ENEPEX
ENCONTRO DE ENSINO, PESQUISA E EXTENSÃO
8° ENEPE UFGD • 5° EPEX UEMS

MODELAGEM DE DADOS COM *DATA WAREHOUSE* E OLAP: UM ESTUDO DE CASO

Maycon Henrique Trindade[1]; Francy Helder Silva de Almeida[2]; Everton Castelão Tetila[3]

UFGD/FACET – Caixa Postal 364, 79.804-970 – Dourados – MS, E-mail: mayconhenrique@live.com
[1]Bacharelando do Curso de Sistemas de Informação da UFGD. [2]Bacharel em Sistemas de Informação UFGD
[3]Orientador. Professor FACET

RESUMO

Os *data warehouses* (ou armazém de dados) oferecem acesso a dados para análise complexa, descoberta de conhecimento e tomada de decisão. O modelo de dados multidimensional utilizado nos *data warehouses* dá aos tomadores de decisão informações no nível correto de detalhe, com base em organização e perspectiva apropriadas. Além disso, é uma boa escolha para Processamento Analítico *On-line* (OLAP). Nesse sentido, este trabalho propõe criar uma base de conhecimento, a partir de um *data warehouse*, para investigar as causas potenciais dos problemas do curso de Bacharelado em Sistemas de Informação da UFGD, tais como, reprovação, evasão e baixo índice de diplomação. Para a realização deste trabalho, foram utilizados os aplicativos *MySQL 5.6*, *MySQL Workbench 6.0* e *Pentaho BI-Server CE-4.8.0*.

Palavras-chave: *data warehouse*, OLAP, Inteligência Empresarial.

1 INTRODUÇÃO

O amadurecimento do conceito de *Data Warehouse* na década de 1980 aconteceu junto com a evolução dos sistemas de informação desenvolvidos pelo mercado. Estimulados pelos mercados de massa, os sistemas transacionais cresceram de tal forma que passaram a apresentar dificuldades para atender as necessidades mais evoluídas de análise de negócio, principalmente devido à dificuldade de conciliar, em ambientes de alta concorrência transacional, a possibilidade de geração de relatórios em tempo real [GONÇALVES, 2012]. Nasceu, então, o conceito de D*ata Warehouse* (DW), que pode ser definido como “uma coleção de dados

orientada a assunto, integrada, não volátil, variável no tempo para o suporte de apoio às decisões da gerência" [INMON, 1992].

Gonçalves (2012) acrescenta que para atender essas finalidades, um dos requisitos do DW é que esse repositório seja totalmente separado do ambiente de produção, com modelagem própria e armazenamento distinto. Dessa forma, as consultas envolvendo dados com visões analíticas ou resumidas em um contexto temporal se tornam possíveis, contribuindo para o processo de tomada de decisões.

À medida que o DW é um ambiente projetado para execução de consultas analíticas, os dados em um DW não podem ser alterados. Portanto, somente duas operações podem ser realizadas: a carga e a consulta dos dados.

Segundo Elmasri & Navathe (2011), os DW oferecem acessos a dados para análise complexa, descoberta de conhecimento e tomada de decisão. Eles dão suporte a demandas de alto desempenho sobre os dados e informações de uma organização. Vários tipos de aplicações – OLAP, DSS e aplicações de mineração de dados – são aceitos. Definimos cada uma delas a seguir:

- **OLAP (Processamento analítico on-line):** termo utilizado para descrever a análise de dados complexos do DW.
- **DSS (sistemas de apoio à decisão):** também conhecido como **EIS – sistemas de informações executivas,** ajudam os tomadores de decisões de uma organização com dados de nível mais alto com decisões complexas e importantes.
- **Mineração de dados:** usada para *descoberta do conhecimento*, o processo de procurar novo conhecimento imprevisto nos dados.

Nesse contexto, este trabalho propõe criar um DW e analisar os dados com o Processamento Analítico On-line (OLAP), com o intuito de investigar o perfil discente, as disciplinas com maior índice de reprovação, assim como avaliar os parâmetros que influenciam na qualidade do curso de Bacharelado em Sistemas de Informação (BSI) da Universidade Federal da Grande Dourados (UFGD).

1.1 PROBLEMÁTICA E JUSTIFICATIVA

Com o passar dos anos, os cursos da área de computação passaram a ter grande procura nos vestibulares e processos seletivos de diversas instituições, públicas e privadas, no Brasil e

exterior. Essa procura deve-se, principalmente, ao crescimento da área de computação e, sobretudo, ao interesse da sociedade às questões relacionadas à tecnologia.

Apesar dessa grande procura, existe um dilema: muitas estudantes ingressam nas universidades, cursos técnicos e institutos federais, no entanto, uma pequena parcela de graduados nessa área apresentam-se ao mercado de trabalho para exercer os cargos disponíveis nas empresas de tecnologia. "Segundo o IBGE, o setor de serviço da informação cresceu quase 5% no ano de 2011, ficando à frente de setores importantes da economia como o da construção civil, indústria e comércio. Por outro lado, no setor de Tecnologia da Informação (TI) existe um déficit de 115 mil vagas de trabalho" [REDE GLOBO DE TELEVISÃO, 2012].

1.2 METODOLOGIA

Segundo Gil (2010), uma pesquisa pode ser classificada em: exploratória, descritiva ou explicativa.

A proposta metodológica deste trabalho se pauta na ação de levantamento bibliográfico, contemplada em corpo conceitual mais amplo da pesquisa exploratória, conforme anunciado por Gil (2010).

Ainda segundo o autor, existem várias estratégias de pesquisa. Uma delas, aqui utilizada, é o Estudo de Caso. "O estudo de caso envolve o estudo profundo e exaustivo de um ou poucos objetos de maneira que se permita o seu amplo e detalhado conhecimento" [GIL 2010]. De tal modo, um estudo de caso foi realizado com a base de dados do curso de BSI da UFGD. O pré-processamento desses dados foi realizado para a carga no DW, por meio do Sistema Gerenciador de Banco de Dados (SGBD) *MySQL 5.6*. Depois disso, os dados do DW foram projetados para a execução do OLAP com o aplicativo *Pentaho BI Server CE-4.8.0* (PENTAHO, 2014). Por fim, os resultados foram analisados com base na experiência pessoal dos participantes do estudo.

A seguir, são descritos os softwares utilizados para os processos de coleta, armazenamento, modelagem e consultas analíticas *on-line*.

- ***Mysql Database Server*** versão 5.6.12 – utilizado para armazenar os dados coletados, disponível em: <http://dev.mysql.com/get/Downloads/MySQLInstal-ler/mysql-instal-ler-community-5.6.12.0.msi/from/http://cdn.mysql.com/>
- ***Mysql Workbench*** versão 6.0.7 – utilizado para criar o modelo multidimensional (*data warehouse*), disponível em: <http://cdn.mysql.com-/archives/mysql-workbench-/mysql-workbench-community-6.0.7-win32.msi>

- ***Pentaho open source business intelligence*** versão 4.8.0.stable – utilizado para executar o OLAP, disponível em: <http://ufpr.dl.sourceforge.net/project/pentaho-/Business%20Intelligence%20Server/4.8.0-stable/biserver-ce-4.8.0-stable.zip>

2 ARMAZÉM DE DADOS (DATA WAREHOUSE)

Os bancos de dados tradicionais têm suporte para o Processamento de Transações On-line (OLTP), que inclui inserções, atualizações e exclusões, enquanto também têm suporte para requisitos de consulta de informação. Os bancos de dados relacionais tradicionais são otimizados para processar consultas que podem tocar em uma pequena parte do banco de dados e transações que lidam com inserções ou atualizações no processo de algumas tuplas por relação. Assim, eles não podem ser otimizados para OLAP, DSS ou mineração de dados. Ao contrário os DW são projetados exatamente para dar suporte à extração, processamento e apresentação eficientes para fins analíticos e de tomadas de decisão.

Em comparação com os bancos de dados transacionais, os DW não são voláteis. Isso significa que as informações no DW mudam com muito menos frequência e podem ser consideradas não de tempo real com atualização periódica. Em sistemas transacionais, as transações são a unidade e o agente de mudança no banco de dados; ao contrário a informação no DW é muito menos detalhada e atualizada de acordo com uma escolha cuidadosa de política de atualização, normalmente incremental [ELMASRI & NAVATHE, 2011].

Outra característica importante no projeto de um DW é a granularidade (nível de detalhamento nos dados). Quanto menos detalhes, mais alto o nível de granularidade. A granularidade nos primeiros sistemas operacionais era tida como certa, pois quando os dados eram atualizados, certamente seria ao mais baixo nível de granularidade, sendo que no ambiente de DW, a granularidade não é um pressuposto [INMON 1997].

3 ESTUDO DE CASO

Esta seção apresenta o estudo de caso realizado a partir dos dados referentes aos históricos acadêmicos dos cursos de BSI e Análise de Sistemas (AS) da UFGD. Inicialmente, foram coletados os dados referentes à aprovação, reprovação, diplomação e evasão dos discentes, entre o período de 2006 a 2012 (Seção 3.1). Em seguida, os dados foram pré-processados para a carga no DW (Seção 3.2). Depois disso, o OLAP foi realizado com o aplicativo *Pentaho BI Server CE-4.8.0*. Por fim, os resultados foram analisados com base na experiência pessoal dos participantes do estudo (Seção 3.3).

3.1 COLETA DE DADOS

Para a coleta de dados, uma solicitação dos históricos acadêmicos do curso de BSI e AS foi encaminhada à Pró-reitoría de Graduação (PROGRAD) da UFGD. Essa solicitação foi formalizada por meio de uma Comunicação Interna (CI) e os relatórios, indispensáveis para a realização deste trabalho, foram entregues em formato digital (.xls e .csv) com os seguintes atributos: ano de ingresso, curso, acadêmico (nome do acadêmico), sexo, data de nascimento, disciplina (nome da disciplina), semestre (primeiro ou segundo), RGA (registro acadêmico), nota (nota final na disciplina), falta (quantidade de faltas), resultado (aprovado, reprovado por nota, reprovado por falta), tipo estado final (indica o estado final do acadêmico: diplomação, transferência, evasão ou regularmente matriculado), tem filhos, atividade remunerada, etnia (branco, pardo, amarelo, indígena, negro) e cidade.

Os dados dos históricos acadêmicos são representados através de uma planilha eletrônica com 14.413 registros, referente a 400 acadêmicos ingressos entre os anos 2006 e 2012. Os valores de registro dos atributos "Acadêmico", "Logradouro" e "Bairro" foram alterados para garantir a privacidade dos dados acadêmicos.

3.2 PRÉ-PROCESSAMENTO DE DADOS E MODELO MULTIDIMENSIONAL

O Pré-processamento de dados refere-se ao processo de extração das informações coletadas a partir de diferentes sistemas, incluindo a transformação dos dados sistêmicos em informações de negócios e a carga desses dados de origem para o repositório central. Nesse contexto, para a consolidação dos dados transacionais no repositório do DW, os históricos acadêmicos foram pré-processados: alguns atributos (campos) não foram selecionados para a carga no DW, outros foram transformados para melhor análise. Por exemplo, o atributo data_nascimento foi transformado no atributo maior_25 anos para analisar as métricas (ex: diplomados) em apenas duas categorias – acadêmicos com mais ou menos de 25 anos.

O DW utiliza o modelo multidimensional, baseado em tabelas fato e dimensão. A Figura 1 mostra o modelo multidimensional criado com o aplicativo *MySQL Workbench 6.0*[1] para o OLAP no *PentahoBI-Server CE-4.8.0*.

[1]O *MySQL Workbench* também pode ser utilizado para converter o modelo multidimensional em *script* SQL. Então, a partir das instruções SQL é possível criar o DW no servidor de banco de dados *MySQL*.

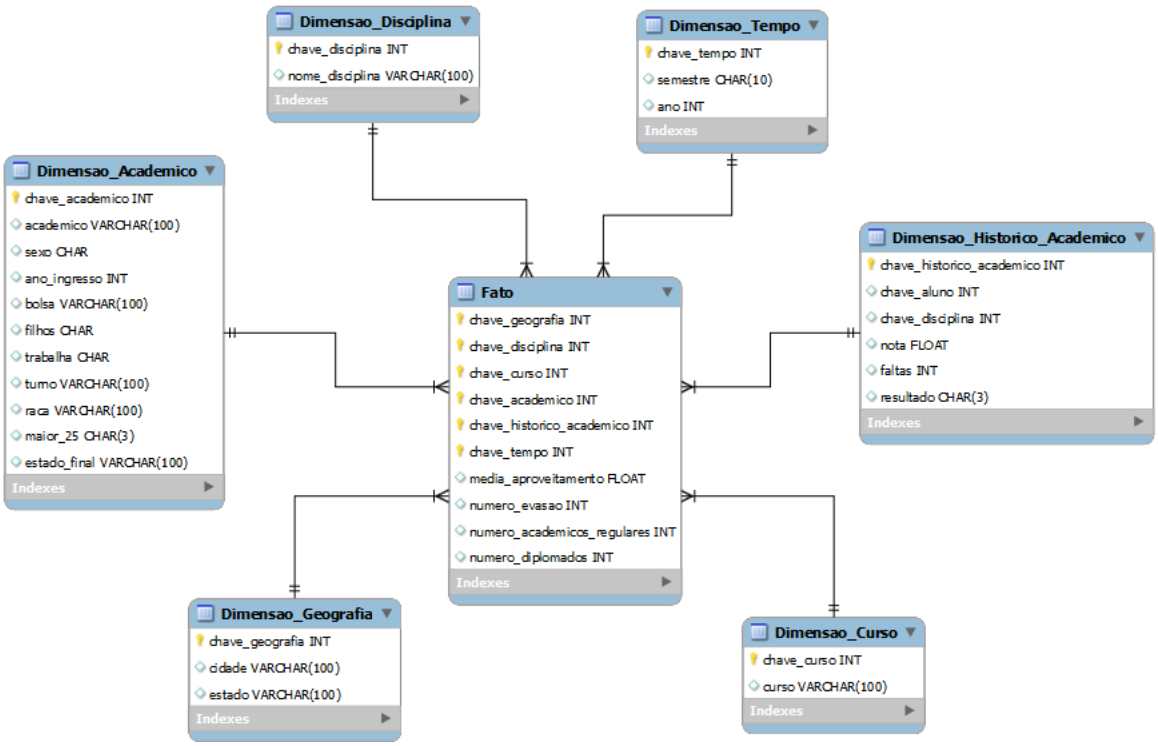

Figura 1 – Projeto lógico do modelo multidimensional de dados.

Kimball (1998) informa que para se distinguir quais campos de dados serão fatos e quais serão atributos (de dimensão) ao projetar um banco de dados, podemos usar a regra: se o dado for numérico e alterar a cada amostragem é fato, se for uma descrição constante de um item será um atributo de dimensão.

Observe que o modelo utiliza seis tabelas dimensão e uma tabela fato para armazenar os dados dos históricos acadêmicos. A tabela fato contém as métricas ou os fatos que estamos interessados em medir. As tabelas dimensão, por sua vez, relacionam-se com a tabela fato e contém os atributos da dimensão. Assim, visões analíticas envolvendo as métricas media_aproveitamento, numero_evasao, numero_academicos_regulares e numero_diplomados podem ser resumidas em um contexto temporal ou outro nível de detalhe específico, como geografia e curso.

3.3 ANÁLISE E INTERPRETAÇÃO DOS DADOS

Esta seção apresenta a análise dinâmica e multidimensional dos dados através de visões criadas no OLAP *Mondrian*. Os resultados, aqui apresentados, são discutidos a seguir.

		Measures
Tempo	Curso	Media_aproveitamento
2006	SISTEMAS DE INFORMAÇÃO - BACHARELADO	4,614
2007	SISTEMAS DE INFORMAÇÃO - BACHARELADO	4,972
2008	SISTEMAS DE INFORMAÇÃO - BACHARELADO	5,474
2009	SISTEMAS DE INFORMAÇÃO - BACHARELADO	3,852
2010	SISTEMAS DE INFORMAÇÃO - BACHARELADO	4,774
2011	SISTEMAS DE INFORMAÇÃO - BACHARELADO	4,348
2012	SISTEMAS DE INFORMAÇÃO - BACHARELADO	3,541

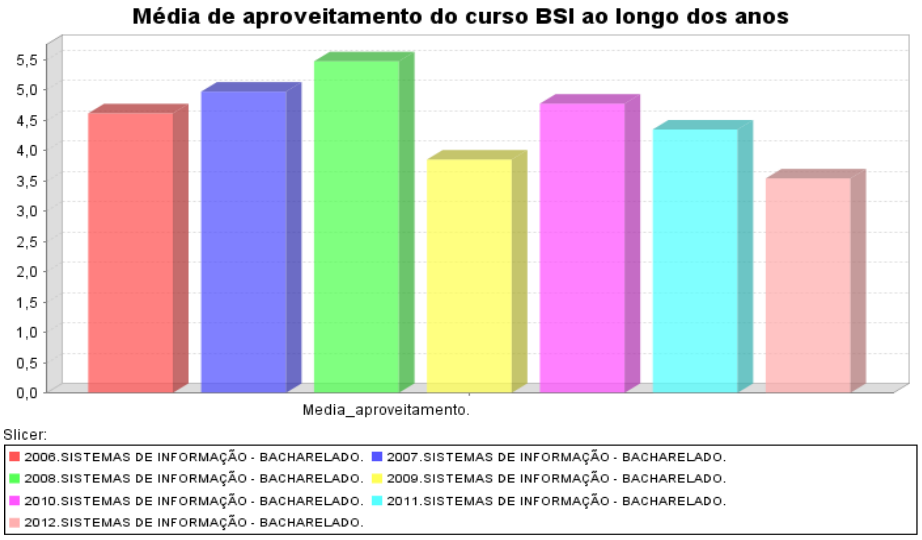

Figura 2 - Média de aproveitamento ao longo dos anos.

A Figura 2 apresenta a média de aproveitamento do curso BSI entre os anos 2006 e 2012. Após três anos de crescimento (2006, 2007 e 2008) a média de aproveitamento apresenta uma queda expressiva a partir de 2010. São causas prováveis desse declínio, passíveis de investigação: (1) o aumento do número de acadêmicos ingressos no vestibular pela Lei das Cotas (Lei nº 12.711, de 29 de agosto de 2012); (2) as mudanças nas metodologias de ensino e aprendizagem (3) as contratações de novos professores; (4) outras causas.

	Measures
Tempo	Numero_evasao
2006	43
2007	56
2008	52
2009	43
2010	24
2011	20
2012	9

Slicer:

Figura 3 - Número de evasão ao longo dos anos.

A Figura 3 apresenta o número de evasão ao longo dos anos. Note que o ano de 2007 registrou o maior número de evasão. Este número tem caído expressivamente ao longo dos anos, sendo o seu menor valor registrado em 2012. É razoável supor que uma causa potencial

para esse fenômeno tem sido o aumento da demanda por profissionais de Tecnologia da Informação, conforme discutido na Seção 1.1.

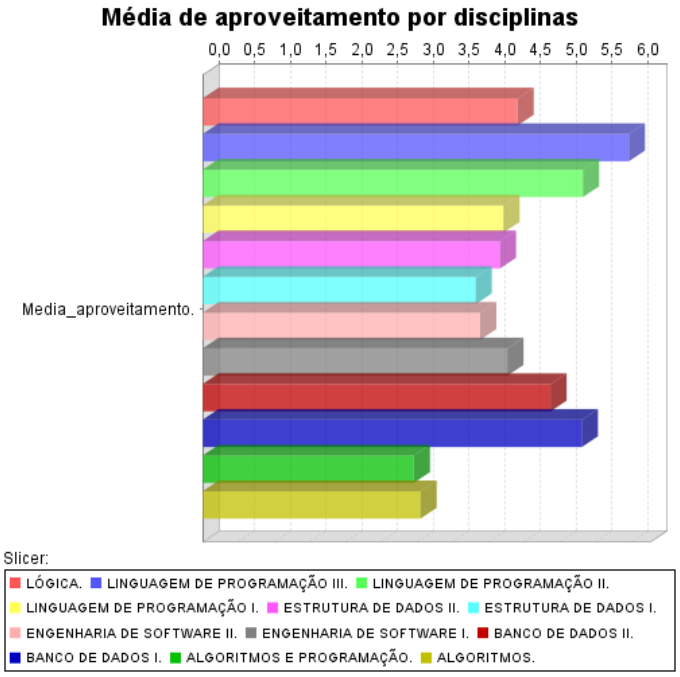

Figura 4 - Média de aproveitamento por disciplinas.

A Figura 4 compara a média de aproveitamento das principais disciplinas do curso de BSI. Essas disciplinas fazem parte da grade curricular obrigatória. É possível observar que as menores médias foram obtidas pelas disciplinas Algoritmos e Programação (2.9) e Algoritmos (3.0). Ao contrário, as disciplinas Linguagem de Programação III (5.9), Linguagem de Programação II (5.3), Banco de Dados I (5.3) e Banco de Dados II (4.9) obtiveram as maiores médias. As médias das demais disciplinas foram: Lógica (4.4), Estrutura de dados I (3.9), Estrutura de dados II (4.1), Engenharia de software I (4.2) e Engenharia de software II (3.8).

Figura 5 - Número de diplomados ao longo dos anos.

De modo semelhante à Figura 2, o número de diplomados tem diminuído ao longo dos anos, sendo a maior alta registrada em 2007, como mostra a Figura 5. De tal modo, podemos presumir que o número de diplomados tem uma relação diretamente proporcional à média de aproveitamento, representada na Figura 2. Note que ambas as variáveis – média de aproveitamento e número de diplomados – tiveram uma queda considerável nos últimos anos.

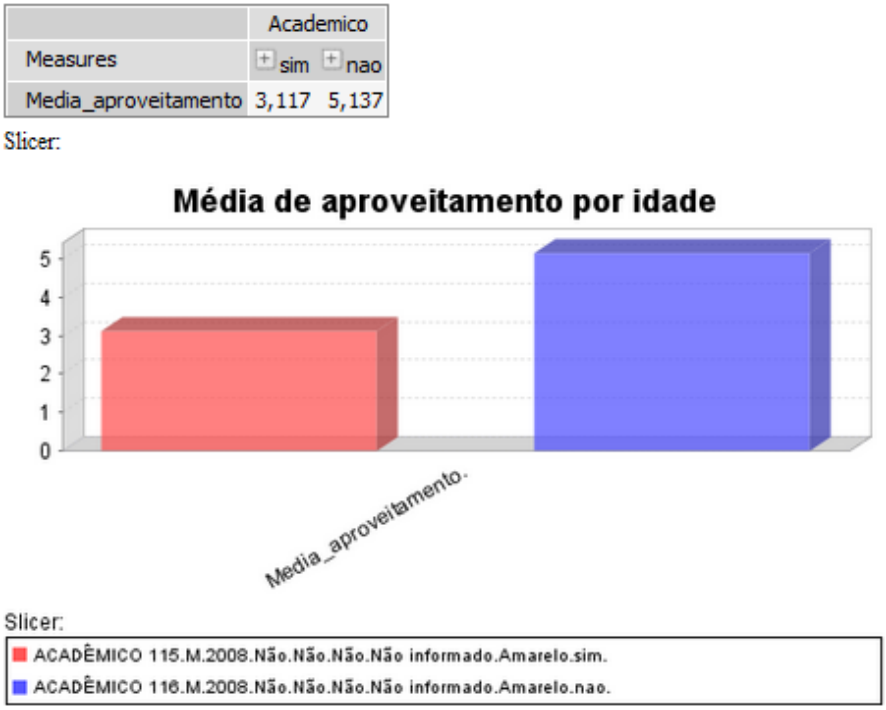

Figura 6 - Média de aproveitamento por alunos maiores e menores que 25 anos.

A Figura 6 apresenta a média de aproveitamento dos acadêmicos distribuídos por idade. Neste trabalho, consideramos 25 anos a linha de corte para melhor análise das métricas a partir de dois grupos: acadêmicos com mais ou menos de 25 anos. Esse valor é baseado nas corretoras de seguros de automóveis que classifica valores de seguro com base nos históricos de acidentes por idade.

Observe que a média de aproveitamento dos acadêmicos mais jovens é maior que a média de aproveitamento dos acadêmicos com idade superior a 25 anos. Isso confronta a tese de que acadêmicos mais experientes tem melhores médias de aproveitamento (em geral). Prováveis causas potencias para esse comportamento são: (1) trabalho, (2) tempo reduzido para o estudo; (2) vida conjugal; (3) filhos, (4) outras causas.

A Figura 7 apresenta a média de aproveitamento distribuída por 5 grupos de etnias: Amarelo, Branco, Pardo, Negro e Indígena. Essa classificação é importante para verificar o desempenho dos acadêmicos a partir da Lei de Cotas. O gráfico mostra que as etnias Indígena e Negro - que possuem reservas das vagas oferecidas pela Lei nº 12.711, de 29 de agosto de

2012 - tiveram as menores médias de aproveitamento, com os índices 2,51 e 3,47, respectivamente.

	Academico				
Measures	Pardo	Negro	Indígena	Branco	Amarelo
Media_aproveitamento	4,483	3,473	2,518	4,494	5,768

Slicer:

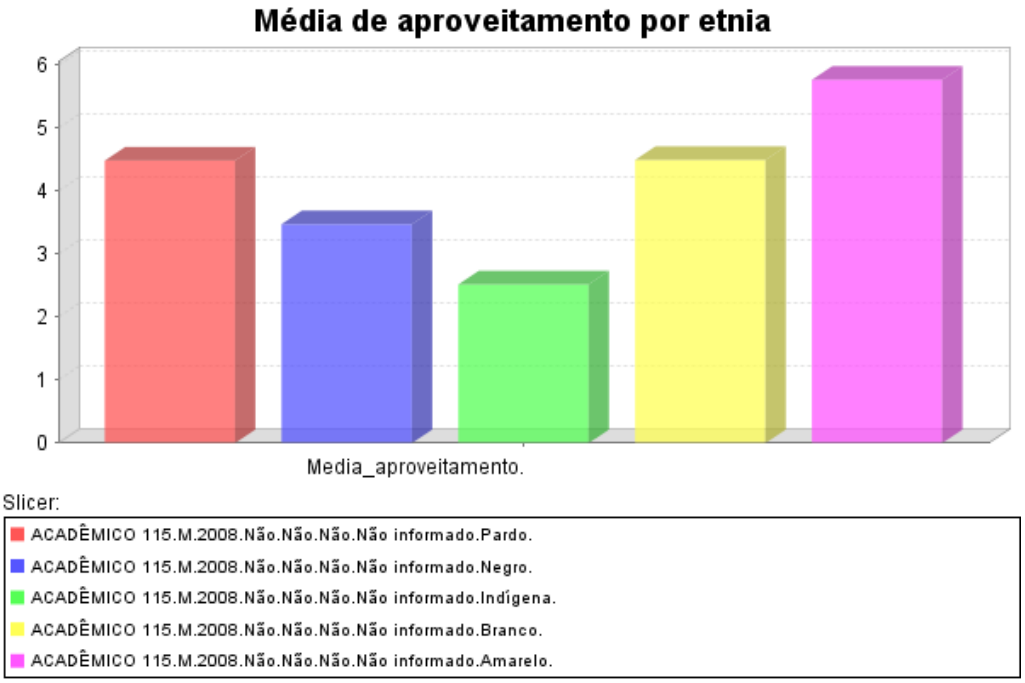

Figura 7 - Média de aproveitamento por etnia.

4 CONSIDERAÇÕES FINAIS

Para a realização deste trabalho, foram pesquisados diversos materiais da literatura, constituídos principalmente de livros e artigos científicos. Essa pesquisa bibliográfica mostrou-se necessária e suficiente para a obtenção dos conhecimentos científicos e técnicos para desenvolver o trabalho proposto.

Conforme discutido na Seção 3.3, a média de aproveitamento dos acadêmicos do curso de BSI é 15,61% superior ao curso de AS. No entanto, a média de aproveitamento dos acadêmicos no curso de BSI apresenta uma queda expressiva a partir de 2010.

O número de diplomados e o número de evasões no curso de BSI tem diminuído consideravelmente a partir de 2007. Isso significa que os acadêmicos estão permanecendo mais tempo retidos no curso, o que contribui para o aumento na demanda de professores e infraestrutura, como salas, livros e laboratórios de informática.

A Figura 4 apresentou as médias de aproveitamento das disciplinas obrigatórias do curso de BSI. É possível observar que algumas dessas disciplinas apresentam médias de aproveitamento bem abaixo das demais, como Algoritmos e Programação (2.9) e Algoritmos (3.0). Certamente, essas disciplinas contribuem diretamente para o alto número de evasão e a retenção dos acadêmicos no curso. Programas de monitoria com as disciplinas que possuem as piores médias de aproveitamento seguramente seria uma boa solução para melhorar o rendimento desses acadêmicos, assim como a média geral do curso.

Outro dado relevante discutido na Figura 6 mostra que acadêmicos com idade superior a 25 anos tem média de aproveitamento menor que os acadêmicos abaixo dessa faixa etária. Nesse contexto, medidas e programas que contribuam para a permanência do discente no curso e melhore o seu aproveitamento nas disciplinas poderiam ser adotados. Por exemplo, programas como bolsa permanência e PIBIC poderiam reforçar os fundamentos conceituais para esse grupo.

Em relação à média de aproveitamento por etnias, presume-se que a Lei das Cotas deverá ocasionar a redução da média de aproveitamento do curso de BSI até 2016. Isso porque a lei obriga as universidades, institutos e centros federais a reservarem para candidatos cotistas metade das vagas oferecidas anualmente em seus processos seletivos até 30 de agosto de 2016, ou seja, 12,5% do total das vagas em 2013, 25% para 2014, 37,5% para 2015, até chegar aos 50% em 2016. Conforme discutido na Seção 3.3 e ilustrado na Figura 7, a média de aproveitamento das etnias com direito a essas vagas é significativamente inferior às demais.

REFERÊNCIAS

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados.** 6ª ed., Addison Wesley, 2011. 788p.

GIL, A. C. **Como elaborar projetos de pesquisa?** 5. ed., São Paulo: Atlas, 2010. 200 p.

GONÇALVES, R. R. **Integração de dados na prática:** Técnicas de ETL para Business Intelligence com o Microsoft Integration Services 2012. São Paulo: Érica, 2012. 158p.

INMON, W. H. **Building the date warehouse.** Wiley, 1992.

INMON, W. H. **Como construir o Data Warehouse.** 2ª ed. Rio de Janeiro: Campus, 1997.

KIMBALL, Ralph. **Data Warehouse Toolkit.** Tradução Mônica Rosemberg; Revisão Técnica Ronal Stevis Cassiolato. São Paulo: Makron Books, 1998.

PENTAHO. Business analytics and business intelligence leaders. 2014. Disponível em: <http://www.pentaho.com>. Acesso em: 13/07/2014.

REDE GLOBO DE TELEVISÃO. **Setor de tecnologia da informação tem déficit de 115 mil trabalhadores.** 2012. Disponível em: <http://g1.globo.com/jornal-hoje/noticia/2012/05/setor-de-tecnologia-da-informacao-tem-deficit-de-115-mil-trabalhadores.html>. Acesso em: 13/08/2013.