Uma Abordagem de *Data Warehouse* Educacional para Apoio à Tomada de Decisão

José G. de Oliveira Júnior[1], Laudelino C. Bastos[1], Celso A. Alves Kaestner[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
CEP 80.230-901 - Curitiba - PR - Brasil

josjun@alunos.utfpr.edu.br,{bastos,celsokaestner}@utfpr.edu.br

Abstract. *Currently the Higher Education Institutions use of academic management systems and the data generated by these transactional systems has not been effectively exploited. The objective of this research is to present a Data Warehouse approach for these institutions, in order to support decision making by educational managers. The approach is generic and can be applied to a wide range of Educational Institutions. We present a case study applied at the Federal University of Technology - Paraná. The proposal is based on the use of materialized views to build the Data Warehouse and the use of table functions for aggregation of business rules.*

Resumo. *Atualmente as Instituições de Educação Superior utilizam sistemas de gestão acadêmica e os dados gerados por esses sistemas transacionais não tem sido explorados de forma eficaz. O objetivo desta pesquisa é apresentar uma abordagem de Data Warehouse para essas instituições, com o intuito de subsidiar a tomada de decisão pelos gestores educacionais. A abordagem é genérica e pode ser aplicada a uma grande quantidade de instituições de ensino. É apresentado um estudo de caso aplicado na Universidade Tecnológica Federal do Paraná. A proposta baseia-se no uso de visões materializadas para construção do Data Warehouse e no uso de table functions para agregação das regras de negócio.*

1. Introdução

Atualmente, os gestores universitários, em especial os coordenadores de curso, estão convivendo com um ambiente cada vez mais dinâmico, em que precisam tomar decisões, não só em maior número, mas de forma cada vez mais rápida. Isso se deve a grande quantidade de alunos de diferentes perfis de formação, diversidade de cursos e currículos, opções de matrículas mais flexíveis, entre outros fatores.

Por outro lado, as Instituições de Educação Superior (IES) estão armazenando quantidades cada vez maiores de dados, mas a extração de conhecimento dessas bases tem se tornado um desafio. Muitas IESs utilizam processos descentralizados de extração de informações gerenciais que são, na maioria das vezes, demorados, dispendiosos, cansativos e imprecisos, pois reúnem uma grande quantidade de dados que precisam ser coletados de diversas fontes e convertidos em um formato apropriado que possibilite a sua análise. Em contrapartida, as informações disponibilizadas para a tomada de decisão são poucas, dispersas e muitas vezes não possuem acompanhamento histórico.

DOI: 10.5753/cbie.wcbie.2015.1064

Problemas similares foram encontrados em aplicações comerciais, onde o *Data Warehouse* (DW) surgiu como uma ferramenta para consultas analíticas. O conceito de DW traz a abordagem de separação do ambiente de processamento operacional do ambiente analítico. Posteriormente surgiu o conceito de *Data Warehousing* abrangendo o conjunto de tecnologias empregadas nestes ambientes [Inmon 2005].

O DW proporciona um ambiente confiável para o processo de tomada de decisão, facilitando a aplicação de técnicas estatísticas e análises para identificar relações, que a primeira vista podem parecer ocultas.

A proposta deste trabalho é apresentar uma abordagem de um DW Educacional, com o intuito de subsidiar a tomada de decisão pelos gestores educacionais, em especial os coordenadores de curso. A ideia de utilizar um DW é manter o conhecimento corporativo prontamente disponível e em um formato adequado para os tomadores de decisão, para facilitar a posterior aplicação de algoritmos de mineração de dados, pois o DW é um precursor muito útil para exploração de dados [Witten et al. 2011].

O restante do artigo está organizado da seguinte forma: na seção 2 estão descritos alguns conceitos de DW; na seção 3 estão elencados alguns trabalhos correlacionados com esta pesquisa; na seção 4 está detalhado o estudo de caso; e a seção 5 descreve as conclusões e os trabalhos futuros.

2. Conceitos em *Data Warehouse*

Segundo [Inmon 2005], um DW é um conjunto de dados baseado em assuntos, integrado, não volátil, variável em relação ao tempo, de apoio às decisões gerenciais. Orientado a assunto significa que o DW é identificado ou desenvolvido com base no tema principal em um ambiente corporativo.

Uma das tecnologias muitas vezes discutidas no contexto de DW é o processamento de sistemas de gerenciamento de banco de dados multidimensionais, às vezes chamado de processamento *On-Line Analytical Processing* (OLAP) [Inmon 2005]. Sistemas de gerenciamento de banco de dados multidimensionais fornecem uma estrutura que permite que a organização tenha acesso muito flexível aos dados, para separá-los de várias maneiras, e explorar de forma dinâmica a relação entre os dados resumidos e detalhados [Inmon 2005].

A modelagem multidimensional, uma técnica de projeto lógico normalmente usada para construção de DW, é definida sobre dois pilares: tabelas fato e tabelas dimensão. Tabela fato é a tabela primária em um modelo dimensional, onde as medidas de desempenho numéricas do negócio são armazenadas [Kimball e Ross 2011]. As tabelas de dimensão são complementos integrais para uma tabela fato. As tabelas dimensão contêm os descritores textuais do negócio.

No ambiente de Sistemas Gerenciadores de Banco de Dados Relacionais (SGBDR), uma tabela fato é construída geralmente com um registro para cada medição discreta. Esta tabela fato é rodeada por um conjunto de tabelas de dimensão, descrevendo precisamente o que é conhecido no âmbito de cada registro da medição. Este modelo é chamado de *star schema*, ou modelo estrela [Kimball e Ross 2011]. As tabelas de dimensão muitas vezes representam relações hierárquicas no negócio. Assim, a informação descritiva hierárquica é armazenada de forma redundante, mas isto é feito

com o espírito de facilidade de uso e desempenho da consulta. Este modelo é chamado *snowflake schema*, ou modelo flocos de neve [Kimball e Ross 2011].

3. Trabalhos Relacionados

Existem vários trabalhos relacionados a DW aplicados na gestão de instituições de ensino.

[Mansmann e Scholl 2007] apresentam uma metodologia para a avaliação da capacidade educacional e planejamento, inspirada pela reforma do sistema de ensino superior na Alemanha, sendo implementada como um sistema de apoio à decisão que permite a simulação e avaliação de várias propostas e cenários. O sistema proposto integra dados de entrada a partir de um DW e a apresentação da informação é feita usando gráficos, gerando a saída adequada para os tomadores de decisão, revelando detalhes e dependências significativas nos dados.

[Dimokas et al. 2008] introduzem um protótipo de DW educacional e mineração de dados. O trabalho detalha o projeto e desenvolvimento de uma solução de DW que facilita a análise de dados departamentais, propondo uma análise estatística com um conjunto de técnicas de mineração e testes de hipóteses estatísticas.

[Yan e Li 2011] apresentam um modelo de dados multidimensionais, aplicado em um sistema de análise de dados multidimensionais para um DW educacional e mineração de dados. A aplicação é implementada usando arquitetura J2EE com tecnologia RIA Adobe Flash.

O trabalho de [Miranda et al. 2014] propõe um modelo de DW, *dashboard* e uso técnicas de mineração de dados para uma ferramenta analítica aplicada em IESs. O resultado obtido foi um modelo que permite melhorar o desempenho e ajudar no processo de tomada de decisão.

Outros estudos aplicam o conceito de DW em IESs, como nos trabalhos de [Clemes et al. 2001], [Di Domenico 2001], [Heise 2005] e [Dong et al. 2006].

4. Estudo de Caso

O modelo proposto nesta pesquisa foi aplicado na Universidade Tecnológica Federal do Paraná (UTFPR), com a finalidade de analisar os dados acadêmicos da instituição, com o foco inicial de analisar a evasão e a retenção de alunos em cursos de graduação.

A UTFPR é uma IES pública especializada no ensino tecnológico, composta por 13 câmpus distribuídos no Estado do Paraná. Em 2014[1] a instituição atendia a 34.415 discentes, sendo 1.693 alunos matriculados nos 19 cursos técnicos, 5.024 nos cursos superiores de tecnologia, 20.134 nos cursos de bacharelado e licenciaturas, 6.019 nos seus 91 cursos de especialização, 1.251 alunos nos 40 cursos de mestrado e 294 nos seis cursos de doutorado.

4.1. Origem dos Dados

A UTFPR utiliza como principal repositório de dados o SGBDR Oracle desde 1999, sendo os sistemas corporativos desenvolvidos usando Oracle Web Access - OWA, uma extensão web em linguagem PL/SQL.

[1] http://www.utfpr.edu.br/estrutura-universitaria/diretorias-de-gestao/diretoria-de-gestao-da-avaliacao-institucional/relatorio-de-gestao/2014_relatorio-de-gestao

Um DW torna-se efetivo na avaliação de dados com granularidade temporal. Sendo assim, para permitir a análise de dados de anos anteriores foram criadas inicialmente duas séries históricas. A primeira delas representa o vínculo do aluno com o curso em cada período letivo, conforme o modelo relacional mostrado na Figura 1. Essa série histórica contém os principais dados acadêmicos dos alunos em cada período letivo de vínculo com o curso, como: situação do aluno (regular, trancado, etc.), coeficiente de rendimento, número de disciplinas matriculadas, número de disciplinas reprovadas, entre outros atributos. Foi utilizada a granularidade do período letivo (semestral, anual, quadrimestral) por ser uma medida que representa a evolução do aluno na matriz curricular do curso. A série histórica do vínculo do aluno com o curso foi inferida entre os anos de 1979 e 2014, baseado no histórico escolar dos alunos, contendo dados de 105.117 discentes, em um total de 839.223 tuplas. A partir de 2015 os dados são gerados no final de cada período letivo.

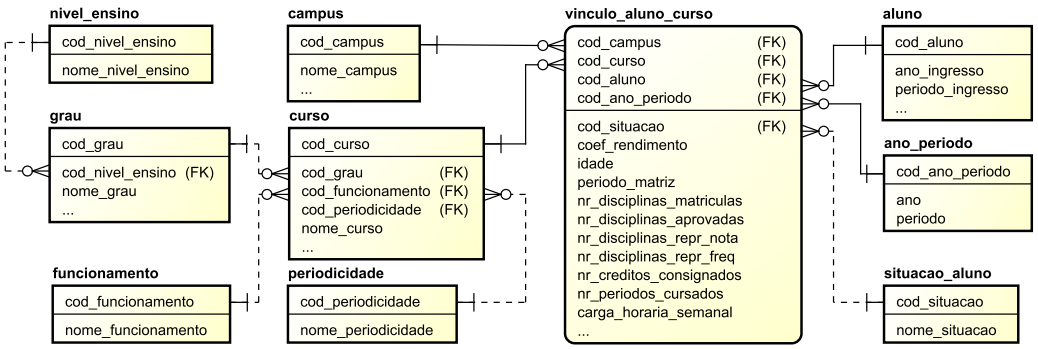

Figura 1. Modelo Relacional da série histórica do vínculo do aluno com o curso

A segunda séria histórica criada representa as disciplinas/turma cursadas em cada período letivo, conforme mostrado na Figura 2, contendo informações como: média e desvio padrão das notas, número de alunos matriculados, número de alunos reprovados por nota, número de alunos reprovados por frequência, carga horária semanal, entre outros atributos. A série histórica foi inferida entre os anos de 1979 e 2014 contendo 430.857 tuplas de 13.110 disciplinas.

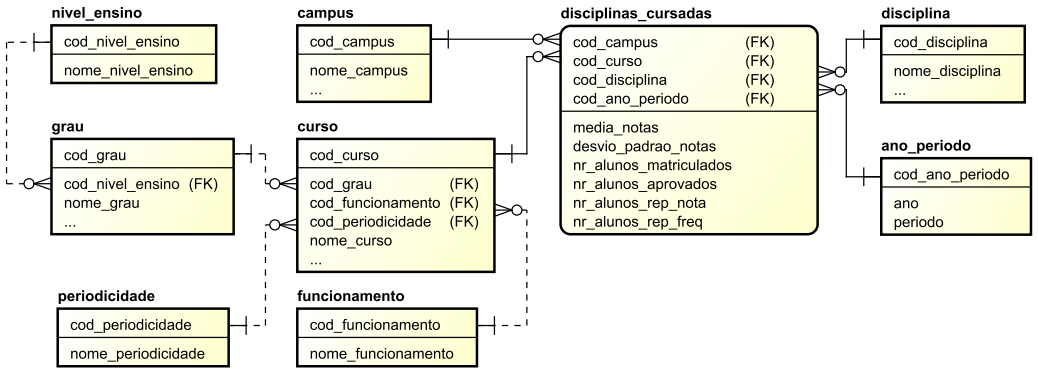

Figura 2. Modelo Relacional das disciplinas cursadas em cada período letivo

4.2. Arquitetura

A arquitetura proposta neste trabalho está baseada em três camadas: dados, negócio e apresentação, conforme mostrado na Figura 3.

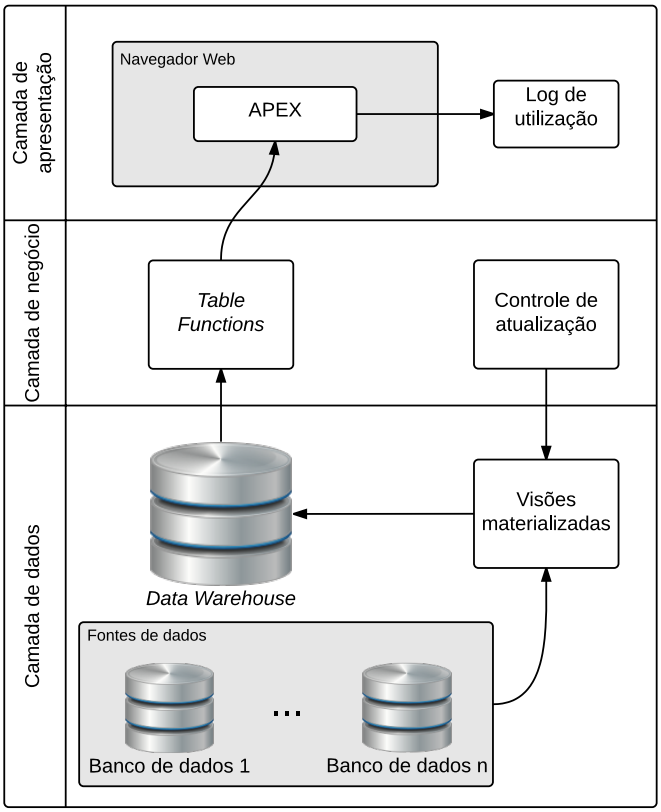

Figura 3. Arquitetura de DW usando visões materializadas e *table functions*

Na camada de dados temos os bancos de dados transacionais de origem, as visões materializadas[2] e o DW. O DW proposto neste projeto é um *Virtual Warehouse*, isto é um conjunto de visões sobre bancos de dados operacionais onde apenas algumas das possíveis visões são materializadas [Han et al. 2011]. Foi escolhida esta abordagem pois permite flexibilidade na criação e administração do DW.

Na camada de negócio temos o controle de atualização das visões materializadas. As visões materializadas são atualizadas com periodicidade diária, semanal ou mensal, conforme a necessidade. Além disso temos as *table functions*, que são funções que produzem um conjunto de tuplas (ou uma tabela aninhada) que podem ser consultadas como uma tabela de banco de dados físico [Murthy et al. 2006]. As vantagens no uso das *table functions* são a agregação das regras de negócio e a facilidade de construção dos relatórios na camada de apresentação, passando somente os parâmetros para uma função PL/SQL. A seguir é mostrado um exemplo de *table function* que gera o relatório mostrado no gráfico da Figura 7.

[2]Uma visão é uma relação derivada, definida em termos de relações base, que é computada todas as vezes em que uma referência a ela é feita. Uma visão é dita materializada quando ela é realmente armazenada na base de dados em vez de ser computada a partir das relações base em resposta a consultas [Quass et al. 1996].

```
select *
from table(mpDW.fcAlunosPorSituacao(:campus,
                                    :nivel_ensino,
                                    :tipo_curso,
                                    :periodicidade,
                                    :funcionamento,
                                    :curso);
```

Em que:
- *mpDW* é uma *package* em PL/SQL que encapsula todas as *table functions*;
- *fcAlunosPorSituacao* é uma *table function* que possui a regra de negócio para retornar as tuplas conforme o filtro selecionado;
- os campos de filtro (dimensões) estão prefixados com ':'.

A declaração da *table function* é mostrada a seguir. Foram ocultados trechos do código (...) para simplificação.

```
function fcAlunosPorSituacao(p_campus  integer,
                             ...
                             p_curso   integer)
  return tarr_AlunosPorSituacao pipelined
is
begin
  for c in c_AlunosPorSituacao(p_campus => p_campus,
                               ...
                               p_curso  => p_curso)
  loop
    pipe row(c);
  end loop;
end fcAlunosPorSituacao;
```

Abaixo temos a declaração do cursor em PL/SQL, onde fica a regra de negócio da consulta.

```
cursor c_AlunosPorSituacaof(p_campus  integer,
                            ...
                            p_curso   integer)
is
  select nomesituacao,
         sum(total) total
  from
  (
    select s.nomesituacao,
           count(*) total
    from alunocurso a
    inner join curso     c on c.curso    = a.curso
    inner join tipocurso t on t.tipo     = c.tipo
    inner join situacao  s on s.situacao = a.situacao
    where (p_campus = 0 or a.campus = p_campus)
      ...
      and (p_curso  = 0 or t.curso  = p_curso)
    group by s.nomesituacao
  )
  group by nomesituacao;
```

Na camada de apresentação foi construída uma aplicação desenvolvida no ambiente de desenvolvimento RAD Oracle Aplication Express - APEX, que é um ambiente web de desenvolvimento de aplicações 4GL que vem integrado com o SGBDR Oracle. Instituições como o CERN (*Conseil Européen pour la Recherche Nucléaire*) também utilizam o ambiente APEX como ferramenta de geração de relatórios baseados em dados somente leitura [Zaharieva e Billen 2009]. O ambiente APEX torna-se uma excelente escolha quando a instituição possui *expertise* em PL/SQL. Na camada de apresentação temos também o log de utilização dos relatórios, que registra o tempo de geração dos

relatórios, permitindo assim que se realize a melhoria no desempenho das consultas.

4.3. Modelagem Multidimensional

Para a criação das tabelas fato foram padronizadas as dimensões mínimas, a fim de manter a uniformidade nos relatórios. As dimensões mínimas utilizadas em todos os relatórios acadêmicos são:

- campus → unidades acadêmicas da instituição;
- nivel_ensino → níveis de ensino: técnico, graduação, stricto sensu;
- grau → graus acadêmicos: bacharelado, licenciatura, tecnologia, etc.;
- curso → cursos vinculados em cada câmpus;
- funcionamento → funcionamento do curso: em atividade, em extinção ou extinto;
- periodicidade → cursos semestrais, anuais ou quadrimestrais.

Para exemplificar, a Figura 4 mostra a modelagem da tabela fato de alunos por situação acadêmica (regular, trancado, etc.). O gráfico que usa essa tabela fato é mostrado na Figura 7.

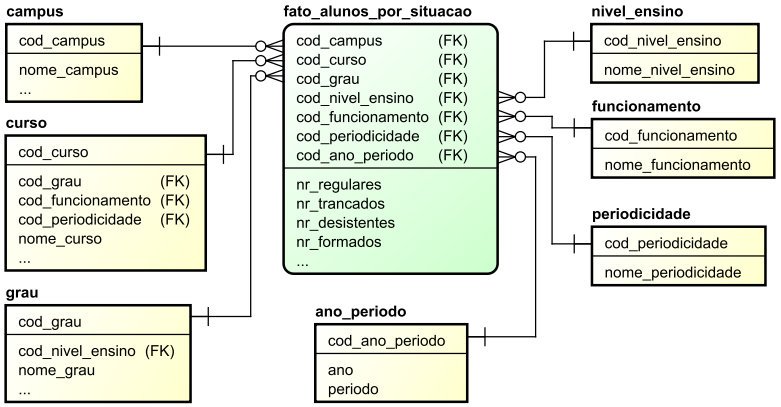

Figura 4. Modelo estrela do número de alunos por situação

A Figura 5 mostra a modelagem da tabela fato do percentual de aprovação das disciplinas cursadas. O gráfico que usa esta tabela fato é mostrada na Figura 8.

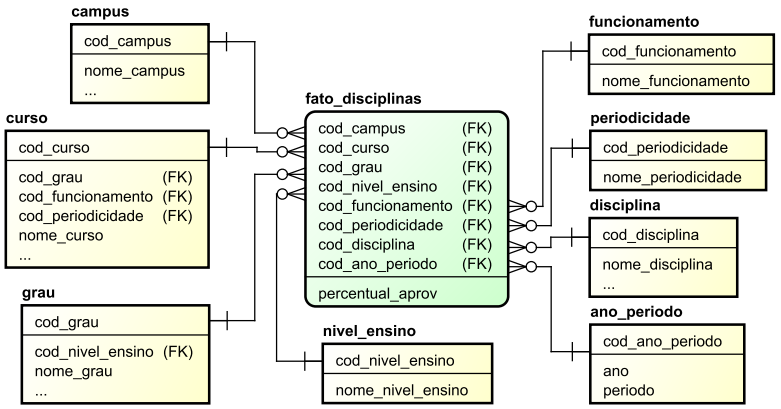

Figura 5. Modelo estrela do percentual de aprovação nas disciplinas

4.4. Relatórios

Os relatórios permitem que os gestores educacionais obtenham informações em quaisquer dos filtros selecionados, desde informações de toda a instituição até informações de um curso específico. Os relatórios disponíveis na aplicação estão agrupados em dois conjuntos: dados atuais e séries históricas. Todos os relatórios disponíveis têm como padrão a visualização em formato tabular, conforme exemplo mostrado na Figura 6, com opção de exportação em formato CSV, e a visualização gráfica, conforme exemplo mostrado na Figura 7, que exibe a quantidade de alunos por situação em cada ano período. Cada usuário pode explorar qualquer combinação de dimensões disponíveis nos filtros.

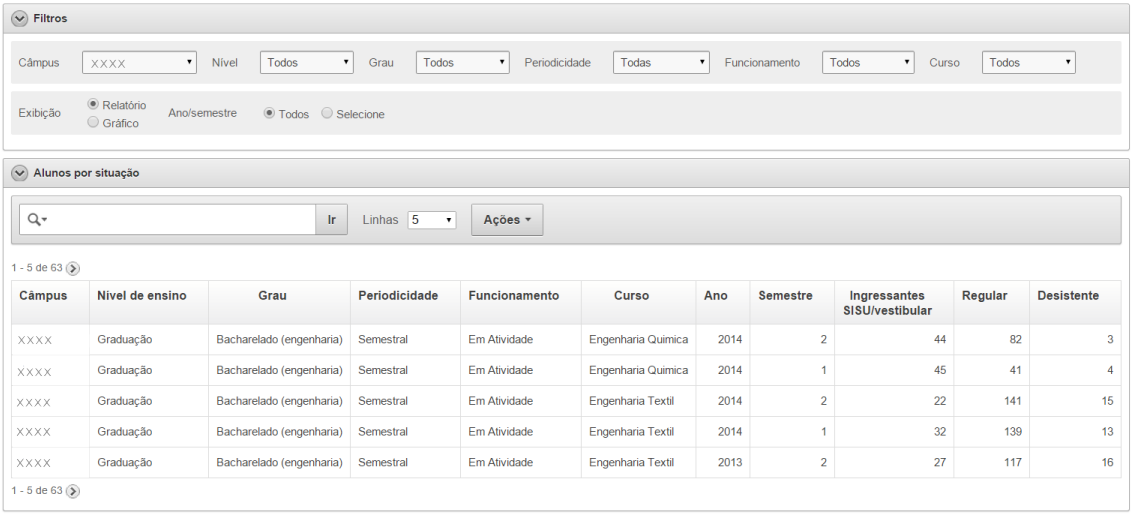

Figura 6. Relatório da quantidade de alunos por situação

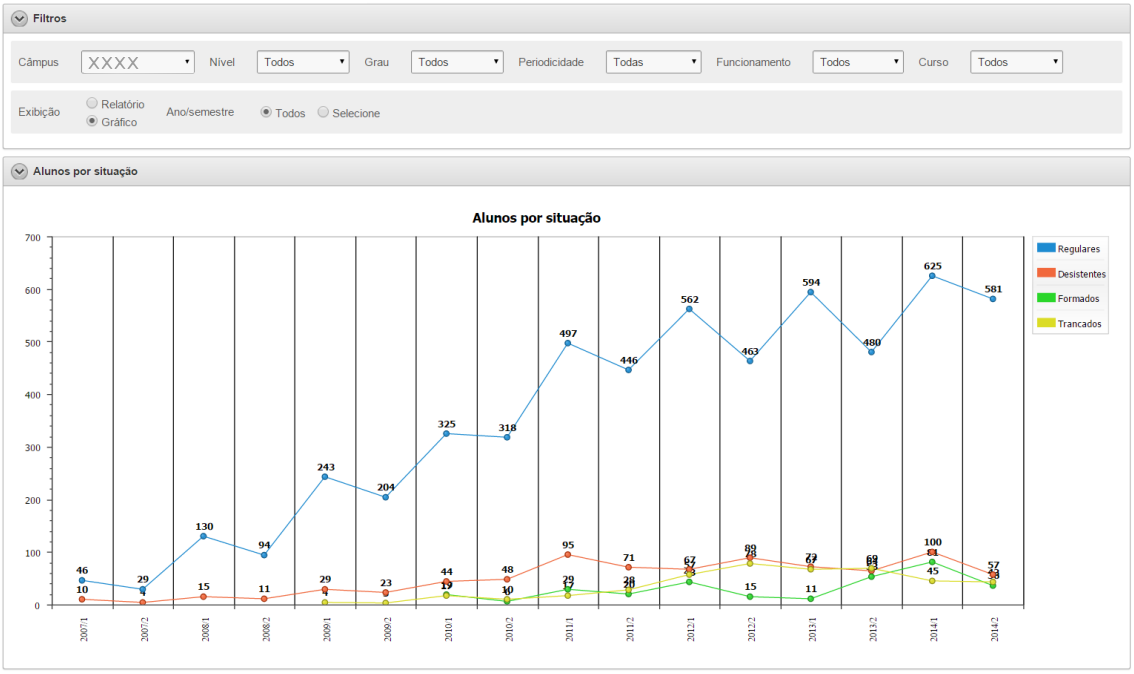

Figura 7. Gráfico da quantidade de alunos por situação

A Figura 8 mostra o percentual de aprovação nas disciplinas, utilizando os filtros de ano/período de início e fim.

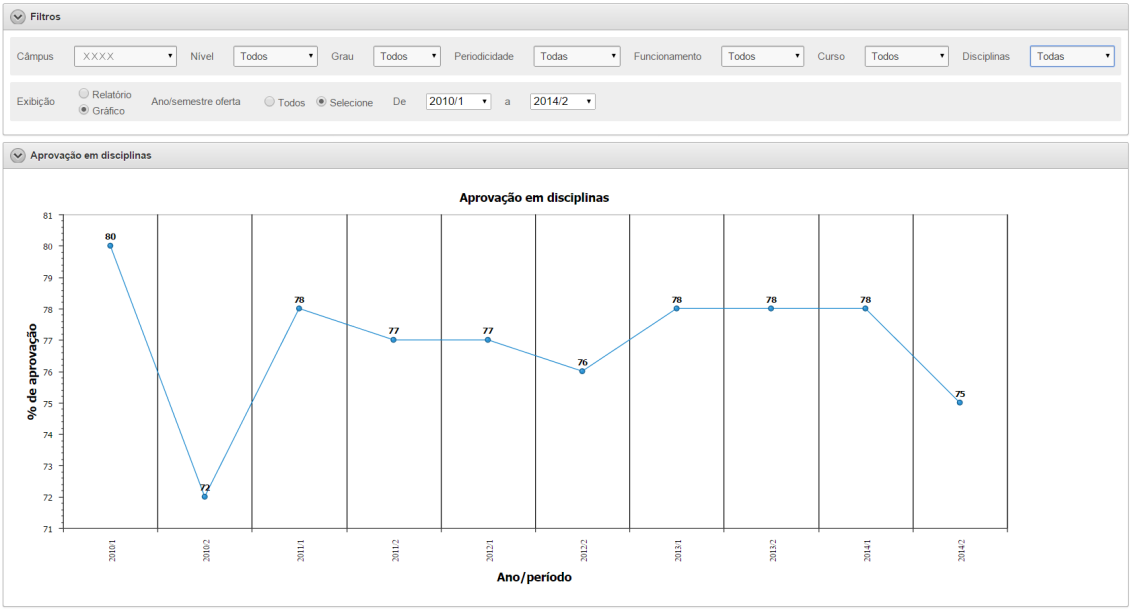

Figura 8. Gráfico do percentual de aprovação nas disciplinas

5. Conclusão e Trabalhos Futuros

A abordagem da modelagem dimensional proposta artigo possui muitas vantagens. A construção do *Data Warehouse* baseado em visões materializadas propicia flexibilidade e facilidade de manutenção, pois se alguma estrutura no banco de dados for alterada é de fácil verificação. O uso das *table functions* permite uma independência da camada de apresentação, permitindo que a interface seja construída em outra linguagem de programação, bastando apenas utilizar a mesma API.

O modelo proposto por este trabalho, que está sendo utilizado pelos gestores da UTFPR como ferramenta OLAP, pode ser aplicado em outras Instituições de Educação Superior, pois tanto as visões materializadas quanto as *table functions* estão disponíveis em outros bancos de dados. A aplicação da modelagem multidimensional permite aos gestores educacionais a obtenção de informações em diversos níveis, auxiliando na tarefa de tomada de decisão.

Este estudo preliminar representa uma parte de um projeto com a finalidade de identificar padrões para a análise da evasão escolar usando mineração de dados, pois o *Data Warehouse* é um precursor muito útil para exploração de dados. Além disso, pretende-se trabalhar na inclusão de mais assuntos no *Data Warehouse* e a criação *Dashboards* para os principais perfis de gestores educacionais.

Referências

Clemes, M. et al. (2001). *Data warehouse como suporte ao sistema de informações gerenciais em uma instituição de ensino superior: estudo de caso na UFSC*. Dissertação de mestrado, UFSC, Florianópolis, SC, BR.

Di Domenico, J. A. (2001). *Definição de um ambiente data warehouse em uma instituição de ensino superior*. Dissertação de mestrado, UFSC, Florianópolis, SC, BR.

Dimokas, N., Mittas, N., Nanopoulos, A., e Angelis, L. (2008). A prototype system for educational data warehousing and mining. In *Panhellenic Conference on Informatics (PCI)*, pages 199–203. IEEE.

Dong, P., Dong, J., e Huang, T. (2006). Application of data warehouse technique in educational decision support system. In *IEEE International Conference on Service Operations and Logistics and Informatics*, pages 818–822.

Han, J., Kamber, M., e Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Heise, D. L. (2005). *Data warehousing and decision making in higher education in the United States*. Tese de doutorado, Andrews University.

Inmon, W. H. (2005). *Building the data warehouse*. John wiley & sons, Indianapolis, IN, USA, 4rd edition.

Kimball, R. e Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

Mansmann, S. e Scholl, M. H. (2007). Decision support system for managing educational capacity utilization. *IEEE Transactions on Education*, 50(2):143–150.

Miranda, E., Suryani, E., et al. (2014). Implementation of datawarehouse, datamining and dashboard for higher education. *Journal of Theoretical & Applied Information Technology*, 64(3).

Murthy, R., Sethi, A., Ghosh, B., Thusoo, A., Agrawal, S., e Yoaz, A. (2006). Method and system for pipelined database table functions. US Patent 7.103.590.

Quass, D., Gupta, A., Mumick, I. S., e Widom, J. (1996). Making views self-maintainable for data warehousing. In *Fourth International Conference on Parallel and Distributed Information Systems*, pages 158–169. IEEE.

Witten, I. H., Frank, E., e Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Yan, X. e Li, X. (2011). A multidimensional data analysis system based on mda for educational data warehousing. In *6th International Conference on Computer Science & Education (ICCSE)*, pages 88–94. IEEE.

Zaharieva, Z. e Billen, R. (2009). Rapid development of database interfaces with oracle apex, used for the controls systems at cern. In *International Conference on Accelerator and Large Experimental Physics Control Systems*. ICALEPCS.