Revista EaD & tecnologias digitais na educação

Modelagem de Dados com Data Warehouse e OLAP: um estudo de caso

Everton Castelão Tetila (UFGD)

evertontetila@ufgd.edu.br

***Resumo**: Os data warehouses (DW) oferecem acesso a dados para análise complexa, descoberta de conhecimento e tomada de decisão. O modelo de dados multidimensional utilizado nos DW 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 artigo apresenta uma base de conhecimento criada para investigar as causas potenciais dos problemas do curso de Bacharelado em Sistemas de Informação da UFGD, tais como, alto índice de reprovação, evasão e baixo índice de diplomação. Para a realização desta pesquisa, foram utilizados os aplicativos MySQL Community Server 5.6.21, MySQL Workbench 6.2.3 e PentahoBI-Server CE-4.8.0.*

***Palavras-chave**: descoberta de conhecimento, data warehouse, OLAP.*

***Abstract**: Data warehouses offer data access to complex analysis, knowledge discovery and decision making. The multidimensional data model used in the data warehouses provide decision makers information on the correct level of detail, based on organization and appropriate perspective. Moreover, it is a good choice for Online Analytical Processing (OLAP). In this sense, this paper proposes create a knowledge base, to investigate the potential causes of problems in the Baccalaureate in Information Systems course of the UFGD, such as, high rate of school failure, dropout and low graduation rate. For this research, the MySQL Community Server 5.6.21, MySQL Workbench 6.2.3 and Pentaho BI Server CE-4.8.0 applications were used.*

***Keywords**: knowledge discovery, data warehouse, OLAP.*

1. Introdução

Com o aumento da capacidade de armazenamento de dados e com a crescente automação dos processos por meio de mecanismos sistêmicos, o volume de informa-

ções disponíveis está cada vez maior. Contudo, os dados operacionais provenientes dos processos transacionais das organizações contribuem pouco para a tomada de decisão. Para que os dados gerados pelos processos operacionais possam ser utilizados de forma estratégica pelas organizações, se tornando subsídio para o processo decisório, é fundamental que exista uma transformação natural em seu conteúdo e forma.

O conjunto de dados operacionais deve ser coletado a partir dos diferentes sistemas transacionais existentes para um repositório único, o qual é capaz de consolidar e sincronizar as informações sob a ótica do cliente, de receita, ou de um processo de negócio específico. Este conjunto de informações concentrado em um único repositório permite uma visão não apenas corporativa dos dados associados à operação da empresa, mas também fornece uma visão analítica dos cenários de mercado. Estes dados são armazenados em ambientes conhecidos como *Data Warehouse*[2] e o processo referente à extração das informações, incluindo a transformação dos dados sistêmicos em informações de negócios e a carga destes dados de origem para o repositório central, é denominado ETL (Extração, Transformação e Carga). Após a consolidação dos dados transacionais no repositório *data warehouse*, é possível se criar visões mais agregadas das informações, separadas e formatadas segundo determinados contextos de negócios, auxiliando fortemente nos processos decisórios (PINHEIRO, 2008).

Segundo Elmasri & Navathe (2011), vários tipos de aplicações – OLAP, DSS e aplicações de mineração de dados – são aceitos no repositório *data warehouse*. Para tanto, definimos cada uma delas a seguir:

- **OLAP (Processamento analítico on-line)** termo utilizado para descrever a análise de dados complexos do *data warehouse*.
- **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 artigo propõe criar um repositório *data warehouse* para analisar os dados com o processamento analítico on-line (OLAP). Isso permite 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: muitos estudantes ingressam nas universidades, cursos técnicos e institutos federais, no entanto, poucos egressos na área

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

de Computação 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

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

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 *data warehouse*, por meio do Sistema Gerenciador de Banco de Dados (SGBD) *MySQL 5.6.21*. Logo após, os dados do DW foram projetados para a execução do OLAP com o aplicativo *Pentaho BI Server CE-4.8.0* (PENTAHO, 2014) e, depois, os relatórios gráficos produzidos foram analisados.

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

- **Mysql Community Server:** utilizado para armazenar os dados coletados, disponível em (MYSQL COMMUNITY SERVER, 2014).
- **Mysql Workbench:** utilizado para criar o modelo multidimensional (*data warehouse*), disponível em (MYSQL WORKBENCH, 2014).
- **Pentaho Open Source Business Intelligence:** utilizado para executar o OLAP, disponível em (PENTAHO, 2014).

2. Armazém de dados (Data Warehouse)

Os bancos de dados tradicionais têm suporte para o processamento de transação *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 *Data Warehouse* (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 transacio-

nais, 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 (atributos), mais alto é o nível de granularidade. Por exemplo, a métrica valor_venda poderia ser consultada em um contexto temporal sob múltiplas perspectivas diferentes, a partir dos atributos Ano, Semestre, Mês, dia, hora, etc. Nos primeiros sistemas operacionais a granularidade era tida como certa, pois quando os dados eram atualizados, certamente seria ao mais baixo nível de detalhe, sendo que no ambiente de DW, ela 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 discutidos na 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-reitoria de Graduação (PROGRAD) da UFGD. Essa solicitação foi formalizada por meio de uma Comunicação Interna (CI) e os relatórios, indispensáveis para a realização desta pesquisa, foram entregues em formato digital (.xls e .csv) com os seguintes campos: 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 foram recebidos em uma planilha eletrônica com 14.413 registros, referente a 400 acadêmicos ingressos entre o período de 2006 a 2012. Os valores de registro dos campos “Acadêmico”, “Logradouro” e “Bairro” foram alterados para garantir a privacidade da informação.

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. Desse modo, 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 6.2.3*[3] para executar o OLAP no *PentahoBI-Server CE-4.8.0*.

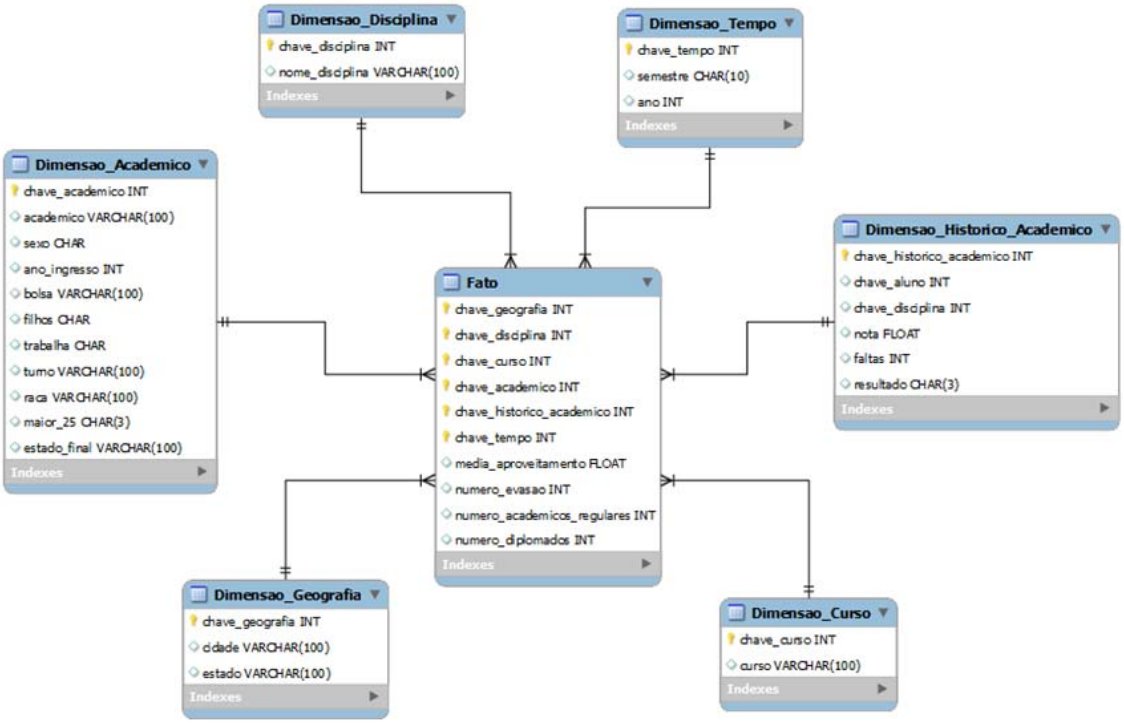

Figura 1 – Projeto lógico do modelo multidimensional de dados.

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.

Kimball (1998) informa que para se distinguir quais campos de dados serão fatos e quais serão atributos (de dimensão) ao projetar um modelo multidimensional, 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.

[3]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*.

3.3 Análise e interpretação dos dados

Esta seção apresenta a análise dinâmica e multimensional dos dados por meio 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

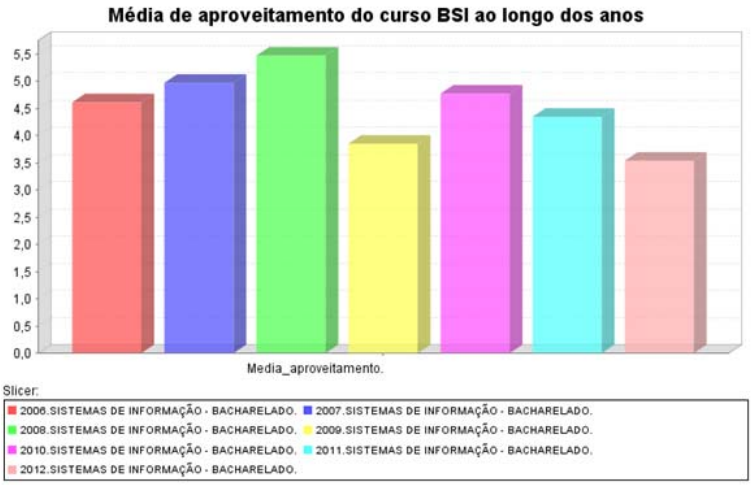

Figura 2 - Média de aproveitamento ao longo dos anos.

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) mudanças nas metodologias de ensino e aprendizagem; (3) 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

Figura 3 - Número de evasões ao longo dos anos.

Revista: EaD & Tecnologias Digitais na Educação, Dourados, MS, Jan/Nov 2014 – n° 3, Vol. 2
ISSN 2318-4051

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.

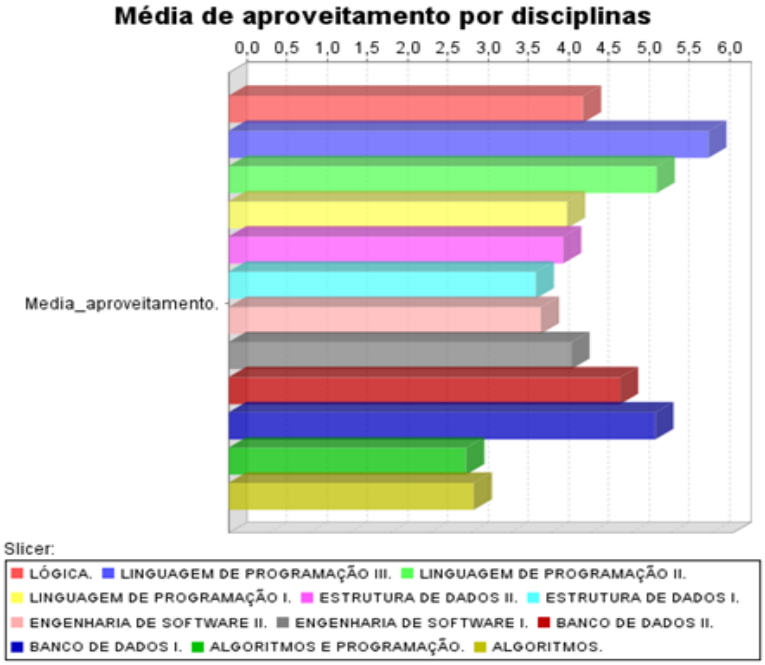

Figura 4 - Média de aproveitamento por disciplinas.

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). Por outro lado, 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.

Revista: EaD & Tecnologias Digitais na Educação, Dourados, MS, Jan/Nov 2014 – n° 3, Vol. 2
ISSN 2318-4051

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. Logo, podemos presumir que o número de diplomados tem uma relação diretamente proporcional à média de aproveitamento, apresentada na Figura 2. Note que ambas as variáveis – média de aproveitamento e número de diplomados – tiveram uma queda considerável nos últimos anos.

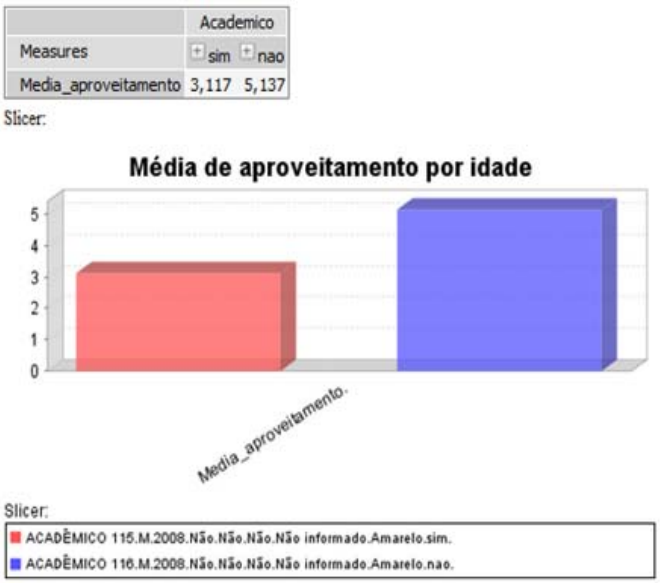

Figura 6 - Média de aproveitamento por alunos maiores e menores que 25 anos.

A Figura 6 apresenta a média de aproveitamento dos acadêmicos distribuídos por idade. Nesta pesquisa foi considerado 25 anos a linha de corte para melhor análise das métricas a partir de duas categorias: acadêmicos com mais ou menos de 25 anos. Assim, podemos utilizar uma métrica (ex: média de aproveitamento) para comparar os acadêmicos mais jovens em relação aos veteranos. Esse valor (25) 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 com menos de 25 anos (5,137) é melhor que a média de aproveitamento dos acadêmicos com mais de 25 anos (3,117). Isso confronta a tese de que acadêmicos veteranos 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 em 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 reserva de 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.

Revista: EaD & Tecnologias Digitais na Educação, Dourados, MS, Jan/Nov 2014 – n° 3, Vol. 2
ISSN 2318-4051

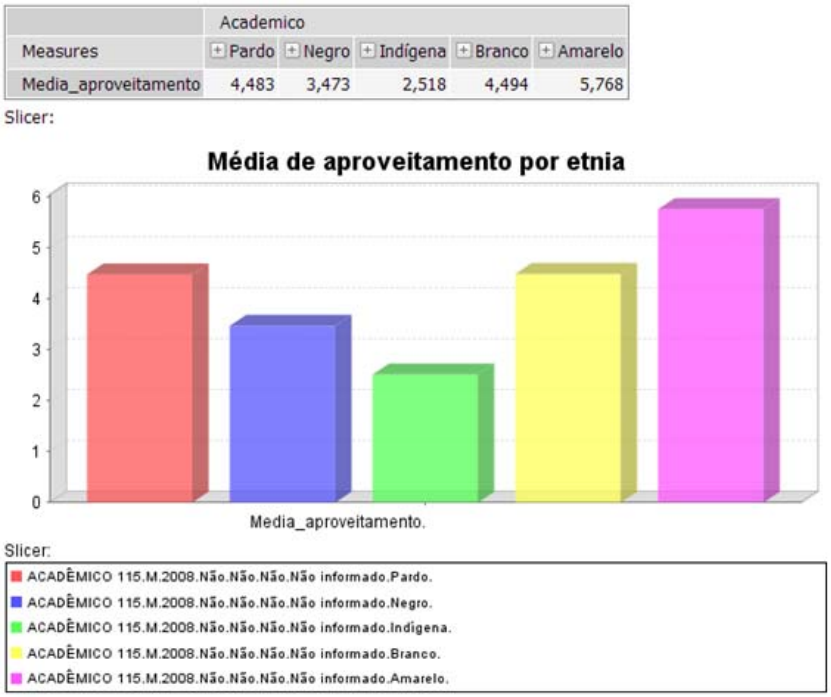

Figura 7 - Média de aproveitamento por etnia.

Conforme discutido nesta seção, diversas variáveis (métricas) que impactam diretamente na qualidade do curso podem ser calculadas por meio de visões analíticas criadas no OLAP *Mondrian*. Como exemplo, as métricas: média de aproveitamento, número de evasão e número de diplomados foram calculadas ao longo do tempo (de 2006 a 2012). Ao mesmo modo, a métrica média de aproveitamento foi calculada em diferentes contextos: por disciplina, por idade e por etnia.

A partir da análise dinâmica e multimensional dos dados, as causas potenciais dos problemas que afetam o rendimento dos acadêmicos podem ser identificadas, bem como, previsões podem ser realizadas de maneira consistente. Esse conhecimento gerado deve ser transformado em ações factíveis e exequíveis que busquem a melhoria da qualidade do curso de modo geral.

4. Considerações finais

Conforme discutido na Seção 3.3, a média de aproveitamento dos acadêmicos do curso de BSI é 15,61% superior ao curso de AS. Apesar disso, a média de aproveitamento dos acadêmicos do 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, por sua vez, 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

Revista: EaD & Tecnologias Digitais na Educação, Dourados, MS, Jan/Nov 2014 – n° 3, Vol. 2
ISSN 2318-4051

uma boa solução para melhorar o rendimento dos acadêmicos retidos, 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 contribuir para 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.

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.

MYSQL COMMUNITY SERVER. Versão 5.6.21, 2014. Disponível em: <http://dev.mysql.com/downloads/mysql/>. Acesso em: 26/09/2014.

MYSQL WORKBENCH. Versão 6.2.3, 2014. Disponível em: <http://dev.mysql.com/downloads/workbench/>. Acesso em: 26/09/2014.

PENTAHO. **Pentaho Open Source Business Intelligence.** Versão 4.8.0.stable, 2014. Disponível em: <http://ufpr.dl.sourceforge.net/project/pentaho/Business%20Intelligence%20Server/4.8.0-stable/biserver-ce-4.8.0-stable.zip>. Acesso em: 26/09/2014.

PINHEIRO, C. A. R. **Inteligência Analítica:** Mineração de Dados e Descoberta do Conhecimento**.** Rio de Janeiro: Ciência Moderna. 2008.

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.