (notas no ENADE), e índices de avaliação das instituições (Conceito Institucional – CI).

Portanto, uma das principais motivações das análises realizadas nesse trabalho, por meios do DW e OLAP, é compreender o perfil desse novo estudante, visando obter uma visão contextualizada e realista da EaD no Brasil, em especial da área de conhecimento “ciências exatas e da terra”, subárea “computação”, e a partir dessa investigação, mediante trabalhos complementares a este, indicar medidas para evitar a evasão dos estudantes e propor melhorias na gestão e práticas pedagógicas das instituições e cursos da EaD.

O DW proporcionou um ambiente analítico e de apoio à decisão eficiente, sendo bastante útil para exploração dos dados educacionais de diferentes granularidades. Por meio do DW os dados detalhados ou sumarizados ficam acessíveis às instituições, que podem categorizá-los de várias maneiras, elaborar suas próprias consultas não triviais, bem como investigar, de forma dinâmica, possíveis correlações. Contudo, como trabalhos futuros pretende-se ampliar o escopo a fim de expandir as análises no contexto da EaD, buscando utilizar, além dos dados abertos do e-MEC e UAB, dados específicos dos sistemas de gestão acadêmica das IFES. Almeja-se também a aplicação de técnicas de mineração de dados por meio dos paradigmas do aprendizado de máquina supervisionado (algoritmos de classificação) e não-supervisionado (algoritmos de associação e *clustering*), a fim de identificar padrões sobre evasão, reprovação, rendimento acadêmico e perfil socioeconômico dos estudantes.

IMPLEMENTATION OF A DATA WAREHOUSE FOR ANALYSIS OF OPEN GOVERNMENTAL DATA IN DISTANCE EDUCATION

Abstract: The paper presents an implementation of a multidimensional educational database to analysis of Brazilian higher education, focusing on Distance Education - Distance Education. To support the decision support process, a Data Warehouse (DW) was designed, based on the facts constellation dimensional modeling and the application of OnLine Analytical Processing (OLAP) technologies. Open data of high granularity of the e-MEC and Open University of Brazil systems were used. The work is based on the development of an analytical computational environment aiming at tracing the profile of educational institutions active in EaD, as well as university and courses, especially the area of knowledge "exact sciences and land" and subarea "computing". Therefore, after the analyzes carried out in the work by means of DW, it is verified that the modality of distance education is increasingly accessible in Brazil, aiming at the interiorization and democratization of higher education.

Keywords: Distance Education. Open Data. Data Warehouse. Dimensional Modeling.

Referências

- ABED – Associação Brasileira de Educação a Distância. **Censo EAD BR**: relatório analítico da aprendizagem a distância no brasil. Curitiba: InterSaberes, 2016.
- FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery. **AI Magazine**, p. 37-54, 1996.
- FREITAS, C. N.; GOUVEIA, R. M. M.; SILVA, A. Online Analytical Processing em ambientes virtuais de aprendizagem da educação a distância. In **Anais do 4º DesafIE - Workshop de Desafios da Computação Aplicada à Educação – XXXV Congresso da Sociedade Brasileira de Computação**, 2015.
- GOTTARDO, Ernani; KAESTNER, Celso A. A.; NORONHA, R. V. Estimativa de desempenho acadêmico de estudantes: análise da aplicação de técnicas de mineração de dados em cursos a distância. **Revista Brasileira de Informática na Educação**, v. 22, p. 45-55, 2014.
- GOUVEIA, R. M. M. **Mineração de Dados em data warehouse para sistema de abastecimento de água**. Dissertação (Mestrado) - Universidade Federal da Paraíba, João Pessoa, PB, 2009.
- GUÉRCIO, Hugo; MARQUES, Phillipe; STRÖELE, Victor; PEREIRA, Crystiam Kelle; BARRERE, Eduardo. (2014) Análise do desempenho estudantil na educação a distância aplicando técnicas de mineração de dados. In: **Anais do III Congresso Brasileiro de Informática na Educação**, Dourados/MS. Workshop de Mineração de Dados em Ambientes Virtuais de Ensino/Aprendizagem, p. 641-650, 2014
- INMON, William H. **Building the Data Warehouse: getting started**. 4a Edição. Editora: Wiley Publishing, Inc, 2005.



JOSÉ JÚNIOR, Oliveira; BASTOS, Laudelino; KAESTNER, Celso. Uma abordagem de data warehouse educacional para apoio à tomada de decisão. In: **Anais do Workshop de Mineração de Dados Educacionais do Congresso Brasileiro de Informática na Educação – CBIE**, 2015

KANASHIRO, Augusto. **Um data warehouse de publicações científicas: indexação automática da dimensão tópicos de pesquisa dos data marts**. Tese (Doutorado). Universidade de São Paulo, 2007

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit: the complete guide to dimensional modeling**. John Wiley & Sons, 2011.

PÊGAS, Daniel dos Santos; YANO, Edgar Toshiro. O uso de data warehousing no processo de aprendizagem de programação. In: **Anais do Simpósio Brasileiro de Informática na Educação**. p. 534-543, 2004.

PONTICELLI, Fernanda A.; ZUCOLOTTO, Andreia M.; BELUCO, Alexandre. A Educação a Distância como auxílio na formação de professores. **#Tear: Revista de Educação, Ciência e Tecnologia**, v. 2, n. 2, 2013.

PRADO, Marcelo Vinaud. **Data warehouse para apoio à gestão da operação em empresas do transporte rodoviário interestadual de passageiros**. 2006. 111 f. Dissertação (Mestrado) - Universidade de Brasília, Brasília, 2006.

PURIFICAÇÃO, Marcelo Máximo; PESSOA, Teresa. O ensino da matemática em meio à tecnologia: desafio aos programas de formação de professores. **#Tear: Revista de Educação, Ciência e Tecnologia**, v. 4, n. 2, 2015