Uma Proposta de Aplicação de Business Inteligence para o Censo das IES

André Monteiro[1], Fernando Cezar Borges[1], Hugo Alex Conceição Silva[1], Mário Jorge Pereira[1], Mauricio Cesar Santos da Purificação[1]

[1]Centro de Processamento de Dados – Universidade Federal da Bahia (UFBA)
40.170.110 – Salvador – BA – Brazil

{andremp,fercez,mariojp,mauricioc}@ufba.br

Abstract. *This paper presents a proposal to optimize the chain of routines performed during the preparation of the census of Higher Education Institutions. This improvement will be achieved by adopting a model that uses Business Intelligence tools during the process of collecting and validating data from the National Institute of Educational Studies Anisio Teixeira. The proposal is being developed at the Federal University of Bahia and could be applyed to other Higher Education Institutions in the future.*

Resumo. *Este artigo apresenta uma proposta para otimizar a cadeia de rotinas executadas durante a elaboração do censo das Instituições de Ensino Superior. Esta melhoria será dada através da adoção de um modelo que utiliza ferramentas de Business Intelligence nas etapas de coleta e validação dos dados junto ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. A proposta apresentada está sendo desenvolvida na Universidade Federal da Bahia e poderá ser aplicada em outras Instituições de Ensino Superior posteriormente.*

1. Introdução

O censo das Instituições de Ensino Superior (IES) é realizado anualmente e seus dados são disponibilizados para a sociedade em geral. O tempo entre a disponibilização dos indicadores obtidos ao final do censo e a próxima coleta não permite que as IES procedam com as intervenções necessárias para a melhoria dos indicadores. A proposta deste trabalho é definir um modelo único de Data Warehouse (DW) para todas as IES que atenda aos indicadores coletados durante o censo. Como resultado, espera-se facilitar e agilizar a coleta dos dados, minimizar o tempo entre a coleta e a divulgação dos resultados, permitir um monitoramento constante dos indicadores por parte das IES e consequentemente melhorar a qualidade da informação fornecida.

Este artigo encontra-se organizado da seguinte forma: Na seção 2 é detalhado o processo do censo que é realizado pelas IES anualmente. Na seção 3 é apresentado o conceito de *Business Intelligence* (BI) e é detalhada uma arquitetura geral para a construção de soluções de BI. Na seção 4 é apresentada a suíte Pentaho BI que tem sido utilizada como plataforma de desenvolvimento na Universidade Federal da Bahia (UFBA) e faz parte da proposta deste trabalho. Na seção 5 é apresentado o modelo proposto e na Seção 6 é apresentada a conclusão e os trabalhos futuros referentes a esta proposta.

2. Censo

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) realiza uma coleta anual de dados juntos às IES. A coleta dos dados tem como referência as diretrizes gerais previstas pelo Decreto nº 6.425 [1] de 4 de abril de 2008 [2]. A partir dos dados obtidos são elaborados indicadores educacionais. Este processo de coleta e disponibilização dos dados (na forma de indicadores educacionais) é conhecido como censo da educação superior.

Ao fim do censo as informações são disponibilizadas oficialmente tanto para a comunidade acadêmica como para a sociedade em geral. Esta divulgação reúne informações sobre as instituições de ensino superior, e entre elas estão: dados dos cursos de graduação presenciais ou à distância, cursos sequênciais, vagas oferecidas, inscrições, matrículas, ingressantes e concluintes além de informações sobre docentes nas diferentes formas de organização acadêmica e categorias administrativas.

Para chegar aos indicadores, as informações são obtidas por meio de um questionário eletrônico que fica aberto durante um período para o preenchimento pelas IES. O preenchimento ocorre, normalmente, dentro do primeiro trimestre do ano. Após este período, os dados são novamente colocados a disposição das IES para validação e/ou correção. O Inep pode então realizar uma análise na base de dados gerada pela coleta para verificar a ocorrência de inconsistências nas informações obtidas. Depois desta verificação o Inep divulga os indicadores gerados durante o censo, normalmente, no último trimestre do ano.

Durante este processo é necessário o envio de informações ao Ministério da Educação (MEC), o que ocorre através de uma extração de dados, seja esta feita por meio de planilhas eletrônicas, relatórios e/ou carga de dados obtidos a partir dos sistemas operacionais das IES. Devido ao volume de dados essa etapa é bastante custosa. Outro problema no uso dos dados dos sistemas operacionais é a sua constante atualização, o que dificulta a auditoria após a carga, por conta dos relatórios refletirem o estado atual do sistema. Deste modo, é necessário arquivar e manter relatórios que são gerados no momento da extração dos dados ou do preenchimento das planilhas. A Figura 1 detalha o processo do Censo realizado pelas IES.

3. *Business Intelligence*

BI pode ser visto como um processo sistemático de aquisição, tratamento e análise de informações em que os dados internos e externos de uma organização são integrados para gerar informação pertinente para o processo de tomada de decisão. O papel do BI é criar um ambiente informacional com processos através dos quais os dados operacionais possam ser coletados, tanto dos sistemas transacionais como de fontes externas, e analisados, revelando dimensões ”estratégicas”do negócio [Petrini et al. 2006].

Analisando a Figura 2, podemos visualizar um modelo tradicional para o desenvolvimento de soluções de BI. Segundo este modelo, existem basicamente dois componentes representados pela Arquitetura Técnica de Povoamento (ATP) e pela Arquitetura

[1] http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2008/Decreto/D6425.htm
[2] http://www.censosuperior.inep.gov.br/

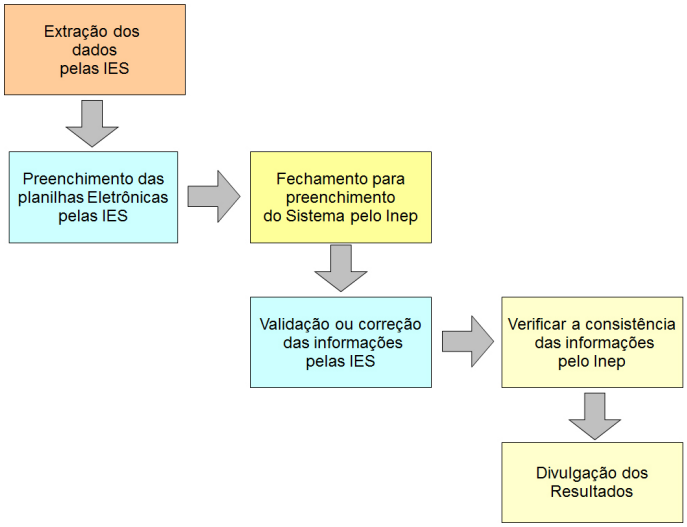

Figura 1. Processo de Coleta e Divulgação de Dados do Censo.

de Acesso aos Dados (AAD). A ATP define como os dados oriundos dos sistemas operacionais (fontes de dados externas e internas) são carregadas no DW através do processo de Extração, Transformação e Carga (ETC) (do inglês *Extract, Transform and Load* (ETL). No processo de ETL, os dados operacionais são uniformizados, selecionados, transformados, e carregados no DW da solução de BI.

A AAD define como os dados armazenados no DW serão acessados pelos usuários. Por exemplo, pode-se construir Data Marts (DMs), que são como um DW armazenando dados a níveis departamentais ou vínculados a áreas específicas do negócio, de modo que, caso o DW necessite de algum suporte ou manutenção, os departamentos continuem acessando seus dados nos respectivos DMs. Por conseguinte, um novo processo de ETL será realizado para que do DW possa realizar-se a carga nos dados dos DMs.

Além desses dois componentes, existe uma camada de metadados que engloba toda a solução. Os metadados devem possibilitar o registro de todas as informações sobre os dados como suas origens e os processos de transformação sofridos. São de extrema importância dentro do ambiente de DW, pois representam uma visão integrada das bases de dados que fazem parte deste ambiente. São utilizados para construir, manter, gerenciar e utilizar o DW [da Costa and Anciães 2001].

4. Suíte Pentaho BI

O Pentaho é uma alternativa comercial e de código fonte aberto para projetos de BI [Almeida 2010]. É oficialmente formado por cinco componentes que foram criados separadamente com o objetivo de solucionar etapas específicas do processo de construção de uma solução de BI. Por isso, também podem ser utilizados individualmente, interagindo com outras ferramentas. Estes componentes são complementares entre si e fornecem suporte completo a uma solução de BI quando utilizados em conjunto.

Estes componentes e suas funcionalidades são detalhados abaixo [Bouman and van Dongen 2009]:

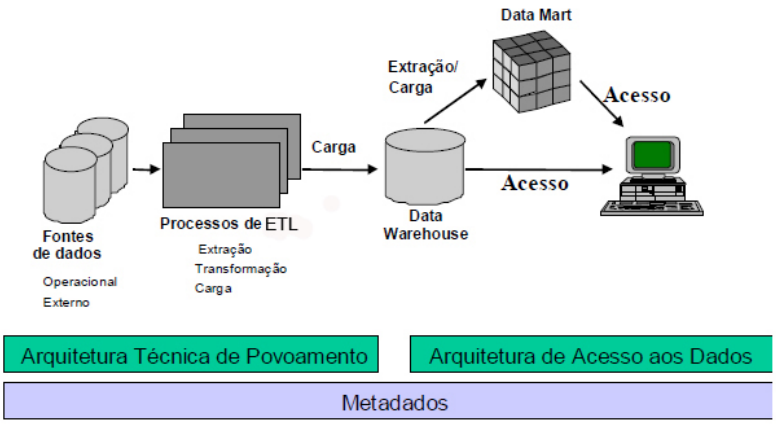

Figura 2. Arquitetura de BI - Visão Simplificada. Adaptado de [da Costa and Anciães 2001].

- **Pentaho BI Server**: Contém a infra-estrutura responsável por integrar a solução de BI. Ele oferece serviços essenciais para construir, implantar, executar e dar suporte aos relatórios, análise de cubos, dashboards, etc. Também provê mecanismos de autenticação e controle de acesso;
- **Pentaho Data Integration (PDI)**: Também conhecido por Kettle, é responsável pela construção e execução do processo de ETL;
- **Pentaho Analysis**: Também conhecido como Mondriam, é um servidor OLAP [3] que permite analisar interativamente grandes conjuntos de dados armazenados em bancos de dados relacionais (pois é baseado na arquitetura ROLAP);
- **Pentaho Reporting**: É uma coleção de projetos de código fonte aberto, focados principalmente na criação, produção e publicação do conteúdo de relatórios de forma rica e sofisticada;
- **Pentaho Dashboards**: Este módulo da suíte permite a criação de painéis de controle, mais conhecidos como *Dashboards* [4] e através dele é possível reunir em uma mesma tela os principais indicadores de um departamento ou de toda a empresa;
- **Weka**: É um conjunto abrangente de ferramentas para aprendizado de máquina e mineração de dados. Pode ser usada para ajudar a entender melhor o negócio da instituição. O Weka não faz parte originalmente da suíte, mas pode ser usado em conjunto para a análise de dados.

5. Modelo Proposto

A proposta deste trabalho é que seja definido um modelo único de DW que atenda aos indicadores utilizados pelo MEC durante o censo das IES. Este modelo seria replicado em todas as IES sendo que suas dimensões seriam preenchidas pelo MEC cabendo às IES o preenchimento das tabelas fato do modelo. As IES iriam então extrair os dados dos

[3]O termo OLAP refere-se a um conjunto de ferramentas voltadas para acesso e análise *ad-hoc* de dados, com o objetivo final de transformar dados em informações capazes de dar suporte às decisões gerenciais de forma amigável, flexível e em tempo hábil ao usuário [Araújo et al. 2007]

[4]Painel com um conjunto de indicadores gráficos que, por estar num formato visual, facilita a compreensão e assimilação das informações. Geralmente estes indicadores gráficos são integrados entre si [Almeida 2010].

seus sistemas operacionas, tratar/unificar e fazer a carga em seu modelo dimensional e no modelo dimensional do censo. Deste modo, a coleta dos dados para o MEC se resumirá a extração dos dados nas tabelas fato do modelo das IES e a carga nas tabelas fato do Inep (modelo do censo).

Observa-se a existência de duas camadas dimensionais no modelo proposto, a primeira camada refere-se ao próprio DW que seria mantido por cada IES segundo às suas necessidades e a segunda camada, a um DW de modelagem unificada para todas as IES de modo que os indicadores necessários ao censo sejam contemplados. Para as IES será entregue uma infraestrutura pronta de BI (Um *live cd* contendo o sistema operacional Debian[5] e a suíte Pentaho BI, previamente configurada com o banco de dados PostgreSQL [6]). Apenas a configuração do preenchimento (carga) dos dados seria necessária. A própria IES poderia fazer consultas através da ferramenta de BI e assim validar os dados gerados.

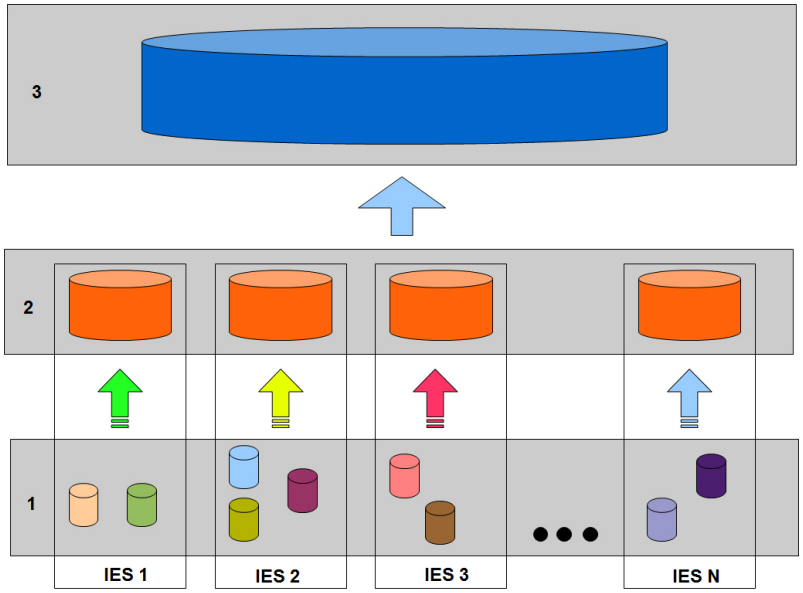

Figura 3. Arquitetura do Modelo Proposto.

A Figura 3 apresenta como estão definidas as camadas da solução proposta: Na camada 1 temos os diferentes bancos operacionais das diferentes IES. Entre a camada 1 e a camada 2 temos o primeiro ETL, devido a pluralidade de sistemas e a necessidade de uniformização dos dados. O modelo da camada 2 deve contemplar os indicadores necessários ao censo, podendo também atender à necessidades especifícas das IES. Neste caso, pode se considerar que cada IES mantenha um DW particular e a partir dele seja realizada a carga no DW que será utilizado pelo censo.

Na camada 3, o DW possui o mesmo modelo para as diferentes IES. Entre as camadas 2 e 3 temos o segundo ETL que deve ser o mesmo para todas as IES, onde podem ser efetuadas validações além da carga dos dados 4. Na 3° e última camada temos o DW geral que contempla todas as cargas das IES.

Atualmente a camada 1 e 2 tem sido trabalhada pela Universidade Federal da

[5]http://www.debian.org/
[6]http://www.postgresql.org/

Bahia (UFBA). A modelagem feita para o DW da UFBA já contempla a maioria dos indicadores do censo relacionados à alunos e docentes (este DW contempla ainda outras análises realizadas pela instituição, como o planejamento acadêmico, carga horária semestral de docentes etc...). O que precisa ser feito no momento é a modelagem do restante dos indicadores, a carga dos dados, a modelagem final da camada 3 e a disponibilização da solução completa para outras IES.

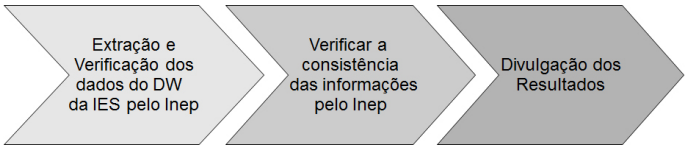

Figura 4. Etapas para Coleta e Divulgação do Censo com Adoção do Novo Modelo.

Com a adoção dessa proposta por parte da IES, a coleta e divulgação dos dados do censo passa a ser mais direta, sendo necessário apenas a realização de 3 etapas por parte do Inep (Figura 4). As fases de extração de dados seriam facilitadas pelo uso da ferramenta PDI e as etapas de validação e análise de dados seriam facilitadas pelas ferramentas *Pentaho Analysis*, *Pentaho Reporting* e *Pentaho Dashboards*, na criação de cubos OLAP, relatórios gerenciais e *dashboards* com os indicadores educacionais.

6. Trabalhos Futuros e Conclusão

A adoção do modelo proposto por este artigo vai permitir um processo mais ágil e de melhor qualidade para geração dos dados relacionados ao censo das IES. O acompanhamento dos indicadores por parte das IES e as intervenções a serem realizadas pelas mesmas podem ser feitas antes mesmo do término de todas as etapas do censo. O uso de ferramentas *Open Source* permite uma maior facilidade na replicação da proposta e na disponibilização das ferramentas para as IES, além de seguir a recomendação do Governo Federal na adoção de programas caracterizados por esse tipo de licenciamento.

Com a disponibilização destas ferramentas livres já configuradas, as IES podem inclusive trabalhar no sentido de desenvolver soluções customizadas de acordo com suas necessidades específicas.

Referências

Almeida, G. R. (2010). mpentaho uma ferramenta de acesso às análises olap da suíte de business intelligence pentaho em dispositivos móveis android. Technical report, Universidade Federal da Bahia.

Araújo, E. M. T., de Lourdes Souza Batista, M., and de Magalhães, T. M. (2007). Olap: Características, arquitetura e ferramentas. Technical report, Instituto Vianna Júnior - Faculdades Integradas Vianna Júnior.

Bouman, R. and van Dongen, J. (2009). *Pentaho Solutions Business Intelligence and Data Warehousing with Pentaho and MySQL*. Wiley Publishing, Inc.

da Costa, A. F. and Anciães, F. C. (2001). Aspectos de criação e carga de um ambiente de data warehouse. Technical report, Rio de Janeiro.

Petrini, M., Freitas, M. T., and Pozzebon, M. (2006). Inteligência de negócios ou inteligência competitiva? noivo neurótico, noiva nervosa.